

PREPARATION MOTRICE ET SELECTION DE CIBLES ARTICULATOIRES ACCENTUEES

L. Crevier-Buchmann, J.F. Bonnot*
C. Chevrie-Muller et C. Arabia-Guidet

INSERM, Laboratoire de Recherche sur le Langage
Paris, France.

*Département de Linguistique Générale et URA CNRS 668, Strasbourg - France

ABSTRACT - The influence of preparatory mechanisms on reaction time (TR) was measured in two different situations: one with simple choice (TRS) and one with random choice (TRC). The different signals measured were the acoustic signal and 3 EMG signals. The material was quadrisyllabic non words and stress was placed on one of the 4 syllables, either predetermined (TRS) or at random (TRC). 1/ TR was significantly longer in the random situation than in the simple choice situation. 2/ In the random situation when compared with the simple choice situation, TR was significantly longer when the stress was on the first and to a lesser extent on the second syllable.

1 - INTRODUCTION - Bien qu'il ne soit pas aisé d'évaluer l'importance respective des divers processus contrôlant la mise en oeuvre du mouvement, on a des raisons de penser que leur paramétrisation est assez strictement hiérarchisée: il est nécessaire de distinguer des unités de planification (13) (phonèmes, mots et syntagmes) et des unités d'exécution (syllabes et, plus généralement, unités prosodiques). Une telle hiérarchie a d'importantes conséquences en ce qui concerne les commandes. Il peut exister une indépendance des articulateurs à ces divers niveaux d'organisation (5). Il a aussi été montré que, dans l'élaboration d'un modèle de production, l'organisation temporelle est sensible: - à la complexité intrinsèque de la séquence (nombre de syllabes, assemblage consonantique déclenchant des effets coarticulatoires plus ou moins durs (1), ou même propriétés subphonémiques (4),

- et à des contraintes "externes" (modification de la vitesse d'élocution ou opposition voix normale/voix chuchotée/voix criée (2)). Ces contributions motrices diversifiées illustrent bien l'extrême plasticité des systèmes de régulation du timing. L'une des questions capitales, en psychophysologie de la motricité (7), que l'on commence seulement à aborder en production de la parole (3) concerne les processus de préprogrammation et la régulation "on line". Depuis quelques années, on s'est particulièrement attaché à préciser les mécanismes préparatoires définis comme des modifications physiologiques "which not only anticipate the action (...) but can be experimentally manipulated and have a predictive value for performance efficiency" (10). Le traitement de l'information nécessaire à l'encodage est donc ajustable en fonction des caractéristiques intraséquentielles. Nous nous proposons d'examiner les relations de ces mécanismes préparatoires avec la durée des temps de réaction simple (TRS) et avec choix (TRC). On s'est placé dans le cadre théorique d'un modèle rythmique du langage (9). Certes, en français, on ne peut pas parler d'accent de mot; il n'en reste pas moins qu'au plan de la mise en oeuvre des effecteurs, l'alternance de temps forts et de temps faibles est primordiale. On testera en conséquence les hypothèses à partir de logatomes quadrisyllabiques: les temps de réaction devraient être plus longs dans le cas où la localisation des "accents" n'est connue du sujet qu'au moment où le stimulus lui est fourni, parce qu'il est vraisemblable que la tâche cognitive est alors plus complexe. On fait par ailleurs une deuxième hypothèse: les variations de la

force affectant la phase la plus périphérique de la préparation motrice devraient avoir une pertinence particulière. On utilise en conséquence la méthode électromyographique (5, 6, 8, 10). Enfin, dans la mesure où il semble que les stratégies musculaires des structures dotées d'un impact syllabique ou séquentiel (mandibule) seront assez différentes de celles dotées d'un impact segmental ou infrasegmental (lèvres lors de l'occlusion), on examinera ces deux structures.

2. MATERIEL, METHODES

Le matériel informatique utilisé nécessite 2 micro-ordinateurs PC, 1 imprimante, 1 carte convertisseur 4 voies; 3 programmes sont utilisés successivement.

Le matériel linguistique se composait de logatomes quadrisyllabiques CVCVCV CV (mamamama) dont une syllabe devait être accentuée. La méthode mise au point pour mesurer l'intervalle stimulus-réponse est la suivante: chaque logatome est présenté sur un écran, un point apparaît sous la syllabe à accentuer (tirée au hasard, pour les TR à choix complexe) à un temps T_0 , séparé de l'affichage par un intervalle variable (1,25 sec. + temps aléatoire de 0 à 0,25 sec.). Le temps d'exposition à partir de T_0 est de 0,25 sec., puis l'écran s'éteint. Le logatome suivant est présenté après un intervalle de durée aléatoire à partir de l'effacement du précédent (4 sec. + durée aléatoire de 0 à 1 sec.). En situation de temps de réaction simple (TRS) une expérience comprend pour chacune des syllabes à accentuer, 8 présentations (pour 2 expériences réalisées ici, $8 \times 4 \times 2 = 64$); le sujet dès la première présentation sait que les 7 suivantes porteront sur la même syllabe. En situation de temps de réaction à choix complexe (TRC) l'affichage du point se fait de façon aléatoire sous l'une des 4 syllabes. Le nombre de présentations pour chaque syllabe, en 3 expériences, a été de 21 (soit $21 \times 4 = 84$ présentations). Compte tenu des échecs (faux départs, erreurs dans la place de l'accent) 54 TRS ont été mesurés (syll. A = 14, syll. 2 = 14, syll. 3 = 13, syll. 4 = 13; taux de réussite global 84%) et 52 TRC (syll. 1 = 16, syll. 2 = 13, syll. 3 = 11, syll. 4 = 12; taux de réussite global 62%). Les effectifs complets de réponse sans erreur

ont été utilisés pour les comparaisons de variances en plan factoriel (2 facteurs: position de la syllabe accentuée à 4 niveaux: TRS + TRC). Le nombre de répétitions conservé pour l'analyse a été de 11 (total des TR analysés = 88). TRS et TRC ont été mesurés pour le signal acoustique de parole (intervalle entre T_0 et début du signal) et de la même façon pour les signaux électromyographiques (EMG) enregistrés au niveau des 3 muscles: digastrique (DIG), orbiculaire supérieur des lèvres (OOS) et orbiculaire inférieur (OOI) grâce à des électrodes de surface.

3. RESULTATS

3.1 Influence de la situation sur les TR (Tableau 1)

- La différence de durées de TR en fonction de la situation est très significative (test de F pour le signal acoustique et chaque signal EMG, $p < 0.01$).

- L'analyse "syllabe par syllabe" de la différence entre TRS et TRC par le test de Student (tableau 1) montre que celle-ci n'est significative pour tous les signaux (acoustique, EMG) que lorsque l'accent est sur la 1ère syllabe. Sur la 2ème syllabe, la différence n'est significative que pour le signal acoustique et le digastrique. Pour l'accent en 3ème et 4ème syllabe aucune différence n'est significative.

3.2. Influence de la place de l'accentuation sur le TR

- L'analyse de variance montre qu'il existe globalement une différence significative entre les TR en fonction de la place de l'accent, pour les mesures effectuées sur le signal acoustique ($F^3_{80} = 3.0$ $p < 0.05$).

- La figure 1 montre bien, par ailleurs, qu'en 3ème et à un moindre degré en 4ème syllabe la différence entre les moyennes de TRS et TRC tend à diminuer.

4. DISCUSSION

4.1. Les phénomènes préparatoires sont bien mis en évidence dans la mesure où les durées de TR pour l'accent en 1ère syllabe, sont toujours plus élevées dans la condition avec choix par rapport à la situation sans choix. Autrement dit, le sujet a besoin d'un surcroît de temps pour mettre en oeuvre son action dans les

cas où le signal de départ coïncide avec l'information concernant la syllabe à accentuer. Ceci montre que la tâche cognitive est plus complexe et confirme que, dans la situation simple, la séquence est partiellement préprogrammée (7). Cet effet décroît lorsque l'accent frappe la 2ème et surtout la 3ème syllabe. Il est possible de proposer 2 explications complémentaires :

- de nombreux travaux ont clairement mis en évidence la particulière complexité des ajustements articulatoires initiaux ; Kent (10) note par exemple que "the generation of response specifications is more complex for initial than medial or final consonants". Il faut souligner que dans l'expérience que nous présentons, l'exécutant est obligé de définir à la fois les paramètres spatio-temporels concernant les segments à venir, et de contrôler les indices aérodynamiques générant une augmentation de l'énergie acoustique.

- d'autre part, l'attaque recèle très vraisemblablement une information à partir de laquelle s'effectue une bonne partie de la paramétrisation intraséquentielle (3).

4.2. On constate que l'écart entre TRS et TRC diminue considérablement en 3ème syllabe. On peut suggérer qu'il s'agit là d'un effet du "planning réparti" (7) qui se poursuit alors que l'exécution a débuté : il est vraisemblable que le système d'encodage est en mesure de calculer qu'au moment où la cible sera atteinte, une grande partie des spécifications sera disponible. Il s'ensuit que l'activité de

préprogrammation peut être quelque peu allongée. En 4ème syllabe, on relève toutefois une nouvelle augmentation de la différence quoique non significative. On sait qu'en français, la dernière syllabe du groupe est accentuée. Bien qu'il s'agisse de non-mots, on pourrait avoir affaire ici à une telle manifestation. D'autre part, on peut penser, qu'outre le signal de début d'augmentation d'énergie, il est nécessaire de programmer un signal de fin (8) : dans les autres cas, l'excédant de force peut en quelque sorte être "absorbé" par les segments subséquents. Au moins en ce qui concerne les items accentués sur les syllabes 1 et 2, les différences de durée les plus fortes entre TRC et TRS sont obtenues pour le digastrique (abaisseur du maxillaire). On peut avancer que la programmation de l'activité mandibulaire est, sinon plus complexe que celle des lèvres, du moins plus liée à un modèle rythmique global : en effet, la mandibule peut être considérée comme le "pacemaker" pour la parole ; c'est elle qui impose l'organisation temporelle syllabique et plus généralement prosodique (14). Au contraire, l'activité labiale est ici étroitement liée à la phase subsegmentale d'occlusion. Les commandes des articulateurs sont donc fortement dépendantes de la dynamique de la production, et nos résultats font clairement ressortir la nécessité d'une analyse pluridimensionnelle des contrôles temporels (5, 9).

Tableau I : Influence de la situation sur les temps de réaction
* p<0.05 ; ** p<0.01 ; ***p<0.001 ; NS : Non Significatif

Syllabe accentuée	I		II		III		IV	
	TRC	TRS	TRC	TRS	TRC	TRS	TRC	TRS
Signal Acoustique	1103	1020*	1099	1016*	1105	1090	1113	1073
						NS		NS
DIG	1147	965***	1047	963**	1066	1044	1083	1024
						NS		NS
OOS	903	834*	875	814	875	877	896	878
				NS		NS		NS
OOI	934	844**	921	861	893	918	955	893
				NS		NS		NS

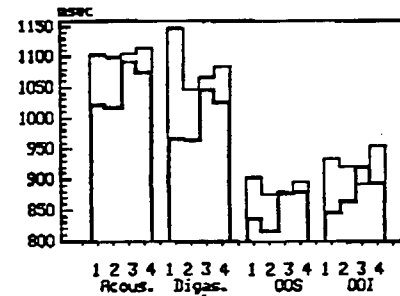


Fig. 1 : Influence de la place de l'accent sur le temps de réaction

— TRC ; - - - TRS.

5. REFERENCES

- [1] BONNOT, J.F., BOTHOREL, A., CHEVRIE, C. (1989), "De l'invariance relationnelle à la modélisation de la variabilité", TIPS 21, 265-302.
- [2] BONNOT, J.F., CHEVRIE, C., (1991), "Some effects of shouted and whispered conditions on temporal organisation", J. Phonetics (in press).
- [3] BONNOT, J.F., CHEVRIE, C., MATON, B., ARABIA, C., GREINER, G.F., (1986), "Coarticulation and motor encoding of labiality and nasality in CVCVCV nonsense words" Speech Comm., 5, 83-95.
- [4] BOYCE, S., KRAKOW, R., BELL-BERTI, F., GELFER, C.E. (1990), "Converging sources of evidence for dissecting articulatory movements into core gestures" J. Phonetics, 18, 173-188.
- [5] BROWMAN, C., GOLDSTEIN, L. (1990), "Tiers in articulatory phonology", Papers in Lab. Phonology Cambridge University Press, 341-376.
- [6] FUJIMURA, O. (1988) "Some remarks on consonant clusters", Ann. Bull. Rilp., 22, 59-66.
- [7] GARCIA-COLERA, A., SEMJEN, A. (1988), "Distributed planning of movement sequences", J. Motor Behavior, 20, 341-367.
- [8] IVRY, R. (1986), "Force and timing components of the motor program", J. Motor Behavior, 18, 449-474.
- [9] KELLER, E. (1989), "Predictors of subsyllabic duration in speech motor control", JASA, 85, 322-326.
- [10] KENT, R. (1983), "The segmental organization of speech", The Production of Speech, Berlin, Springer, 57-89.

- [11] PITT, M., SAMUEL, A. (1990), "The use of rhythm in attending to speech", J. of Exhp. Psychology, 16, 573-654.
- [12] REQUIN, J., LECAS, J.C., BONNET, M. (1984), "Some experimental evidence of a three-step model of motor preparation", Preparatory States and Processes, Hillsdall Erlbaum, 259-284.
- [13] STERNBERG, S., KNOLL, R., MONSELL, S., WRIGHT, C.E. (1988), "Motor programs and hierarchical organization in the control of rapid speech", Phonetica, 45, 175-197.
- [14] WORLEY, C. (1989), "Organisation temporelle articulatoire acoustique des gestes vocaliques et consonantiques", Thèse de Doctorat, Grenoble.

THE DISTINCTION OF CENTRAL AND PERIPHERAL SPEECH TIMING MECHANISMS

Eric Keller

Département de Linguistique, Université du Québec à Montréal,
Montréal, Qc. H3C 3P8, Canada

ABSTRACT

This study examines the multiple and conjoint prediction of speech timing events by central and more peripheral mechanisms. Phonemic ("central") distinctions showed greater predictive power for VOT segments, while rate ("more peripheral") distinctions showed greater predictive power for syllable intervals and vocalic durations. In patients with cerebellar disorders ("ataxic dysarthrics", patients suffering from a "relatively peripheral" motor disorder), the predictive power of speech rate was more strongly impaired than that of consonant distinction.

1. INTRODUCTION

Traditionally, phonetic science has considered variability to be a nuisance variable. This has been particularly so with respect to timing, where considerable variability is observed in intra- and inter-syllabic durations over repeated productions of the same utterance by the same or by different speakers.

However in line with most contemporary behavioral and social sciences, an alternative theoretical approach to variability is possible. In this view, variability is the result of the combined effects of a multiplicity of factors, some of which may be related to central speech processing, others to more peripheral motor processing and yet others to muscular execution (Figure 1).

In the domain of speech timing, a variety of potential predictors can be proposed for time segments measured at the periphery (e.g., in an acoustic wave-

form). On the one hand, lengthening or shortening effects can be due to linguistic factors, such as semantic emphasis, syntactic pauses, or phonemic distinctions. On the other hand, some timing effects are related to overall speech rate and to rhythmic variations.

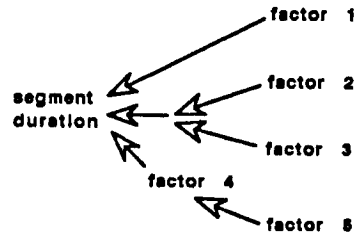


Figure 1. A theoretical approach to explaining variability in speech timing. Variability is seen as the outcome of the combined effects of a multiplicity of factors.

The present experiment examines the predictive interplay of two such factors. The first factor is phonemic distinctiveness (specifically, the ternary distinction between /p/, /t/, and /k/ in CV syllables). Since this distinction is relevant to lexemic distinction, it is considered to be representative of *linguistic control*.

The second factor is speech rate (e.g., normal or fast rate in a simple, repeated CV paradigm like /papapa/). Since many repeated motor actions such as walking and tapping can be produced at a faster or slower rate, this factor is considered to be representative of *general motor control*.

Normal speech probably involves concurrent processing at linguistic and general motor control levels. Therefore,

the two types of factors should exert a combined influence on durational segments in speech. In addition, the linguistic control factors should predict the greatest proportion of variance in those time segments that serve most directly in the acoustic distinction of syllables, such as VOTs. Conversely, general motor control factors should predict the greatest proportion of variance in other time segments in speech.

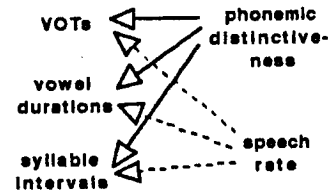


Figure 2. The predictive pattern examined in this study. Three phoneme categories (/p/, /t/, /k/ in CV syllables) and two speech rates (normal, fast) predict the duration of three speech periods (VOTs, vowel durations, syllable intervals).

A further test of this theoretical framework is possible. Patients suffering from neurological lesions affecting predominantly general motor control should show the greatest reduction in predictive power in speech rate. After all, if speech rate is indeed processed by a structure similar to that which controls the rate of production of other motor actions, an impairment affecting such a structure should have similar effects on speech motor control as on limb control.

2. METHOD

Seven dysarthric patients with cerebellar and/or ponto-cerebellar lesions (mostly diagnosed as Friedreich's ataxia) and six control subjects were asked to produce either /pa/, /ta/ or /ka/ stimuli repeatedly until the examiner held up his hand (minimum: 5 seconds). Tasks were performed at fast and conversational speech rates. Patients had been selected from a larger group of 13 patients for their particular severity of impairment.

Recordings were digitized at 10.4 kHz

with a 12-bit MacAdios Model 411 system. Time measures were taken at three points in the acoustic waveform (see Figure 3), and three speech segments were calculated from these measures. Points 1 and 2 are defined in traditional manner for VOT at the burst and at the onset of voicing. Point 3 is defined by the loss of vocalic oscillation, as judged against a noise threshold of the succeeding resting signal segment. Three representative durational measures derived from these observation points (VOT, vowel duration and syllable interval) were selected from an original 10 time measures.

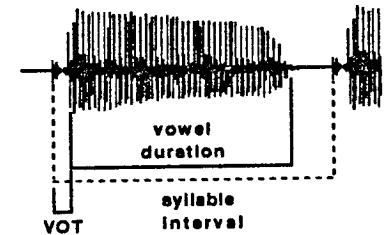


Figure 3. Duration measurements on the acoustic waveform. Out of ten durational measures performed within syllables and between adjoining syllables, VOTs, vowel durations and syllable intervals were retained as representative time segments.

In the subset of the data discussed here, there were 5,733 observations (out of an original 23,586 measurement points). Interjudgemental agreement on 2,880 re-measured pairs of measurement points was 98.6%.

Because of moderate to severe positive skewness, all measured time segments were log transformed, then z transformed, and measures exceeding ± 3.0 s.d. were eliminated (33 out of 23,586, or 0.13%). Subsequently the probability that data was not normally distributed was $< .05$.

Standard multiple regressions of the form:

$$\text{speech segment} = \text{phoneme category} + \text{speech rate category} + \text{constant}$$

were performed separately for each subject or patient, as well as for each measured speech segment (39 cells). There were an average of 147 observations per cell. Predictors were ternary-valued for the phoneme category [p/, /t/, /k/] and binary-valued for the speech rate category [normal, fast]. Degree of prediction was derived from the multiple regressions' averaged absolute beta coefficients for each type of predictive relationship.

3. RESULTS

The results of the regression analyses were in agreement with the hypotheses specified above.

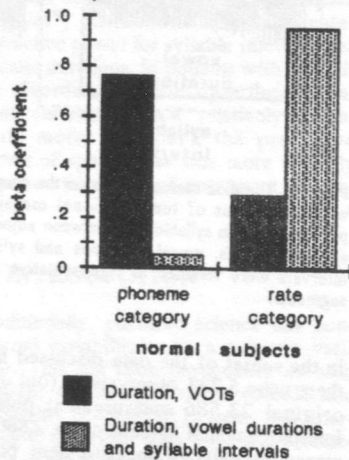


Figure 4. The prediction of time segments in six normal subjects. On the average, the ternary phoneme distinction showed the best prediction for VOT durations, while the binary speech rate distinction had the best prediction for vowel durations and syllable intervals. Results also indicate that time segments tended to be jointly predicted by phoneme and rate differentiations (illustrated in this figure for the prediction of VOT). Results for vowel durations and syllable intervals were similar throughout the study and were combined for presentation here.

(1) *The hypothesis of joint prediction.* Results for the normal subjects showed that time segments tended to be

jointly predicted by phoneme category and speech rate category (Figure 4). Although the predictive relations were weak in some cases (e.g., a beta of 0.057 for the relation "phoneme category predicts vowel duration/syllable interval"), the majority of the 400 original cells (77.3%) showed predictive relations significant at $p < .05$, and most (63.5%) were significant at $p < .001$, indicating that the two predictors tended to co-vary with all of the three time measures.

(2) *The hypothesis of distinct prediction.* The phoneme category had good predictive power for VOTs (beta 0.762), while rate category had excellent explanatory power for vowel durations and syllable intervals (beta 0.955, Figure 4). Crossed correlations ("speech rate predicts VOT" [beta 0.294], and "phoneme distinction predicts vowel durations and syllable intervals" [beta 0.057]) showed less predictive capacity.

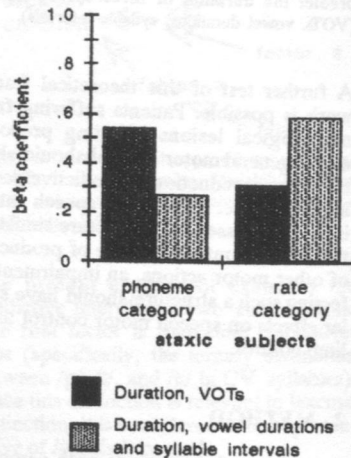


Figure 5. The prediction of time segments in seven patients with cerebellar disorders (patients with ataxic dysarthria). In comparison to the normal subjects, the (more peripheral) cerebellar disorder affected the predictive capacity of speech rate more strongly than the predictive capacity of consonant distinction. This offers some support for the notion that phonemic distinctions are part of a central programming mechanism, while speech rate is more directly related to a general motor programming mechanism.

(3) *The hypothesis of select impairment.* The prediction for patients with impairment of general motor control is also supported. It was expected that such patients would show a greater reduction of control over the relation "speech rate predicts vowel durations and syllable intervals", and a lesser impairment of the relation "phoneme category predicts VOT". Indeed, the first type of prediction was more diminished than the second (a reduction of 39%, beta 0.585 instead of beta 0.955, Figure 5). The relation "consonant distinction predicts VOT" showed a reduction of 29% (beta 0.541 instead of beta 0.762).

4. DISCUSSION

The present experiment illustrates a small fraction of the entire framework of predictive relations that is likely to characterize timing relations in speech.

The results support a view that considers phonetic variability at the periphery to be the predictive outcome of a multiplicity of factors, including linguistically relevant determiners like phonemic distinctiveness and general motor control determiners like speech rate. There is also some support for the notion of considering some of these factors, like the linguistic factors, to be more "central" and others, like the general motor control factors, to be more "peripheral" in nature.

The view supported by the present data thus contradicts earlier approaches to speech timing that attempted to view timing variations due to changes in rate and stressing as simple metrical variations of a basic temporal organization (the "proportional timing" hypothesis proposed predominantly by Kelso and colleagues, e.g. [2], see also [1] and [3]).

At the same time, the interesting, but somewhat limited results (39% against 29% reduction) from the pathological populations induce some caution. It is recalled that the present group of seven patients with presumed cerebellar and ponto-cerebellar lesions had been selected for the great severity of their impairment. Although these are patients whose cortical processing should not be affected by a direct lesion, their relation "phoneme

category predicts VOT" was also reduced (by 29%). And while their excessive timing variabilities and great difficulties in controlling limb movement betrays extensive cerebellar impairment, 11 of 14 cells illustrating the relation "rate category predicts vowel duration and syllable interval" were still significant at $p < .05$ (8 of 14 at $p < .001$). Lesions affecting a portion of the motor output system presumably interferes with the entire system, particularly the processing of events "upstream". At the same time, lesions presumably affecting a specific process rarely succeed in obliterating its entire functionality.

Finally, the study illustrates one of several interesting statistical techniques that can be used to explore the complete timing framework. Multiple regression and its more sophisticated outgrowth, path analysis, would seem to be the natural analysis techniques for a complex structure consisting of multiple predictor categories and a large number of predicted time measures in the speech utterance.

5. REFERENCES

- [1] KELLER, E. (1990). Speech motor timing. In W.J. Hardcastle and A. Marchal, *Speech Production and Speech Modelling* (pp. 343-364). Amsterdam: Kluwer.
- [2] KELSO, J.A.S., SALTZMAN, E.L., & TULLER, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, 14, 29-59.
- [3] MUNHALL, K.G. (1985). An examination of intra-articulatory relative timing. *Journal of the Acoustical Society of America*, 78, 1548-1553.

CONSTRAINTS ON THE BEHAVIOR OF THE TONGUE BODY: VOWELS AND ALVEOLAR STOP CONSONANTS

Simon D. Levy

University of Connecticut, Storrs, and Haskins Laboratories, New Haven, Connecticut

ABSTRACT

X-ray microbeam data from two male subjects were examined in order to test the hypothesis that words with identical vowels but different places of articulation (alveolar versus bilabial) contain the same underlying tongue body activity. Statistical analyses of the microbeam data failed to support this hypothesis. Comparing tongue body activity across consonantal contexts revealed that the alveolar contexts affected different vowels differently. In order to explain this behavior, a computational model was developed, based on robotic models for arm-reaching tasks. The model generated tongue tip and tongue body behavior that was qualitatively similar to the microbeam data.

1. INTRODUCTION

Recent phonological theory has relied heavily on the distinction between the articulatory role of the tongue tip and that of the tongue body (which is often called the tongue dorsum) in the production of speech. The tongue body is considered to carry the burden of articulation for vowels, and the tongue tip is considered to be the articulator primarily responsible for the production of coronal consonants [1], [3], [8]. This phonological dichotomy between tongue tip and tongue body has been paralleled in phonetic theory, including the recently developed task-dynamic model of speech production [9].

Despite the phonological distinction between the two articulators, phonetic investigation has shown a high degree of

correlation between the tongue tip and tongue body [5], [6], [7]. This result calls into question the independence of the articulators that is posited in phonology. The present study was done in order to determine the extent to which actual measurements of tongue tip and tongue body activity support the idea that these two articulators are independent of one another.

2. DATA

The data for this study were obtained at the University of Wisconsin's X-ray Microbeam facility. (I thank Dr. George Papçun of the Los Alamos National Laboratory for making these data available to me.) Two college-age male speakers of American English participated in the study. Gold pellets were placed as shown in Figure 1.

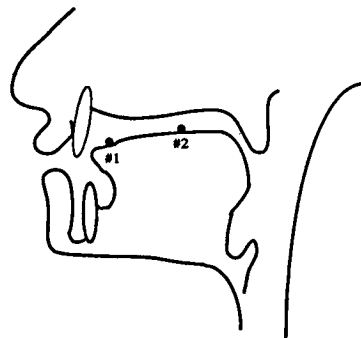


FIGURE 1. Placement of tongue pellets.

The distance from the subjects' tongue tip to pellet #1 (the tongue tip pellet) was

10 millimeters; the distance to pellet #2 (the tongue body pellet) was 35 millimeters.

Each subject was asked to produce three-syllable nonsense words of the form /CV1CəCV2C/. For a given word, all four C's were the same, taken from the set {p,t,b,d}. The two full vowels V1 and V2 were allowed to differ and were taken from the set {i,æ,a,u}. Primary stress was placed on the first syllable. This arrangement yielded words such as /bibəbib/ and /tatətit/. Words were produced in a pre-determined, random order. Visual examination of the microbeam data revealed that the vertical (Y) dimension of motion had a greater range of excursion than the horizontal (X) dimension for both the tongue tip and tongue body pellets. Therefore, only the Y dimension of these pellets' motions was considered for the study.

3. PROCEDURE, FIRST ANALYSIS

For each subject, tongue tip Y and tongue body Y values were extracted for one token of each nonsense word. Tokens were all of the same duration (approximately 1.02 seconds). The extracted tokens for thirteen alveolar consonant utterances (/dədəd/, /dədədid/, /didədə/, /didədid/, /dudədud/, /dədədad/, /didədud/, /dudədud/, /tatətit/, /titətat/, /titətut/, /tutətut/, /tutətut/) were then strung together, making a single large data set. This data set was used as input to the BMDP 6R program for partial correlation/multivariate regression. The tongue tip Y value was used as the independent variable and the tongue body Y as the dependent variable. Thus, the residuals of the partial correlation could be considered to represent the uncorrelated, or independent, behavior of the tongue body with respect to the tongue tip. It was hypothesized that the residual would be identical to the tongue body Y for the corresponding bilabial consonant utterance, in which there was no effect of an alveolar consonant.

4. RESULTS, FIRST ANALYSIS

The results of the first analysis did not support the hypothesis of an independent vocalic component of tongue body motion in the alveolar consonant

utterances. In general, the residual tongue body values were flat, indicating a constant offset of the tongue body with respect to the tongue tip. Deviations from this constant value were either much smaller than the corresponding tongue body displacement in the bilabial utterances, or did not coincide with the vocalic portion of those utterances. An example is given in Figure 2.

5. PROCEDURE, SECOND ANALYSIS

The first analysis suggested that the tongue body is strongly influenced by the tongue tip in the context of alveolar consonants. Therefore, it seemed reasonable to ask whether the alveolar consonants influence the tongue body uniformly across different vowel contexts.

To test the uniformity of the alveolar consonants' influence, it was assumed that the tongue body Y in the bilabial consonant words represented the purely vocalic behavior of the articulator. Therefore, subtracting these tongue body values from tongue body values in alveolar consonant words with the same vowel pattern would isolate the influence of the alveolar consonant. (Subtraction was used instead of a residuals technique because the first analysis suggested that the two techniques were computationally equivalent for the data examined.)

6. RESULTS, SECOND ANALYSIS

The results of the second analysis failed to support the notion that the alveolar context influences the tongue body identically across different vocalic contexts. The degree of excursion of the subtracted tongue body was much greater for low vowels than for high vowels, suggested a strong elevating effect of the alveolar context on the tongue body for low vowels. This result is illustrated in Figure 3.

7. MODELLING THE DATA

Both analyses suggested a high degree of influence of both consonants and vowels on the behavior of the tongue body. In developing a model of these influences, two distinct options presented themselves: first, a model in which the observed behavior is attributable to some unit larger than

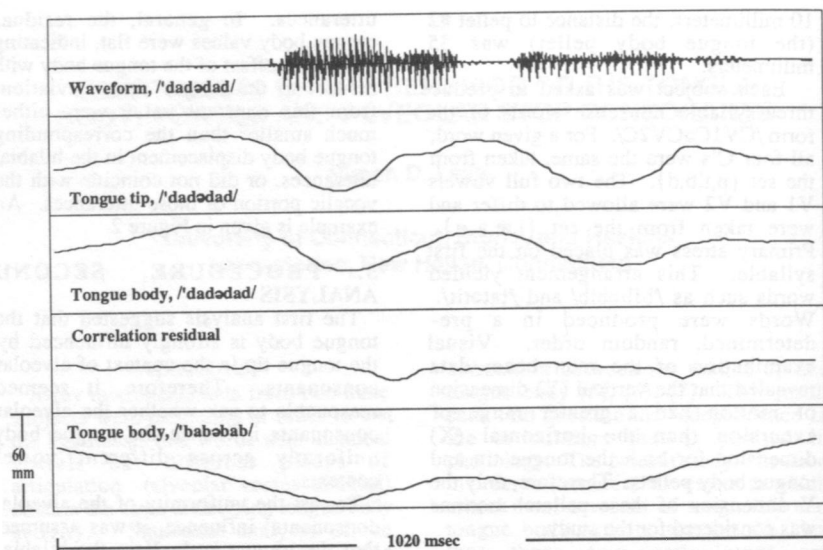


FIGURE 2. Sample results of first analysis. All articulatory channels represent vertical movement in the same scale and range.

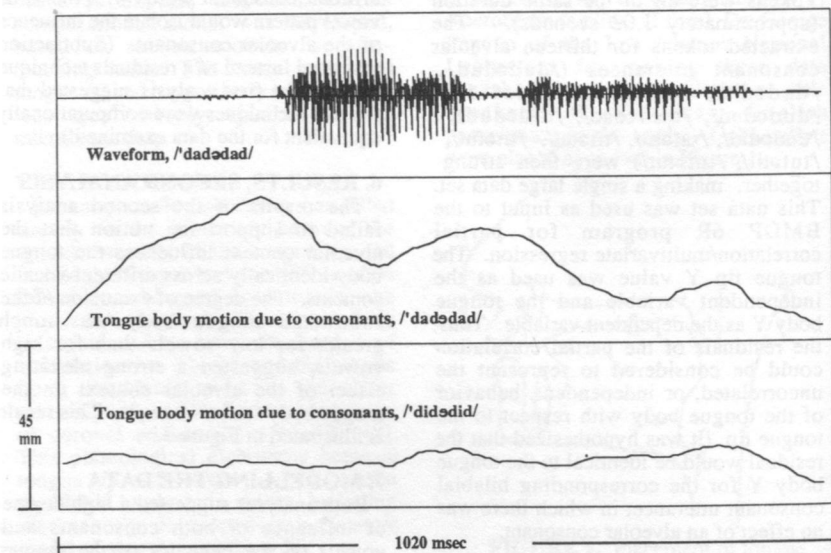


FIGURE 3. Sample results of second analysis. Both articulatory channels represent vertical movement in the same scale and range.

consonants or vowels (demissyllables, perhaps); second, a model in which the behavior results from physical constraints on the tongue's ability to achieve distinct consonantal and vocalic targets. (Öhman [7] developed a set "coarticulation functions" to generate tongue positions from separate consonantal and vocalic influences, but the physical interpretation of these functions is not clear.)

Models of this sort have been developed in the fields of robotics [4] and speech synthesis [2]. A simplified representation of the tongue based on these models is presented in Figure 4. In this model, the distance d between the tongue tip's current position and its target position (for an alveolar stop) is reduced iteratively by modifying the angles $a1$ and $a2$.

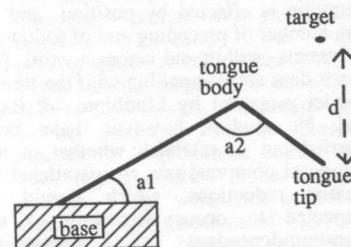


FIGURE 4. A task-based model of the tongue

Preliminary experiments with this model have revealed behavior that is qualitatively similar to certain aspects of the observed behavior of the tongue: The whole model tongue moves in order to support the achievement of alveolar closure, and the difference between the vertical position of the model tongue tip and that of the model tongue body increases as the tip nears its target.

8. CONCLUSIONS

The analyses described here suggest a high degree of interaction between the articulatory goals of the tongue tip and tongue body in the context of alveolar stop consonants. Although it is traditionally associated with the production of vowels (and dorso-velar consonants), the tongue body is strongly

constrained by the articulatory requirements of alveolar stops. Nevertheless, it may be possible to maintain the distinction between the articulators by using a model that takes account of the overall behavior of the tongue in achieving different articulatory goals.

9. ACKNOWLEDGMENTS

I thank Cathie Browman and Ignatius Mattingly for their help and support in this study. Preparation of this paper was supported in part by NIH Grant HD-01994 to Haskins Laboratories.

10. REFERENCES

- [1] BROWMAN, C.P. & L. GOLDSTEIN (1986), "Towards an articulatory phonology", *Phonology Yearbook*, 3, 219-252.
- [2] BROWMAN, C.P., L. GOLDSTEIN, E.L. SALTZMAN, & C. SMITH (1986), "GEST: A computational model for speech production using dynamically defined articulatory gestures," [Abstract], *Journal of the Acoustical Society of America*, 80 (Suppl. 1), S97.
- [3] CLEMENTS, G.N. (1985), "The geometry of phonological features," *Phonology Yearbook*, 2, 225-252.
- [4] HINTON, G. (1984), "Parallel computations for controlling an arm," *Journal of Motor Behavior*, 2, 171-194.
- [5] JOOS, M. (1948), "Acoustic phonetics," *Language*, 24, 2 (Suppl.)
- [6] LINDBLOM, B. (1983), "Economy of speech gestures," In P.F. MacNeilage (Ed.), *The Production of Speech*. New York: Springer Verlag.
- [7] ÖHMAN, S.E.G. (1967), "Numerical model of coarticulation," *Journal of the Acoustical Society of America*, 39, 151-168.
- [8] SAGEY, E. (1986), "The representation of features and relations in non-linear phonology," Unpublished doctoral dissertation, MIT.
- [9] SALTZMAN, E.L. & K. G. MUNHALL (1989), "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, 1(4), 333-382.

WORD- AND PHRASE-LEVEL ASPECTS OF VOWEL REDUCTION IN ITALIAN.

E. Farnetani and M. Vayra

Centro di Fonetica del CNR, Padova, Italy
Scuola Normale Superiore, Pisa, Italy

ABSTRACT

The present study investigates some of the sources of vowel reduction in Italian and the relationship between reduction and acoustic duration. The effects of stress, position within a word, and position within an utterance on the spatial and temporal characteristics of vowel /a/ were quantified and compared. The effects of position within an utterance (initial vs final) were investigated to verify the hypothesis of a progressive early-to-late reduction. The results indicate that phrase-level reduction is unsystematic, (it was observed only in one of our two subjects), and does not appear to be progressive (it is confined to the very edges of the utterance). The most relevant and systematic position effects are instead the lengthening and opening of unstressed vowels in utterance final position. Two findings emerge from the various patterns of interaction between duration and first formant frequency: a) speakers can control the two variables independently and such a control appears to be addressed to the preservation of stress contrast; b) the extent of such control seems to depend on the more or less elaborated speech style that characterizes different speakers.

1. INTRODUCTION

This study concerns articulatory and spectral reduction of vowels in Italian, the former defined as a decrease in the magnitude of gestural displacement, the latter as an increased amount of centralization within the acoustic vowel space.

Two issues are the focus of the present work: the analysis of possible sources of vowel reduction, and the assessment of the relationship between reduction and acoustic duration.

Evidence of spatiotemporal reduction of unstressed vowels in Italian has emerged from a number of studies ([1], [2]). A recent experiment on tongue dorsum and

jaw movements [3] showed that unstressed vowel /a/ is subject to a high degree of reduction, manifested as a decreased displacement of the jaw from the rest position (see also [4]). Other studies on Italian have shown that duration is affected by position and by the number of preceding and of following segments within and across a word [5]. Such data are compatible with the timing model proposed by Lindblom & Rapp [6]. No studies, however, have been carried out to establish whether or not temporal compressions are paralleled by spatial reductions, which should be expected to occur according to the duration-dependent undershoot hypothesis put for by Lindblom [7].

Results of a recent investigation [4] suggest that another source of reduction can be at play at phrase-level: a progressive early-to-late reduction, inferred from a monotonic decrease in first formant frequency and in the amplitude of jaw opening of stressed /a/ along an utterance. If this kind of phrase-level reduction occurs, to what extent will duration be compatible with a duration-dependent undershoot model, which predicts an increased degree of reduction in shorter segments? Investigations concerning the relationship between duration and reduction show that the two variables can be controlled independently. Engstrand [8] showed that in fast speech stressed vowels are not more reduced than their longer counterparts in slow speech. Similarly, Nord [9] showed that finally lengthened unstressed vowels tend to be more centralized than non-final stressed vowels of the same duration.

The present speech material has been constructed in such a way that all the

above mentioned potential sources of reduction can be investigated and their effects quantified and compared, while an assessment of the relationship between duration and reduction has made it possible to identify the conditions under which the two variables appear to be dissociated.

2. METHOD

The corpus consists of trisyllabic nonsense words of the type /CVCVCV/, where C = /t/ and V = /a/, or /i/, or /u/ in symmetric sequences, with stress on initial, medial and final position. The key words were repeated five times in three contexts: in isolation, in sentence initial position (*Ugo... della Torre parti' per la Francia*), and in sentence final position (*Parti' per la Francia col marchese Ugo ...*). Subjects were one female (S1) and one male (S2) Northern speakers of Standard Italian. We simultaneously collected acoustic and electro palatographic (EPG) data. We shall report here the subset of results relative to /a/. For low vowels, a decrease in first formant frequency and an increase in amount of linguopalatal contact in the back region indicate higher tongue/jaw position. They were used as indices of reduction. The durations of vowels were defined as the intervals of periodicity within each syllable; EPG and first formant values were measured at vowel mid points. It must be reminded that EPG provides only partial information on vowel configuration, thus, especially for vowel /a/ we have relied more heavily on F1 than on EPG. The following variables were tested: stress, syllable position within the word, position of the word within the utterance. Series of ANOVAs and t-tests were used for statistical analyses. In the description of the results, we shall refer to differences with a significance level no less than 97.50%.

3. RESULTS AND COMMENTS

3.1 Reduction

In both subjects stressed vowels have higher F1 and less linguo-palatal contact than unstressed vowels. For S1, F1 decreases by 28% in unstressed vowels (average values: 974 vs. 700 Hz), for S2 it decreases by 18% (average values: 669 vs. 548 Hz); for S1 the EPG contact area decreases during stressed vowels by 80% (average contact: 4.39 vs 0.89); for

S2 it decreases by 53% (average contact: 5.68 vs 2.67).

Table 1 shows the mean F1 values.

TABLE 1.

Mean values of F1 (Hz). W1, W2, W3 refer to words with stress on the first, on the second and on the third syllable. Numbers 1, 2, and 3 refer to the context in which the test word was produced: isolated, sentence initial, and sentence final, respectively.

CONTEXT	1	2	3
S1			
W1	980 666 855	990 661 526	983 619 820
W2	850 983 882	554 988 514	626 1000 873
W3	956 711 959	559 644 962	631 656 920
S2			
W1	654 554 654	685 522 479	634 524 591
W2	564 739 655	459 606 528	496 654 617
W3	566 539 738	504 530 690	479 536 674

Globally both stressed and unstressed vowels are affected by position within word or sentence: the higher F1 of unstressed vowels in final position (contexts 1 and 3) indicates that they tend to open, while stressed vowels in final position tend to be more reduced than non-final vowels in S1 (contexts 2 and 3, although only in context 3 the difference reaches the significance level $p < 0.02$), whilst they do not change significantly in S2.

As for the stressed vowels produced by S1, the fact that only those in utterance final position show some significant differences with earlier vowels, and, as shown in Table 1, tend to be more reduced than any other stressed vowels earlier in the sentence, indicates that such reduction is a phrase-level rather than a word-level phenomenon; at the same time it suggests that weakening of stressed vowels is confined to the very last syllable of the utterance. Thus, the present data are only in partial agreement with the hypothesis of a monotonic early-to-late reduction. As for unstressed vowels, the fact they do not tend to open in final position of utterance-initial words, which are not phrase final, suggests that also the decrease in reduction of unstressed vowels in final position is to be regarded as a phrase-level phenomenon. This is corroborated by another observation relative to S1: in isolated words also the unstressed vowels

of initial syllables tend to open (see Table 1), and only in this context the initial syllables of the word are also in phrase- (and utterance-) initial position. Thus, the patterns of reduction for unstressed vowels can be interpreted in very simple terms as tendencies to alterate the degree of vowel height at prosodic boundaries, or, better, in absolute final position for both speakers, and also in absolute initial position for S1. All the other unstressed vowels appear to be equally reduced whatever is their position.

The EPG data do not show the trend observed in F1 for the stressed vowels, but capture the two degrees of reduction observed for unstressed vowels: for both speakers the amount of EPG contact in vowels adjacent to prosodic boundaries is about 23% less than the contact in the more reduced vowels.

3.2 Duration and reduction

The global data show, as expected, that duration and F1 are highly correlated. The correlation coefficients are $r(133) = 0.783$ for S1, and $r(130) = 0.774$ for S2. Duration is also correlated with the amount of EPG contact with $r(130) = -0.704$ for S1 and $r(130) = -0.654$ for S2. We shall examine in detail the relationship between duration and F1, since duration accounts for a larger proportion of variance of F1 than of EPG for both subjects.

TABLE 2.

Mean vowel durations (ms). Captions as in Tab.1

CONTEXT	1	2	3
S1			
W1	140 59 134	127 54 58	131 57 136
W2	68 157 142	55 126 47	43 150 106
W3	65 59 188	52 61 116	60 53 127
S2			
W1	103 60 114	98 56 71	100 58 68
W2	62 142 116	41 109 41	49 117 71
W3	61 53 144	53 64 95	53 65 102

The table shows, as expected, longer durations for stressed than for unstressed vowels. The comparison with Table 1, indicates that the differences in duration between stressed and unstressed vowels are much more often neutralized by position effects than the differences in

F1, and that the relations between duration and F1 differ in the two subjects. For S1, in isolated words, we observe that unstressed vowels undergo final lengthening, which, as seen before, is accompanied by an increase in F1; instead the unstressed vowels in the initial syllables are not significantly longer than medial vowels, in spite of their remarkably high F1 values (cf. Table 1). A second important discrepancy between duration and F1 is observed in the utterance final unstressed vowels: final lengthening makes them as long as the preceding stressed vowels (see Table 2, Context3, W1) or as long as stressed vowels occupying the same position (see Context3 W1 vs W3). They have, however, significantly lower F1 values than the stressed vowels of the same duration even though such values are higher than those of the shorter unstressed vowels (cf. Table 1). Taken together the data indicate that the speaker has allowed the unstressed final vowels to lengthen and to open, but has prevented them from opening as much as the stressed vowels of the same duration. The dissociation between duration and F1 in the first syllables of isolated words could instead be viewed as an effect of a greater articulatory force characterizing the very beginning of an utterance. As for stressed vowels, their variations in duration are not reflected in F1, which appears to be rather stable: only in utterance final position the shortening of the stressed vowel is associated with a decrease in F1. This suggests that the speaker could vary stressed vowel duration without varying their height.

In Figure 1 the mean values of F1 are plotted against the mean durations; the figure also shows the regression line obtained from the row data.

As for S1, (on the left), the figure shows that stressed an unstressed vowels are always distinguished along one or the other dimension. Stressed vowels vary in duration much more than in F1. The unstressed vowels tend to fall into three groups: short (utterance-initial) unreduced vowels, long (utterance-final) moderately reduced vowels, and short highly reduced vowels (those not adjacent to boundaries). Altogether the data indicate that S1 exerted independent control over duration and gestural

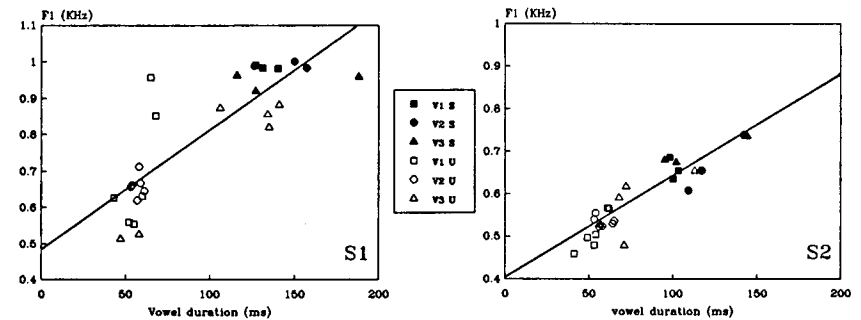


FIGURE 1. F1 plotted against duration; 1, 2, 3: V position within word. coefficients of variation.

amplitude in both stressed and unstressed vowels. As for S2, in isolated words the duration patterns reflect quite well the F1 patterns, with the consequence that unstressed final vowels, which undergo a remarkable final lengthening (and an increase in F1), are no longer distinguished from stressed vowels in initial position (cf. Figure 1 and Tables). In embedded words the very short durations of unstressed vowels parallel their very low F1 values. Instead, stressed vowels, although characterized by longer durations in word medial positions, have F1 frequencies tendentially lower than those of vowels in initial and final positions. This indicates that S2 varied the movement velocity in producing stressed vowels in different word position. The data suggest that S2 exerted an independent control over duration and gestural amplitude mostly in the production of stressed vowels, thus, much less extensively and systematically than S1.

5. CONCLUSION

The overall F1 and EPG data on vowel /a/ indicate that 1) stress seems to be the major factor in vowel reduction; 2) adjacency to prosodic boundary has more systematic influence on unstressed than on stressed vowels; 3) in S1 there is clear evidence of reduction of stressed vowels in utterance final position. As for the intersubject differences in the interactions between F1 and duration, we ascribe the more extensive independent control exhibited by S1 to her more elaborated speech style. Evidence that S2 used a more casual speech style than S1 emerges from the combination of his lower F1 values, larger areas of EPG contact, shorter durations, and higher

The finding that this speaker exerted independent control over duration and F1 less extensively than S1 fits quite well in the global speech style picture we are proposing. The comparison between S1 and S2 data in Fig.1 bears witness to our interpretation: in S2 we note, together with reduced distances between the two stress categories, less extensive deviations from the regression line and more data points below than above it.

6. REFERENCES

- [1] Farnetani, E. & S. Kori (1981), "Italian lexical stress in connected speech". In *Proc. 4th F.A.S.E. Symposium, 1*, Roma: Edizioni Scientifiche Associate, 57-61.
- [2] Farnetani, E. & A. Faber (1991), "Vowel production in isolated words and in connected speech: An investigation of the linguo-mandibular subsystem", in press.
- [3] Lindblom, B. (1963), "A spectrographic study of vowel reduction", *J.A.S.A.*, 35, 1773-1785.
- [4] Lindblom, B. & K. Rapp (1973), "Some temporal regularities of spoken Swedish". *Publ. no. 21, University of Stockholm: Institute of Linguistics*.
- [5] Nord, L. (1986), "Acoustic studies of vowel reduction in Swedish", *STL-QPRS*, 4, 19-36.
- [6] Vayra, M., C. Avesani & C. Fowler (1984), "Patterns of temporal compression in spoken Italian". In *Proc. Xth ICPhS*, Dordrecht, Holland: Foris Publ., 541-546.
- [7] Vayra, M. & C. Fowler (1987), "The word-level interplay of stress, coarticulation, vowel height and vowel position in Italian". *Proc. XIth ICPhS*, 4, Tallinn, Estonia: Academy of Sciences of the Estonian S.S.R., 24-27.
- [8] Vayra, M. & C. Fowler, "Declination of supralaryngeal gestures in spoken Italian", submitted.
- [9] Engstrand, O. (1988), "Articulatory correlates of stress and speaking rate in Swedish VCV utterances", *J.A.S.A.*, 83, 5, 1863-1875.

PRODUCTION DES VOYELLES ET MODELE A REGIONS DISTINCTIVES

René Carré^o and Mohamed Mrayati^{oo}

^o)Département Signal, Unité de Recherche Associée au CNRS
ENST, 46 rue Barrault, 75634 Paris cedex 13

^{oo})Scientific Studies and Research Center, POB 4470, Damascus.

RESUME

On rappelle les propriétés acoustiques optimales d'un tube acoustique divisé en régions distinctives. Pour maîtriser la commande d'un, puis de deux puis de trois formants, on définit deux puis quatre puis huit régions distinctives. On exploite un tel modèle optimal pour produire les sons acoustiquement les plus différents. La production des voyelles est ensuite étudiée à la lumière de ce modèle théorique.

1. LE MODELE EN REGIONS DISTINCTIVES

1.1. Tube fermé à l'une des extrémités, ouvert à l'autre.

On a montré [1], [2], [6], [7] que le tube acoustique fermé à l'une de ses extrémités et ouvert à l'autre pouvait être divisé en régions distinctives permettant des modulations optimales des fréquences de résonance de ce tube. La longueur de ces régions est fixe et correspond à un pourcentage de la longueur totale effective du tube. Les commandes des sections des régions de cette modélisation sont donc transversales. Par modulations optimales, on entend : plages maximales possibles de variations de ces fréquences de résonance, et variations maximales des fréquences pour un déplacement minimal des parois latérales des régions autour de la position neutre. Par ailleurs, on peut souhaiter commander soit le premier formant (le modèle doit être constitué de seulement deux régions R1 et R2), soit les deux premiers formants (on a alors quatre régions, R1, ..., R4), soit des trois premiers formants (on a alors huit régions) (fig. 1), etc.

Le comportement de ce modèle est pseudo-orthogonal c'est-à-dire que, autour de la position neutre, les commandes des sections des différentes régions correspondent à toutes les

combinaisons de sens de variations des formants pris en considération (fig. 1). Ce modèle permet donc la maîtrise des sens de variations des formants.

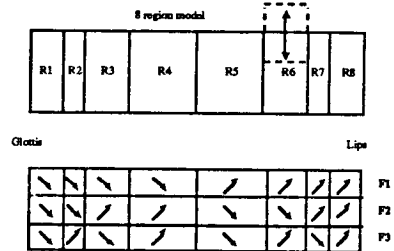


Figure 1. Modèle constitué de huit régions. On note la commande de type transversale et les sens de variations des trois premiers formants pour un accroissement de la section de l'une des régions autour de sa position neutre.

On peut aussi noter que la position de la commande demande de plus en plus de précision selon le nombre de formants considéré : pour le premier formant, les longueurs des deux régions sont de 1/2 si l est la longueur effective du tube ; elles sont de 1/6, 1/3, 1/3 et 1/6 pour prendre en compte les deux premiers formants, etc.

Par ailleurs, le comportement antisymétrique mis en évidence figure 1 peut être exploité pour multiplier les effets de modulation : il suffit de mettre en oeuvre une commande synergétique, c'est-à-dire, par exemple, qu'une commande de réduction de section d'une région de la partie avant du tube va être associée à une augmentation de la section de la région correspondante de la partie arrière. En mettant en oeuvre cette propriété, le nombre de commandes des régions du modèle est divisé par deux. Dans le cas

d'un modèle à deux régions permettant le contrôle du premier formant, un seul paramètre de commande suffit. La figure 2 montre l'évolution du premier formant (et des 2ème et 3ème formants). Dans ce cas, le produit de la section (SR1) de la région arrière (R1) par la section (SR2) de la région avant (R2) est constant. La longueur effective du tube est de 19cm et les différents types de pertes sont pris en compte dans la simulation. On note trois zones principales : deux zones de stabilité et une zone de variation rapide (autour de la position neutre qui est ici de 4cm²). Les zones de stabilité peuvent, en particulier, être utilisées pour coder l'information acoustique. Le passage rapide d'un état à un autre permet un débit maximal d'informations par seconde. La prise en compte du 2ème formant, réalisée dans un modèle à 4 régions, multiplie le nombre de niveaux de codage. Dans ce type de codage à modulation des fréquences de résonance du tube acoustique, on ne prend pas en compte les types de sources d'excitation qui multiplieraient encore les possibilités de produire des informations acoustiquement différentes.

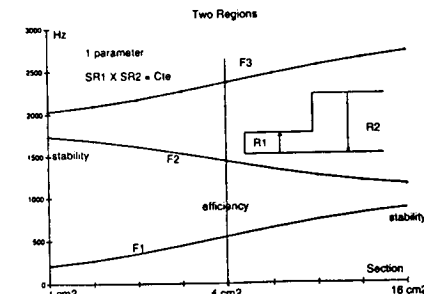


Figure 2. Evolution des trois premiers formants avec un modèle constitué de deux régions. Les sections varient de 0.5 à 32cm² par pas logarithmiques. La position neutre est de 4cm².

1.2. Tube fermé aux deux extrémités.

Selon les principes énoncés dans [7], on peut aussi définir des régions pour un tube acoustique fermé aux deux extrémités. Le premier formant corres-

pond alors à la fréquence de vibration des parois de ce tube, et le deuxième formant peut être contrôlé par un modèle à trois régions. Le comportement du modèle est symétrique et la constriction s'applique au milieu du tube. Ce modèle permet d'atteindre les valeurs les plus basses du couple F1, F2. Mais l'aspect pseudo-orthogonal mis en évidence dans le cas précédent disparaît ici.

Avec ces deux modèles de référence, on peut produire efficacement les sons les plus différents en ce qui concerne les fréquences de formants, tout en faisant appel à un minimum de paramètres de commande. On peut aussi jouer sur le nombre de références retenues. Il faut rappeler que cette modélisation implique intrinsèquement des commandes transversales à des endroits quantifiés du tube acoustique (les endroits les plus efficaces pour la modulation). La question qui se pose maintenant est de savoir si, pour parler, l'homme exploite les propriétés importantes du tube acoustique mises en évidence dans le modèle à régions.

2. LE MODELE A REGIONS DISTINCTIVES ET L'APPAREIL VOCAL.

Tout d'abord, notons que l'appareil vocal est bien adapté pour exploiter les caractéristiques d'un modèle à régions distinctives. C'est un conduit fermé à l'une des extrémités (coté glotte) et ouvert de l'autre (coté lèvres). Dans le cas d'une référence à huit régions, R8 serait contrôlé par les lèvres, R7 par la pointe de la langue, R3, R4, R5, R6 par le corps de la langue laquelle réalise par ailleurs la mise en oeuvre de la synergie grâce à son volume constant. En revanche, R1 et R2 sont relativement constants (larynx). Le modèle fermé-fermé peut être réalisé par une forte labialisation et par une constriction centrale obtenue par le corps de la langue.

En dynamique, le problème de la commande de l'appareil vocal chez l'homme se pose : est-elle transversale et réalise-t-elle des contractions à des endroits spécifiques lors de la production de la parole ? La bonne reproduction de la transition /ai/ par le modèle [2] par exemple est un élément de réponse positive. Un déplacement longitudinal de la constriction donnerait des résultats tout à fait différents.

3. LE MODELE A REGIONS DISTINCTIVES ET LA PRODUCTION DES VOYELLES.

Les hypothèses formulées dans ce paragraphe demandent vérifications et études approfondies. En procédant du plus simple au plus compliqué, en privilégiant une voyelle non labialisée (plus intense) à une voyelle labialisée, en privilégiant la distinction par le premier formant (la position de la commande demande la précision la plus faible) et en tenant compte du fait que l'avant de la langue est plus souple que l'arrière, on peut expliquer le contenu de systèmes vocaliques. Les voyelles extrêmes /a/ et /ɨ/ sont les plus simples à produire dans le cas de la référence fermée-ouverte. Ensuite, la mise en oeuvre de la référence fermée-fermée (avec constriction centrale) permet la production de la voyelle /u/ (F1

et F2 les plus bas) Puis, on retient la position intermédiaire de la transition /ai/, c'est à dire la voyelle /e/. A partir de la configuration de la voyelle /a/, la labialisation entraîne la production de la voyelle /ɔ/. On retrouve ici le système vocalique de 5 voyelles le plus répandu [5]

On a reproduit figure 3, les trajectoires formantiques obtenues par un modèle à quatre régions. R1 est fixé à 1.4cm², R2 et R3 sont commandées synergétiquement. Les sections varient entre 0.7 et 11cm². On a placé différentes voyelles dans le plan F1-F2. A titre indicatif, on a aussi représenté la trajectoire /u/ obtenue par ouverture aux lèvres d'un modèle fermé-fermé avec constriction centrale. Les voyelles /o/ et /ɔ/ et leurs correspondants non labialisés pourraient être obtenue de la même manière.

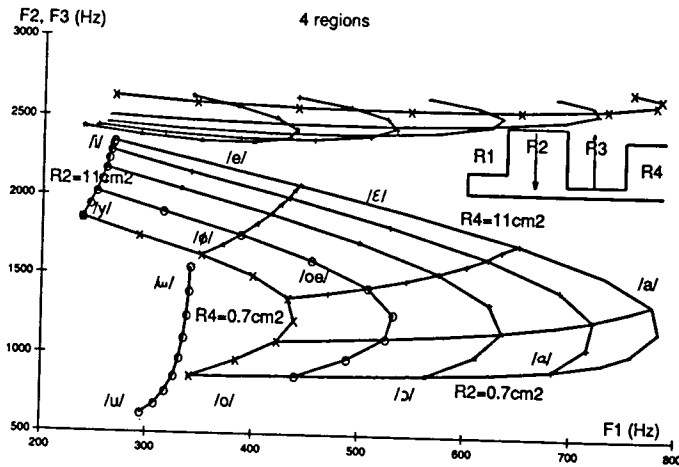


Figure 3. Position de voyelles dans le plan F2, F3(F1) et trajectoires formantiques obtenues avec un modèle constitué de quatre régions distinctives. On a aussi représenté la trajectoire /u/ obtenue avec un modèle fermé-fermé à trois régions.

La prise en considération du troisième formant (pour distinguer les voyelles /i/ et /y/ qui, dans certains cas, ont même deuxième formant par exemple) peut être obtenue avec un modèle en huit régions et commande synergétique de régions symétriques R3 et R6 par exemple (dans ce cas, la plage du deuxième formant est réduite et permet de réaliser les voyelles centrales) ou bien par synergie avant ou bien arrière (R3 et R4 par exemple) ou

bien par une légère asymétrisation de la commande d'un modèle à quatre régions. Un déplacement de 2 cm (figure 4) vers l'avant de l'axe de symétrie permet de stabiliser le deuxième formant lors de la transition /iy/ alors que la plage de variation du troisième formant augmente.

5. DISCUSSIONS

On a montré que l'on pouvait simplement produire les voyelles avec un mini-

num de paramètres de commande. Une stratégie de commande transversale paraît réaliste : elle permet de réaliser simplement des trajectoires que l'on retrouve en

parole naturelle. On peut alors émettre l'hypothèse que l'homme exploite les caractéristiques acoustiques optimales d'un conduit vocal qui serait divisé en régions.

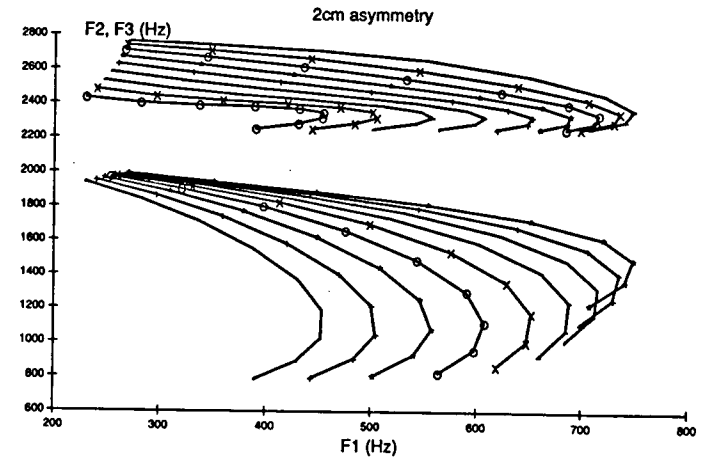


Figure 4. Trajectoires formantiques obtenues avec un modèle à quatre régions pour un déplacement de 2cm de l'axe de symétrie.

Cette interprétation diffère de celle émise par Lindblom [4] qui propose le système de perception comme référence. Dans la logique de l'hypothèse que nous formulons, le lexique serait constitué d'intentions de type réalisation d'une constriction à tel ou tel endroit quantifié du conduit vocal, associées à une intention de type labialisation ou non labialisation. La réalisation des intentions serait obtenue, au niveau périphérique, par une structuration des paramètres articulatoires pour réaliser les gestes phonétiques souhaités. L'objectif du système de perception serait alors de reconnaître les intentions initiales [3]. Le pilotage des systèmes de production et de perception de la parole chez l'homme par des lois physiques n'est pas irréaliste. Si cette hypothèse se vérifie, les conséquences seraient nombreuses ne serait-ce que pour proposer des schémas de l'évolution de l'appareil vocal humain au cours des millénaires.

REFERENCES

[1] Carré, R and Mrayati, M. (1990), "Articulatory-acoustic-phonetic relations and modeling, regions and modes", in

Speech Production and Speech Modelling (W.J Hardcastle and A. Marchal, editors), NATO ASI Series, Kluwer Academic Publishers.

[2] Carré, R. and Mrayati, M. (1991), "Vowel-vowel trajectories and region modeling", Second seminar on speech production: models and data (Leeds, 1990), to appear in The Journal of Phonetics.

[3] Fowler, C.A. and Rosenblum, L.L. (1989), "The perception of phonetic gestures", Haskins Labs., SR-99/100, 102-117.

[4] Lindblom, B. (1986), "Phonetic Universals in Vowel Systems", in Experimental Phonology, J.J. Ohala and J.J. Jaeger ed., Academic Press.

[5] Maddison, I. (1984), "Patterns of Sounds", Cambridge University Press, Cambridge.

[6] Mrayati, M., Carré, R. and Guérin, B. (1988), "Distinctive regions and modes: a new theory of speech production", Speech Communication, 7, 257-286.

[7] Mrayati, M., Carré, R. and Guérin, B. (1990), "Distinctive regions and modes: articulatory-acoustic-phonetic aspects", Speech Communication, 9, 231-238.

CORRECTING ERRONEOUS INFORMATION IN SPONTANEOUS SPEECH:
CUES FOR A.S.R.

P. Howell

University College London, England

ABSTRACT

Previous research has shown that when speakers make a mistake and repair it, they signal that the forward flow of speech has been altered by insertion of a pause, and highlight what has been altered by adding stress on the first word of the alteration. The question of whether these prosodic patterns are specific to repairs or whether they are shared by related grammatical structures was not addressed. In this contribution, the prosody of phrase and word repetitions and conjunctions, taken from a corpus of unrestricted speech, were analyzed. These constructions were chosen because they have certain similarities with repairs and the prosodic analysis procedure can be applied to them. Repetitions behave like repairs whereas conjunctions are dissimilar. It is concluded that the prosody of repairs is a reliable indication that errors are being corrected, and the implications for A.S.R. are discussed.

1. INTRODUCTION

When we listen to speech it is not usually difficult to determine speakers' intentions, even when they make alterations or start a sentence, break off, and recommence. These processes are termed "speech repair" [1]. An example of a repair is "Go left at the, I mean, go right at

the crossroads". There are typically three identifiable parts in a repair - the original utterance (OU), the editing phase and the repair proper. In the example, "Go left at the" is the OU, and "go right at the" is the repair. The OU contains the word or words to be repaired, termed the reparandum, ("left" here). The speaker has gone past the erroneous word and so the repair is said to have an overshoot. The editing phase here is the phrase "I mean". Levelt counts "er" as an editing term, but it is considered here as a form of pause. With "er" excluded, the editing phase is rare in our corpus of repairs from unrestricted speech (4.2%) and is not discussed further. The final phase is the repair which includes the alteration (here the word "right"). Note also that the speaker has backed up to a point prior to where s/he wants to make the alteration and, so, the repair contains a retrace.

Levelt [1] has identified several categories of repair, but attention here is restricted to repairs in which erroneous information has been altered (termed error repairs). Though these may be sub-divided into repairs which occur on different grammatical units, the analysis reported here applies to all categories of error repair so the sub-categories are not discussed

further. The only obligatory parts of error repairs are the alteration and reparandum.

Automatic speech recognition (A.S.R.) systems for timetable or directory enquiries will have to be able to identify when an error has been made in voiced input and what information is being altered. Incorrect recognition could result in erroneous information being generated by the system in response to an enquiry. Though analysis of speech can indicate what speakers do, it is important that these be assessed perceptually to check that listeners use these cues. Clearly information about how human listeners are able to understand (1) that a repair is being made (the forward flow of speech has been stopped) and (2) what erroneous information is being altered may have important implications for A.S.R.

2. PROSODY IN SPEECH REPAIRS

Prosody helps listeners both detect that the forward flow of speech has stopped and locate where the altered information starts [2]. Prosody refers to changes in timing, loudness and pitch movements over groups of segments. The two aspects of prosody that are known to signal information about repairs are pauses and stress. Pauses, as noted previously, include filled pauses as well as periods of silence. Stressed syllables tend to be longer and louder than their unstressed counterparts. Primary and secondary stress are marked in the transcriptions, primary stress indicating a higher level than secondary stress.

Howell and Young [2] analyzed a corpus of 272 repairs drawn from the Survey of English usage (SEU) [3]. This corpus is of unrestricted speech and has

pauses and stresses transcribed. The speech was parsed into the retraced section and the corresponding section that occurred prior to it. There was a marked tendency for sections of pauses to be added before the first word of the retraced section when such sections occurred but no systematic tendency to increase stress on the first word of the retrace compared with its first occurrence. This shows that pauses are used to mark the interruption to the forward flow of speech. The highest degree of stress that occurred on any syllable of the first word of the alteration was compared with this same measure on the reparandum. This analysis showed that stress was added on the first word of the alteration, and this was interpreted as showing that stress is used to highlight the altered information. There was no tendency to add pauses before the alteration in comparison with the reparandum except when no retrace occurred. In the latter cases, the alteration starts immediately after the forward flow of speech has been interrupted, and is consistent with pauses being used to mark such locations.

3. ASSOCIATION OF PROSODIC PATTERNS WITH REPAIR

No data has yet been provided concerning whether the prosodic patterns described are specific to repairs or whether they are shared with related structures. In this presentation two structural types which are closely related to repairs are analyzed. These two types are repetitions and conjunctions. Repetitions include word and phrase repetitions, and the repeated sections can be analyzed in the same way as retraces. A total of 364 word and 168 phrase repetitions were located in the SEU and the results of this

analysis are shown in Table I.

Table I. Analysis of pauses and stresses in repetitions

		Word repetitions	Phrase repetitions
Pauses	Added	107	56
	Dropped	21	4
	N	364	168
		sig.	sig.
Stresses	Added	19	12
	Dropped	13	15
	N	364	168
		ns.	ns.

Basically, the table shows that speakers introduce pauses before the first word of a repetition but there is no increase in stress. This parallels the findings with retraced sections of repairs and offers some support for terming these "covert repairs" [1].

The inclusion of conjunctions in the analysis depends upon a rule described by Levelt for ascertaining whether a repair is well-formed or not (WFR). Levelt's WFR suggests that the two main parts of a repair (original utterance and repair proper) are related in the same manner as the two constituents of a coordination. According to Levelt "An original utterance plus repair (OR) is well formed if and only if there is a string C such that the string (OC or R) is well formed, where C is a completion of the constituent directly dominating the last element of O (or is to be deleted if that last element is a connective such as "or" or "and")." [1, p.486]. Thus, in "There you can park at the left-hand side of the, the right-hand side of the road," the original utterance is "park at the left-hand side of the". The VP "park at the left-hand side of the" can be completed with "road" (=C).

The repair (R) is "park at the right-hand side of the road". The coordination thus becomes: "There you can park at the left-hand side of the road or park at the right-hand side of the road."

Given a co-ordination, this process can be reversed. Thus the co-ordination "a comparative graphology paper or a historical graphology paper" from the SEU could have been the repair "a comparative graphology, a historical graphology paper". Levelt has pointed to the syntactic relationship between co-ordinations and repairs, so it might validly be asked whether this extends as far as the two structures having similar prosodic properties. Using Levelt's WFR, the retrace in a co-ordination starts after the conjunction and the first non-matching word constitutes the alteration ("historical" in the example, which is to be compared with "comparative").

A total of 244 conjunctions were collected from the SEU. These were single words or phrases which were joined by any conjunction. The constraints placed upon conjunction selection was that the word or phrase before the conjunction started at a constituent boundary and that there was a constituent boundary

after the word or phrase after the conjunction. Also, the grammatical category of the word or phrase before the conjunction had to match in type with the grammatical class of the word after the conjunction. No other constraints were applied. As illustrated in the example, this generated material which had sections with retraces and "reparandum/alteration" equivalent pairs and analysis proceeded as described for the repairs. The data are summarised in Table II.

have a retrace, there is a significant tendency to add pauses, as with repairs, but no significant tendency to add stresses, unlike with repairs. Thus, it appears that the different types of constructions, though syntactically related are prosodically dissimilar. The prosody in repairs is dissimilar to that in related grammatical structures such as around conjunctions.

Table II. Analysis of prosodic factors around conjunctions

a) Pauses and stresses on the first word of the "reparandum/alteration" equivalent

		Had no retrace	Had retrace
Pauses	Added	10	2
	Dropped	2	4
	N	198	46
		sig.	ns.
Stresses	Added	61	20
	Dropped	50	4
	N	198	46
		ns.	sig.

b) Pauses and stresses on retraced sections

Pauses	Added	1
	Dropped	2
	N	46
		ns.
Stresses	Added	1
	Dropped	4
	N	46
		ns.

REFERENCES

- [1] LEVELT, W.J.M. (1983), "Monitoring and self-repair in speech", *Cognition*, 14, 41-104.
 [2] HOWELL, P. & YOUNG, K., "The use of prosody in highlighting alterations in repairs from unrestricted speech", *Quarterly Journal of Experimental Psychology*, in press.
 [3] SVARTVIK, J. & QUIRK, R. (1980), "A corpus of English conversation", Lund: Gleerup.
- The prosody around conjunctions differs from that around repairs. For the conjunctions that had a retrace, there is a significant tendency to stress the "alteration", as with repairs, but no significant tendency to add pauses prior to the "retrace", unlike with repairs. For the conjunctions that did not

PROSODIC EFFECTS ON ARTICULATORY GESTURES -- A MODEL OF TEMPORAL ORGANIZATION

Osamu Fujimura, Donna Erickson and
Reiner Wilhelms

The Ohio State University
Division of Speech and Hearing Science
Columbus, OH 43210-1002, U. S. A.

ABSTRACT

A theory of phonetic implementation based on articulatory gestures and their temporal organization is proposed. It is compatible with Ohman's early insight (consonantal perturbation), which in effect assumes a separate tier for vowel to vowel movement as the base, and consonantal gestures superimposed on this base. The segmental constituent units are syllables, each of which is specified by demissyllabic feature values. A generative description is given as a series of computational modules: a converter, a distributor, a concurrent set of actuators, and a signal generator. Implications regarding various conditions of prosodic control are suggested.

1. CONVERTER

A computational procedure is shown in Fig. 1. The converter, as the first component of our model of phonetic implementation, "converts" an abstract phonological representation to an annotated linear pulse train. The converter maps phonological feature specifications for each demissyllable into a set of articulatory gestures. In Fig. 1, τ represents a stop gesture, σ fricative, θ interdental, η glide, N nasal, λ lateral, ρ retroflex, T specifies apical articulation, P bilabial. The letter v stands for consonantal voicing. Vocalic gestures are separately treated, and are represented here by phonemic symbols (none for reduced vowels). Syllable affixes (see Fujimura [4]) are separated by a dot from the final demissyllable. Time and magnitude are assigned to each pulse (represented by vertical lines). The pulses belong to one of two types:

syllables (shown with thick vertical lines), or boundaries (thin vertical lines). Each pulse is associated with minimal phonological feature or boundary type specifications. The phonological tree and the metrical grid, or equivalent abstract representations, constitute the primary basis of computation of the magnitude values of all pulses. Prominence specifications (see an exclamation mark), reflecting factors such as contrastive emphasis, the degree of excitement, etc., are also absorbed into the numerical specifications of time and magnitude. Utterance-related specifications (speaking style, speaker characteristics, etc.) are retained as annotations attached to utterance phrases. An utterance phrase constitutes the domain of motor programming as an integral unit of utterance (see Fujimura [6,7]), and affects both the impulse response functions and the parameters of the signal generator.

The timing characteristics of individual gestures are determined by the converter.¹ In Fig. 1, the i -th syllable pulse is located at t_i , and its height represents the magnitude μ_i assigned to each syllable. Let us assume the interval between two contiguous syllables to be related to the pulse magnitudes by

$$t_i - t_{i-1} = \alpha \mu_{i-1} + \beta \mu_i,$$

where α and β are multiplicative

¹ These time values are presumably readjusted via feedback signals from the signal generation process. For example, the articulatory repulsion, as discussed by Fujimura [5], apparently pertains to temporally adjacent gestures within the same articulator.

constants which determine "shadows" of each syllable pulse on the right and left sides, respectively. A similar shadow is also defined on the left side of a boundary pulse. This results in a leftward shift of the last syllable in the phrase before the boundary, making the decaying effect of the syllable pulse response function part of the overlapping next syllable. This accounts for the preboundary elongation.²

2. DISTRIBUTOR

The distributor distributes the codes produced by the converter to a concurrent set of actuators, each of which represents an articulatory dimension. An articulatory organ may be involved in defining multiple articulatory dimensions. An articulatory dimension may involve more than one organ. The distributor interprets the feature specifications for each demissyllable in terms of articulatory gestures, and distributes relevant syllable pulse information to individual actuators. In the figure, Greek letters in *italic* represent the elementary gestures in the distributor output, and Roman capitals represent the specified articulators (see below for further explanation). A family of mathematical functions prescribes the elementary articulatory gestures as time functions represented in terms of a physical measure of the state in each articulatory dimension. A set of muscular units forms a configuration of physical means for implementing cortical control of a specific dynamic event. This integral configuration of each gesture constitutes an articulatory dimension, such as production of a laminar /s/. Separate articulatory dimensions are defined for different manners of articulation, such as stops vs. fricatives. The output of the distributor is a replica of the input for each actuator to the extent the information is relevant. Thus the code (N, T) standing for /n/ in the final demissyllable of the second syllable /wAn/ is interpreted as { τ , T} for the tongue tip (T) closure (τ) dimension and operates in parallel with {N} for velum lowering. The impulse

² In addition, the parameters of the impulse response functions may be sensitive to the magnitude of the following boundary pulse.

response function for the {N} gesture (in final demissyllable) is implemented with peak event at about t_2 , whereas the { τ , T} gesture has its peak later (see Sproat and Fujimura [10] for similar situations of English laterals). This depiction may appear similar to Browman and Goldstein's gesture score [1,2].

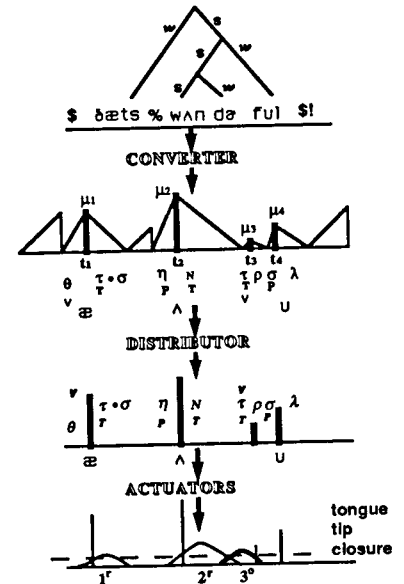


Fig. 1 A computational model of consonant implementation (signal generator omitted)

3. ACTUATORS

Each consonantal actuator receives the time-magnitude pulse information with respect to consonantal gestures for (1) an initial demissyllable, (2) a final demissyllable, (3) syllable affixes (in sequence) as applicable [4]. Different gestures are assigned to separate articulatory dimensions. Multidimensional processings take place concurrently in different actuators triggered by syllable pulses. Each actuator evokes the demissyllabic response, gesture by gesture. Elementary gestures do not require sequencing information within each demissyllable. The impulse response function contains a parameter that shifts the peak activity relative to the time value of the syllable

pulse. The parameter values are prepared in conformity with the inherent vowel affinity [3], [4].³ For consonantal gestures, each syllable impulse evokes an impulse response function in each of the relevant articulatory dimensions for each demissyllable. For example, the i -th syllable pulse in the phrasal domain under consideration has an assigned time t_i and a magnitude μ_i , and it triggers a feature event function $\mu_i g_{\tau}^o(t - t_i)$ in the dimension tongue tip raising, where $g_{\tau}^o(t)$ is the impulse response function for the gesture τ , stop closure, for the initial demissyllable (indicated by the superscript o). The time value t_i is designated for the i -th syllable. For the feature lateral, two articulatory gestures are relevant: (1) tongue tip raising for partial contact and (2) retraction of the back of the tongue body due partly to tongue blade narrowing [10]. The symbol λ in the distributor output in Fig. 1 stands for these two articulatory dimensions in an abbreviated form. Similarly, N , as in $\{N,T\}$, stands for the redundant N and τ of the nasal apical stop. The closure gesture τ is actually specified at the pertinent actuator, but is not shown in the figure. If the final consonant of the syllable is $/m/$, then the syllable pulse evokes the feature event function $\mu_i g_P^f(t - t_i)$ of the final demissyllable in the bilabial closure dimension P , and also $\mu_i g_N^f(t - t_i)$ in the velum lowering dimension N . This function for the final nasality shows a maximum velum lowering at about the peak occurrence of the syllable nucleus, i. e., $t = t_i$. Each actuator, within the pertinent articulatory dimension, compiles the feature event functions evoked by the

syllable pulses within the time domain of utterance phrase. The event time functions are added according to the linear superposition principle, and the resultant time functions for different articulatory dimensions, together with the vocalic base functions, are passed on to the signal generator. The same family of functions is used for prescription of all the consonantal gestures, with parameter values selected for individual articulatory dimensions. Each elementary event starts with the base position moving in the direction of the prescribed vocal tract constriction (in the three-dimensional sense), and then automatically returns to the base position. Different time constants are specified (in the gesture table for each actuator) for starting and ending trips. The lowest panel of Fig. 1 intends to suggest such occurrences of response events, for the final demissyllables of the first and the second syllables, and the initial demissyllable of the third, in the dimension tongue tip closure. The situation of the apparent target reaching short of the roof of the mouth, as in the case of $/s/$, is presumably a mechanical or physiological consequence of saturation in an inherently three-dimensional system.

4. SIGNAL GENERATOR

The signal generator, (not shown in the figure), receives the time functions generated by the total set of actuators, and synthesizes them with the vocalic base functions to materialize articulatory movement in an integrated system. Various types of interaction among articulatory dimensions are automatically treated by the physical model of the total system, both within the same articulatory organ (such as the lips or the tongue) and among different organs (such as the mandible and the lower lip). The system is highly nonlinear. In particular, it contains a strong saturating characteristic (soft clipping) so that a large syllable pulse typically results at the output of the signal generator in a plateau of articulatory position as a function of time. In Fig. 1, at the bottom of the figure, the horizontal dashed line indicates this "soft clipping" that takes place in the signal generating process. The process of generating the (vocalic) base function differs from that for

consonantal gestures. The syllable pulse magnitude is transformed by a nonlinear saturating function into a multiplicative factor that represents the degree of achieving the vocalic gesture target, relative to the neutral vocal tract condition (schwa gesture). The response function parameters are assigned for vocalic gestures in such a way that the peak position occurs with no delay relative to the input impulse. Target positions are specified for the peak.

5. CONCLUDING REMARKS

Our current data, obtained by the Wisconsin microbeam facility [8], concern the pellet positions representing sample flesh points on the surface of the articulatory organs along a particular direction of movement, as a crude approximation for the state variable in each articulatory dimension. More exactly, in our future work, the observed pellet time functions will be compared with predicted output functions using a dynamic three-dimensional computational model of the articulatory system. This computational model is currently being developed and will constitute the principal part of the signal generator.

6. REFERENCES

- [1]BROWMAN, C.P. & GOLDSTEIN, L. (1985), "Dynamic Modeling of Phonetic Structure", in V. A. Fromkin (ed.), *Phonetic Linguistics -- Essays in Honor of Peter Ladefoged* (pp. 35-53), New York: Academic Press.
- [2]BROWMAN, C.P. & GOLDSTEIN, L. (1989), "Tiers in Articulatory Phonology, with some implications for casual speech", in J. Kingston and M.E.Beckman (eds.) *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (pp. 341-376). Cambridge: Cambridge University Press.
- [3]CLEMENTS, G.N. (1989), "The Role of the Sonority Cycle in Core Syllabification", in J. Kingston and M.E. Beckman (eds.) *Papers in Laboratory Phonology I: Between the Grammar and the physics of Speech* (pp. 58-71). Cambridge: Cambridge University Press.
- [4]FUJIMURA, O. (1979), "An Analysis of English Syllables as Cores and Affixes", *Zs. für Phonetik, Sprachwissenschaft und Kommunikations-*

forsch. 32, 417-476.

- [5]FUJIMURA, O. (1986), "Relative Invariance of Articulatory Movements: An Iceberg Model", in J.S. Perkell and D.H. Klatt (eds.) *Invariance and Variability in Speech Processes* (pp. 226-242). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- [6]FUJIMURA, O. (1990a), "Articulatory Perspectives of Speech Organization", in W.J. Hardcastle and A. Marchal (eds.) *Speech Production and Speech Modelling* (pp. 232-342). Dordrecht: Kluwer Academic Publishers.
- [7]FUJIMURA, O. (1990b), "Methods and Goals of Speech Production Research", *Language and Speech* 33, 195-258.
- [8]NADLER, R.D., ABBS, J.H. & FUJIMURA, O. (1987), "Speech Movement Research Using the New X-Ray Microbeam System", *Proc. 11th ICPhS, Tallinn*, Se 11.4.
- [9]ÖHMAN, S.E.G.(1967), "Numerical Model of Coarticulation", *J. Acoust. Soc. Am.* 41, 310-320.
- [10]SPROAT, R.W. & FUJIMURA, O. (1989), "Articulatory Evidence for the Non-categorization of English /l/-Allophones", *Linguistic Soc. Am. Annual Meeting, Dec.89, Washington, D.C.*

³ An autosegmental computation is assumed at the level of distributor for spreading redundant feature values. For example, in the vocal fold adduction dimension, one specification of voicedness (for the entire consonant cluster) indicated for each demissyllable will be distributed throughout the time domain of obstruent gestures and affixes.

STUTTERING AS INDICATION OF SPEECH PLANNING

M. Koopmans, I. Slis and A. Rietveld

Dept. of Language and Speech, University of Nijmegen,
the Netherlands.

ABSTRACT

In this paper we will report the results of two experiments on the distribution of stuttering in spontaneous speech. Our observations support the idea that stuttering is related to syntactic planning in addition to the subordinate process of searching a specific word for a concept.

1. INTRODUCTION

Previous research has shown that in spontaneous speech of young stutterers, stuttering more often occurs during the beginning of clauses than in the remaining part [6]. Moreover, it has been found that stuttering is more likely to occur at those locations where the information load is high [4]. Soderberg [3] studied read out speech instead of spontaneous speech of adult stutterers. He considered the predominance of stuttering on the first words of clauses as grammatical uncertainty. In medial clause positions he observed stuttering predominantly on content words of high information. The explanation was that at the beginning of a clause the first foundations of a grammatical structure are laid. After that the main problem for the speaker is the choice of words. Therefore there is more lexical uncertainty at locations further in the clause. Our aim is to elaborate Soderberg's suggestions, especially by relating grammatical

uncertainty to decision moments in the planning of speech.

Speech planning can be described as a hierarchical process [1] {2}: it is possible to divide speech into segments and nested subsegments that can be related to the different stages in the speech planning process. In the production of speech the clause is considered a grammatical unit. The determination of the grammatical structure of a clause takes place at a hierarchically higher level than the process of word insertion. In this perspective the articulatory realization is of a still lower level than the word insertion level.

Our expectations were (1) that the hierarchical speech planning model gives an explanation for stuttering during the beginning of clauses, (2) that the model is able to explain the differences between stuttering on function words and lexical words depending on their locations in a sentence. We have tested both expectations in a speech experiment with adults. We used spontaneous speech as this kind of speech just requires conceptual and grammatical planning. As will be shown the quantitative analysis of the first experiment supports the speech planning model. The stutter frequency on the first two words (W1 and W2) is higher than on words in the rest of the clause (WR).

Experiment I does not give us any compelling evidence for the model: it is still possible that stutterers and nonstutterers use different segmenting strategies, so that we can't be sure that the positions of W1, W2 and WR within the clause are the same for stutterers and nonstutterers: e.g. if stutterers do not insert a boundary before W1, the labels W1, W2 and WR would be erroneous. Hence a second experiment is set up in which the task was to subdivide a written text into parts. It was assumed that the subdivisions reflect the internal structure of language [5]. We will now discuss the two experiments in more detail.

2. EXPERIMENT I

In this experiment we investigated stuttering on function and lexical words at various locations in a clause. In this study the minimal criterion for a group of words to be called a clause is that it contains one NP and one VP.

2.0 Procedure

2.1 Subjects

25 Stutterers, 7 females and 18 males, aged 17-45, participated in the experiment.

2.2 Speech material

To elicit spontaneous speech, all subjects were interviewed about the same theme i.e. their eating habits. From these interviews 1-minute-samples of spontaneous speech from each subject were selected. In these samples, the clauses were determined and the stuttered words were counted.

2.3 Method

We defined three wordpositions within each clause: the first word (W1), the second word (W2) and the remaining

words (WR). Each word was labeled as function word or lexical word. This distinction was made because the role of function words is largely grammatical, whereas lexical words carry the main semantic content. Besides, we may assume that lexical words have a relative high information load compared to function words. As it is questionable whether auxiliary verbs and copulas are function words or lexical words, we classified them in two ways. In one counting we considered them as lexical words (A), in another counting as function words (B) (table 1a).

2.4 Analysis

Table 1a was submitted to hierarchical loglinear analysis in order to examine the distribution of stutters over the words in a clause. According to the two definitions of lexical and function words two analyses were carried out. In both cases we studied the interactions of word type and word position. The three variables are: word type (lexical [L] and function words [F]), word position (W1, W2 and WR), and frequency of stuttering (number of stuttered words versus nonstuttered words).

Table 1a. The absolute numbers of stuttered and nonstuttered words on word-position (W1, W2, WR) and wordtype (L,F).

Counting A.

	W1		W2		WR	
	L	F	L	F	L	F
+st	7	54	31	41	126	44
-st	85	324	167	225	1140	733

Counting B.

	W1		W2		WR	
	L	F	L	F	L	F
+st	7	54	26	46	119	51
-st	64	345	126	266	1027	846

Table 1b. The percentages of stuttered words as a function of word position (W1, W2, WR) for counting A and B.

W1	W2	WR
13.0	15.6	8.3

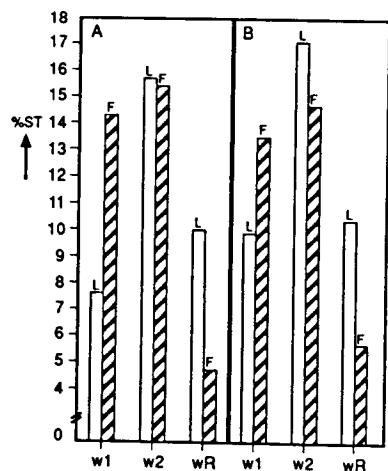


Figure 1. Percentages of stuttered words as a function of word position (W1, W2, WR) for lexical (L) and function words (F) separately. Counting A and B.

2.5 Results and discussion

The first finding is that in spontaneous speech, adults stutter significantly more often at the first and second words of clauses than at the remaining words. For both categorizations A and B defined above, this effect is present (table 1b). This result is in agreement with earlier investigations (see introduction). The second finding is that, using the data of categorization A, there is a significant interaction effect between word position and word type ($A: z=2.20, p=0.028$). Note that function words at the beginning of a clause give rise to more stuttering than in the mid-

dle or at the end of a clause. Besides there is relatively more stuttering on function words than on lexical words at the beginnings of clauses (table 1a, fig 1). The data of categorization B show the same trend as A but the effect is not significant.

Based on these findings we conclude that word position plays an important role in stuttering. As already mentioned, function words have merely a grammatical function. Therefore we might hypothesize that grammatical decisions are made in the beginning of a clause.

3. EXPERIMENT II

In this experiment we investigated segmenting strategies used by stutterers versus nonstutterers.

3.0 Procedure

3.1 Subjects

The subjects are 20 stutterers and 20 nonstutterers. Both groups were matched on education, age and sex.

3.2 Material

We use a printed text of 81 sentences which have been selected from the speech samples from experiment 1. Each word boundary in the text received a theoretical boundary strength from 1 to 5. These strengths were determined with the aid of boundary criteria of Umeda [5], but adapted to Dutch. The values were unknown to the subjects.

3.3 Task of the subjects

The instruction for the participants was to intuitively mark boundaries within the 81 sentences by putting vertical lines between words. The subjects were not given any rules about the number of boundaries per sentence.

Table 2a. Mean number of boundaries placed by stutterers(ST) and nonstutterers(NST).

	X(mean)
ST	0.30
NST	0.33

Table 2b. Mean number of boundaries placed by stutterers(ST) and nonstutterers(NST), for each theoretical boundary strength.

	1	2	3	4	5
ST	0.04	0.09	0.07	0.60	0.76
NST	0.06	0.13	0.09	0.60	0.80

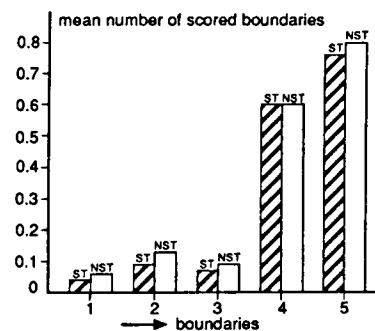


Figure 2. Mean number of boundaries placed by stutterers(ST) and nonstutterers(NST), for each theoretical boundary strength.

3.5 Results and discussion

An analysis of variance is applied on the mean number of boundaries scored by each subject on each theoretical boundary value. No significant difference is observed between stutterers and nonstutterers in the placement of boundaries ($F[1,36]=1.93, p=0.174$). We found no interaction effect between group and scoring of the five boundary strengths ($F[4,144]=0.16, p=0.958$). In both groups there are more boundary placements with increasing bound-

ary strength (table 2a-b, fig 2). These findings support that stutterers use the same linguistic criteria as nonstutterers for structuring language.

4. GENERAL CONCLUSION

The data of experiment I suggest a relation between the stuttering pattern and decision moments in the speech planning process. Moreover it looks more plausible to define auxiliary verbs and copulas as lexical words. The outcome of experiment II shows that the findings in experiment I are not due to different parsing strategies of stutterers and nonstutterers.

5. REFERENCES

- [1] GARRET, M. (1980), "Levels of processing in sentence production", in: Butterworth, B.(ed), *Language production, vol1: Speech and Talk*, New York: Academic Press.
- [2] LEVELT, W. (1989), *From intention to articulation*, Cambridge (Mass.): Bradford Books, the MIT press.
- [3] SODERBERG, G. (1967), "Linguistic factors in stuttering", *Journal of Speech and Hearing Research*, 10, 801-810.
- [4] ST. LOUIS, K. (1979), "Linguistic and motor aspects of stuttering". In: *Speech and Language*, New York: Academic Press, 89-210.
- [5] UMEDA, N. (1982), "Boundary: perceptual and acoustic properties and syntactic and statistical determinants", *Speech and Language*, New York: Academic Press.
- [6] WALL, M., STARKWEATHER, C. & H. CAIRNS (1981), "Syntactic influences on stuttering in young child stutterers", *Journal of fluency disorders*, 6, 283-298.

MRI (MAGNETIC RESONANCE IMAGING) FOR FILMING ARTICULATORY MOVEMENTS.

A.K. Foldvik*, O. Husby*, J. Kværness**, I. C. Nordli*, and P.A. Rinck**

*Department of Linguistics, University of Trondheim, Norway
** MR-Center, Medical Section, University of Trondheim, Norway

ABSTRACT

The range of applications of magnetic resonance imaging is rapidly increasing. Among the promising new techniques is cine-imaging of articulation which offers non-risk insights into speech production.

1. INTRODUCTION

Since speech typically consists of constantly and quickly changing configurations of the vocal tract, imaging of the vocal tract is of interest to speech research.

Cinegraphic X-ray techniques, X-ray computerized tomography, and ultrasound techniques are all in use. However, methodological problems such as side effects of radiation in connection with X-ray and inability to penetrate bone and air in ultrasound studies have caused problems in connection with these techniques.

The advent of magnetic resonance imaging (MRI) opened a new avenue for speech research and it was thought that this method would overcome the disadvantages of earlier techniques [1,5]. MRI shows not only the tongue but all articulators clearly and creates highly accurate three-dimensional images. Its disadvantages lie in the fairly long acquisition times, even when subsecond imaging techniques are used.

2. INSTRUMENTATION

We have used both medium and high field MR systems operating at 0.5 and 1.5 Tesla (Philips Gyroscan S5 and S15; Philips Medical Systems, Eindhoven, The Netherlands) with both head and circular surface coils.

Transverse and coronal spin-echo (SE) sequences were acquired as scout views for double oblique angulation of the desired 5 to 8 mm sagittal scans through the articulators. The sagittal image was used as scout for double oblique transverse and coronal images at different angles through the articulatory tract which provided basis for volume calculations.

For the actual data acquisition either gradient echo, Fast Field Echo, (FFE) or subsecond imaging (SS) sequences were used. In the FFE sequences echo times (TE) ranged from 6 ms to 15 ms, total acquisition times from 3.5 s to 12 s. Acquisition matrix size was 128 times 256, reconstruction matrix size 256 times 256. The acquisition time in SS sequences was 450 ms for a 90 times 128 matrix. Image acquisition could be repeated every 1.5 seconds.

The desired time resolution for speech production imaging is in the order of 10 to 20 ms to cover the movements of the articulators satisfactorily. Subsecond imaging does not provide such speeds. Therefore we returned to FFE sequences with their better signal-

to-noise ratio and higher spatial resolution.

3. SUBJECTS AND METHODS

Due to the special and specific conditions under which articulation had to be produced, only phonetically trained subjects were chosen for these studies. A number of sounds were used for interobserver control, others for intraobserver control, and some were carefully selected as markers of certain Norwegian dialects. Images were acquired both in supine and prone position to compare a possible impact of gravity on the articulators.

The pronunciation was performed in two stages: First a warm-up stage where a predetermined word containing the wanted sound was repeated several times, followed by a five second count-down. Then the speaker "froze" the predetermined word in the middle of the desired sound with a glottal stop, the image data was acquired, and the pronunciation was continued. This procedure allowed the production of static images of certain sounds.

Since speech is a dynamic procedure we wanted to get cine representations of the articulation of complete words.

Since we had no device to trigger the acquisition from spoken sounds, we synchronized an audible beep tone with an artificial ECG that triggered the informant and instrument respectively. The beep was transmitted to the informant through a plastic tube headset as used by airlines. The ECG was picked up by the MR system's telemetry receiver (Hewlett-Packard) and the acquisition was run similarly to a routine cardiac examination.

This meant repeating the word the same number of times as the phase encoding gradients and the spatial resolution. Typically we used 90 to

160 gradient steps resulting in an acquisition time between 1 and 4 minutes.

4. RESULTS

Findings from our MRI studies of Norwegian pronunciation [2,3,4] show that sounds which traditionally have been described as palatals are laminal alveolars and laminal postalveolars, retroflex sounds are in fact apical alveolar sounds, velar stops and nasals are postalpalatals even in front of open back vowels, and Norwegian front vowels [i: e: æ:] traditionally described as close, half-close and half-open respectively, have virtually the same position of the front part of the tongue, but with a marked difference in the degree of tongue root fronting (for [i:]) and tongue root retraction (for [æ:]).

5. DISCUSSION

MRI offers the possibility of recording the dynamics of speech and also has possibilities as regards vocal tract shape and volume mapping. The imaging method itself has to be improved. The inherent problem is related to time resolution. Subsecond imaging does not solve the problem since ideally acquisition time will never be short enough and the signal-to-noise ratio is not good enough.

The triggering can be improved in several ways; e.g. by the use of a microphone to pick up the speech and trigger the MR system, or by a sensor to pick up movements from the lips, chin or larynx.

To reduce acquisition time and dependence on cooperation with the informant, retrospective gating seems to be a promising method. This method could also lead to a further improvement in time resolution.

REFERENCES

- [1] BAER, T., GORE, J.C., BOYCE, S., NYE, P.W. (1987), "Application of MRI to the Analysis of Speech Production", *Magnetic Resonance Imaging*, 5, 1-7.
- [2] FOLDVIK, A.K., HUSBY, O., Kværness, J. (1988), "Magnetic Resonance Imaging.", *Proceedings Speech '88*. 7th FASE Symposium, Edinburgh, 423-428.
- [3] FOLDVIK, A.K., HUSBY, O., Kværness, J. (1989), "An Investigation of Norwegian Vowel Articulation by Means of Magnetic Resonance Imaging (MRI)." *Studies in Languages*, 14, 436-441. Joensuu.
- [4] FOLDVIK, A.K., HUSBY, O., Kværness, J., Nordli, I.C., Rinck, P.A. (1990), "Investigation of Articulation by Magnetic Resonance Imaging.", *Society of Magnetic Resonance in Medicine. Proceedings*. Ninth Annual Meeting, 115. New York.
- [5] VON WEIN, B., DROBNITZKY, M., KLAJMAN, S. (1990), "Magnetresonanztomographie und Sonographie bei der Lautgebung", *Fortschr. Röntgenstr.* 153, 408-412.

We intend to show our MRI films of articulation of Norwegian in connection with the presentation of this paper.

STABILITY OF VOICE FREQUENCY MEASURES IN SPEECH

W. J. Barry (1), M. Goldsmith (2), A. J. Fourcin (1), H. Fuller (2)

(1) University College London, (2) National Physical Laboratory

ABSTRACT

The stability of personal average laryngeal frequency during speech was examined in two-minute and fifteen-minute recordings made at different times of the day and for four differing speech production tasks. One and a half hours of speech from each of four subjects were analysed with respect to mean and modal frequency and frequency range.

The work was funded under Alvey Project MMI/132, Speech Technology Assessment.

1. INTRODUCTION

Voice source characteristics are an important part of the expressive component in the speech communication chain. The most commonly used of these is the frequency of vocal-fold vibration, which is considered both subjectively [1] and objectively ([9], p.82) important for the identification of a speaker. However, although there have been a large number of studies devoted to the measure of average voice fundamental frequency in groups of 50 or more speakers ([9], 2] for survey information), and there is general recognition that a speaker's voice pitch varies with the situation, little has been done to clarify the general question of individual stability in that or other voice frequency measures.

This paper presents results from an in-depth study [4] of the laryngeal patterns in speech of four speakers, and investigates the question of stability in a number of ways. Firstly, on the assumption

that two minutes of continuous speech are sufficient to characterise a speaker's voice frequency, [12, 11, 10, 8], the "stability" of two-minute stretches over a 15-minute period was examined. Secondly, in view of evidence that that system-atic longer term fluctuations occur [7], a series of 15-minute recordings were made at different times of one day. Thirdly, a number of different types of speech tasks were set in order to examine the stability of average voice frequency over tasks.

2. SPEECH MATERIAL

Larynx and speech-signal recordings were made with four speakers (2F, 2M) on two days under anechoic conditions using a two-channel PCM-VCR recorder. Two speakers took part in 6 recording sessions of approximately 15 minutes each during the course of each of the two days. A standard extended reading task (P+B) was given at 9.40am, 12.10pm, and 3.30pm (times ± 20 minutes), namely the reading of the "Environmental Passage" [3] followed by 13 minutes of continuous reading from a book of short stories. These were interspersed (at 10.20, 11.20 and 12.30 ± 20 minutes) with three further tasks for comparison purposes: 1) Eight consecutive readings (RP) of the "Environmental Passage" to provide a constant textual structure and avoid fluctuations in interest and excitement. 2) Free monologue (Mon.) for 15 minutes on a subject of

each speaker's choice to give a longer term comparison between read and freely spoken monologue. 3) A dialogue with pairs of speakers (Dial.) to compare both with reading and with free monologue.

3. ANALYSIS PROCEDURES

Analysis of the larynx signal (Lx) was carried out in the standard way [4, 3, 5] to give laryngeal frequency distributions over 128 logarithmically equal bins on a scale from 30 to 1000Hz. "Cleaned" distributions were used for all quantitative statements and statistical calculations. They were derived by including Lx cycles only if their durations were within 10% of the preceding cycle. This eliminates laryngeal irregularities from the distribution while retaining a maximum number of data points. In addition, the distributions are time-weighted to correspond to our auditory impression of pitch movement as frequency change in time, and to conform to other voice-frequency studies, which have obtained their data from fixed *frame*-based rather than from fundamental- or laryngeal-*period*-based analysis.

Processing of the 15-minute sessions was done on sections of approximately 2 minutes, and overall distributions for the whole session was obtained by cumulation of the shorter sections.

4. RESULTS

Two aspects of larynx frequency stability are addressed: Firstly, the extent to which a "minimal stable period" (i.e. 2 min.) still fluctuates within a longer stretch of speech. Secondly, whether there are genuine systematic baseline changes in the longer term.

4.1. Two-minute sections

There are considerable differences between the mean values for the 2-minute sections both within and across the different tasks. Table 1 gives the range in semi-tones between the lowest and highest mean and modal frequency for a two minute stretch during each of the tasks.

The largest shifts in mean frequency appear to be randomly distributed over tasks and speakers. The comparison of the repeated-passage reading (Task 2) with the other conditions does not support the assumption that shifts in mean frequency are a function of text-type alone. The difference in Task

Table 1. Difference in semi-tones between the highest and lowest mean & (*modal*) frequency for a two-minute stretch during each 15-minute task.

TASK	FC	CP	AH	TJ
1.	1.26 1.01	0.45 1.04	0.53 3.41	1.15 1.68
2.	1.36 1.01	0.19 0.00	0.72 0.86	0.91 0.70
3.	0.51 0.00	1.05 1.04	1.93 0.86	0.60 0.80
4.	0.75 2.10	0.81 1.10	1.35 2.60	2.12 1.60
5.	1.56 1.01	0.95 1.04	0.66 0.85	0.19 0.00
6.	0.91 2.10	0.82 1.04	0.87 0.88	0.87 4.23

2 is very low for speaker CP, but not for the other speakers, and in fact has the highest value of all tasks for FC. In general, the pattern in time for 15 minutes is more regular than for the book readings, but in no uniform way across speakers. CP has one small shift, FC has a steadily rising frequency, and the two male speakers AH and TJ have intermittent fluctuations. Modal values, however, do not reflect the expected task-dependence any more clearly.

The differences between the speakers in the repeated-passage task, and the random distribution of large and small pitch differences for 2-minute stretches across the different tasks, indicate that there is, strictly speaking, no such thing as a generally valid personal voice-frequency value, either mean

or mode (compare [1]). Each situation appears to have its specific and individual effect on a speaker's voice frequency.

4.2 Longer-term stability

To test for variability in the longer term, the values found for the standard passage in tasks 1, 4 and 6 in this study were compared with those for the same speakers reading the same passage (P) some weeks earlier [3]. Table 3 lists the mean and modal values for all the single readings.

Table 3. Mean and (modal) Fx values (Hz) for the standard passage in Tasks 1, 4 and 6 and in P.

TASK	FC	CP	AH	TJ
1	213 197	229 222	114 108	104 103
4	223 209	239 222	122 114	109 103
6	228 209	234 222	123 108	110 114
P	208 209	225 236	130 126	102 98

The overall picture of variation from situation to situation is confirmed by this overview. For each speaker, values for the same text vary from one recording to the next as much as the different 2-minute stretches varied within the 15-minute readings from a book. However, table 3 shows a certain regularity in the values from task 1 to task 6. For all four speakers there is an increase in mean Fx from task 1 to task 4. This is followed by a fall-off in task 6 for CP while TJ and AH maintain approximately the same mean excitation frequency and FC shows another increase. Given the timing of the tasks (morning, mid-day, afternoon) this progression is similar to that reported in [7], namely an increase in fundamental frequency for all speakers during the first part of the day (9am-noon). In partial contrast to our results, however, an increase was found

there for the male speakers during the afternoon with a fall-off restricted to female speakers.

These shifts in the 2-minute standard passages are related to shifts in mean frequency over the 15-minute stretches of which the standard readings are a part. Three of the four speakers show a marked upward shift from morning to mid-day and fall back from mid-day to afternoon. The one speaker (TJ) who has no marked shift in overall mean pitch from morning to mid-day does have a significant shift from mid-day to afternoon (Mann-Whitney, $U' = 6$, $p = 0.05$).

Modal values both for the 2-minute passage and the whole 15-minute reading tasks confuse this fairly clear picture. In the passage (table 3) CP deviates from the pattern, in that her modal value remains constant throughout the day. In the 15-minute values, TJ is the exception, with a steady decrease of mode from 108Hz in task 1, 98Hz in task 4, and 94Hz in task 6.

4.3 Task dependence

Table 4 gives the overall mean values for monologue, dialogue, repeated reading, and tasks 1,4,6 together (compare the latter with table 3 values for the separate free reading sessions).

Table 4. Mean/modal values (Hz) for tasks 3, 5, 2 and 1+4+6.

TASK	FC	CP	AH	TJ
3	191 186	237 222	114 108	89 89
5	209 197	222 222	104 98	100 98
2	228 197	232 222	117 98	101 98
1,4,6	223 209	239 232	121 106	104 98

There is a strong tendency for mean voice frequency to be lower for spontaneous than for read speech. All four dialogue values

and three of the four monologue values (exception CP) are lower than for any of the reading conditions. This finding parallels the results from the STA Normative Study [3] and those from the literature discussed in it. The modal values, however, show a different pattern. Neither the monologue nor the dialogue are consistently lower than the reading tasks, and AH rather than CP is the exception.

5. SUMMARY AND CONCLUSIONS

The results from this investigation of voice-frequency stability indicate very clearly that while mean frequency has been shown to stabilise over a speech sample of 2 minutes or more in duration, fluctuations from one sample to another forbid the use of such a measure as an absolute personal characteristic. It can only serve as a characterisation of the sample in question. Across speech tasks, and even between different 2-minute samples of the same extended task, mean frequency was shown to vary by as much as 15%. Within speakers, some regularity was found in mean-frequency change during the course of the day, and between spontaneous and read speech. This supports previous findings in the literature, but individual variation in the patterns found indicate that these trends also need to be treated with caution.

In the light of these results, the conclusion is unavoidable that reliance on mean voice frequency as an indicator of personal identity is inadvisable. The possible alternative measure, modal frequency, appears less sensitive to task variation, but it still does not offer a reliable voice-frequency characterisation of speakers.

6. REFERENCES

- [1] E. ABBERTON & A.J. FOURCIN, Intonation and Speaker Identification. *Lang&Sp.* 21, 305-318 (1978)
- [2] R. J. BAKEN, *Clinical Measurement of Speech and Voice*. Boston: College Hill Press, (1987)
- [3] W.J. BARRY, M. GOLDSMITH, A.J. FOURCIN, & H. FULLER, H.

(1990): Larynx Analyses of Normative Reference Data. *Project Report, Alvey Project MMI 132*, London: University College.

[4] W.J. BARRY, M. GOLDSMITH, A.J. FOURCIN, & H. FULLER (1990): Stability of Laryngeal Measures in Speech. *Project Report, Alvey Project MMI 132*, London: University College.

[5] A.J. FOURCIN, Laryngographic assessment of phonatory function. In: C.L. Ludlow (ed.) *Conference on the Assessment of Vocal Pathology*, Maryland: ASHA Reports 11, (1981)

[6] H.C. FULLER, A.J. FOURCIN, M.J. GOLDSMITH & M. KEENE (1990): A Database of Normative Speech Recordings. *Proceedings Institute of Acoustics* 12, part 10, 1-6

[7] K.L. GARRETT & E.C. HEAL-EY, An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of the day. *J. Acoust. Soc. Amer.* 82 (1), 58-62, (1987)

[8] S. HILLER, J. LAVER & J. MacKENZIE, Durational aspects of long-term measurements of fundamental frequency perturbations in connected speech. *Work in Progress* 17, 59-76, Dept. of Linguistics, Univ. of Edinburgh, (1984)

[9] H.J. KÜNZEL, *Sprechererkennung. Grundzüge forensischer Sprachverarbeitung*. Heidelberg, (1987)

[10] J.D. MARKEL & S.B. DAVIS, Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. *IEEE Transactions, ASSP-25*, 330-337 (1979)

[11] K. O. MEAD, Identification of speakers from fundamental frequency contours in conversational speech. *JSRU Report 1002*, Ruislip, Middlesex, (1974)

[12] M. STEFFAN-BATOG, W. JASSEM & GRUSZKA-KOSCIELAK, Statistical distribution of short-term F0 values as a personal voice characteristic. In: W. Jassem. (ed.) *Speech Analysis and Synthesis* 2, 195-206, Pol. Acad. Science, (1970)

ELECTROMYOGRAPHIC STUDY ON LARYNGEAL ADJUSTMENT FOR WHISPERING

Koichi Tsunoda, Seiji Niimi, Hajime Hirose,
* Katherine S.Harris and **Thomas Baer

Univ. of Tokyo, Tokyo, Japan.
* Haskins Laboratories, New Haven, U.S.A.
** Cambridge Univ. Cambridge, U.K.

ABSTRACT

In order to clarify laryngeal adjustments for whispering, an electromyographic study of the intrinsic laryngeal muscles was conducted. Subjects were two native Japanese speakers.

Posterior cricoarytenoid muscle (PCA) showed higher background activity during whispering than during ordinary phonation. Furthermore, there were additional segmental activities which seemed to contribute not only to keeping glottis open, but also to performing necessary gesture for relevant phonemes. The activity patterns of the other intrinsic laryngeal muscles are also discussed.

1. INTRODUCTION

It has been said that whispering is an aphonic laryngeal action^[1]. It means that there is no vocal fold vibration. However, even during whispered speech, one can distinguish the difference between "voiced" and "voiceless" segment as well as accent patterns^[2]. There have been only a few reports dealing with the activity of the intrinsic laryngeal muscles during whispering^[3,4]. In order to elucidate the laryngeal adjustment, an electromyographic (EMG) study was performed.

2. PROCEDURE

Subjects were two native speakers of Tokyo dialect of Japanese. EMG activity was recorded from PCA, Cricothyroid muscle (CT), lateral cricoarytenoid muscle (LCA), vocalis muscle (VOC) and interarytenoid muscle (INT). For EMG recordings, hooked wire electrodes were used. The electrodes were inserted

perorally to PCA and INT, while inserted transcutaneously to the other muscles. The method of verification of electrode location described elsewhere^[5]. As an indication of articulatory gestures, intraoral pressure was also recorded on an FM data recorder with EMG signals.

The subjects were required to utter nonsense words of /CVCV/ form in whispering and in ordinary speech with different accent patterns.

The test words were:

/t̚t̚tel/, /t̚e t̚e/, /te tel/
/d̚d̚del/, /d̚e d̚e/, /de del/
/s̚s̚sel/, /se s̚e/, /se se/
/z̚z̚zel/, /ze z̚e/, /ze ze/.

Each test word was embedded in a carrier sentence "i: /CVCV/ desu" (It is a good /CVCV/), and uttered more than ten times in whispering and in ordinary speech.

After the recordings, EMG signals were rectified and computer-processed in order to obtain an average indication of each muscle activity. As the line-up point for averaging EMG signals, the onset of abrupt drop of the intraoral pressure during consonant closure was taken so as to identify the oral release.

3. RESULT and DISCUSSION

Figure 1 shows the averaged EMG patterns for whispering (dotted line) and ordinary phonation (solid line) in subject KT. In ordinary phonation, PCA activity decreases for abduction of vocal folds for phonation from higher activity

level for inspiration. On the other hand, in whispering, there is no suppression of PCA activity but rather increment of the activity for utterance. Furthermore, in addition to the elevated activity level, PCA showed segmental pattern related to each phoneme. (Top panel of the Fig.1)

Each EMG peak which is corresponding to a phoneme is smaller for the whispering than for the ordinary phonation. This difference in EMG peak amplitude between whispering and ordinary phonation is clearly seen on the LCA and CT trace. However, the pattern of the perturbation is similar in both modes of phonation.

Whispering has been defined as "aphonic laryngeal gesture". From this point of view several studies have been conducted to describe laryngeal gesture during whispering mainly by using the fibroscope^[6]. These studies conclude that during whispering, the glottis kept open to prevent vocal fold vibration. The increased PCA activity throughout the utterance as the upward shift of the base line may represent the applied force to open the glottis.

Interestingly, PCA activity showed the segmental pattern which is corresponding to each phoneme. In other words, during whispering PCA showed compatible EMG patterns to the ordinary speech except for higher base line activity. The higher base line activity may contribute to change the speaking mode from the ordinary phonation to the whispering mode. Even in the whispering mode, the general laryngeal attitude to distinguish each phoneme would be preserved. This hypothesis can explain the difference in EMG patterns for the production of "voiced" and "voiceless" segment uttered in whisper. Figure 2 shows the difference in EMG patterns for the production of /t̚t̚e/ and /d̚d̚e/. We can see higher activity of PCA for "voiceless consonant /t̚/" than for "voiced consonant /d̚/" just before the line-up point.

On the other hand, there is high PCA activity at the beginning of utterances around 300 msec before the line-up point. Judging from the intraoral pressure curve (at the bottom panel), this high activity of

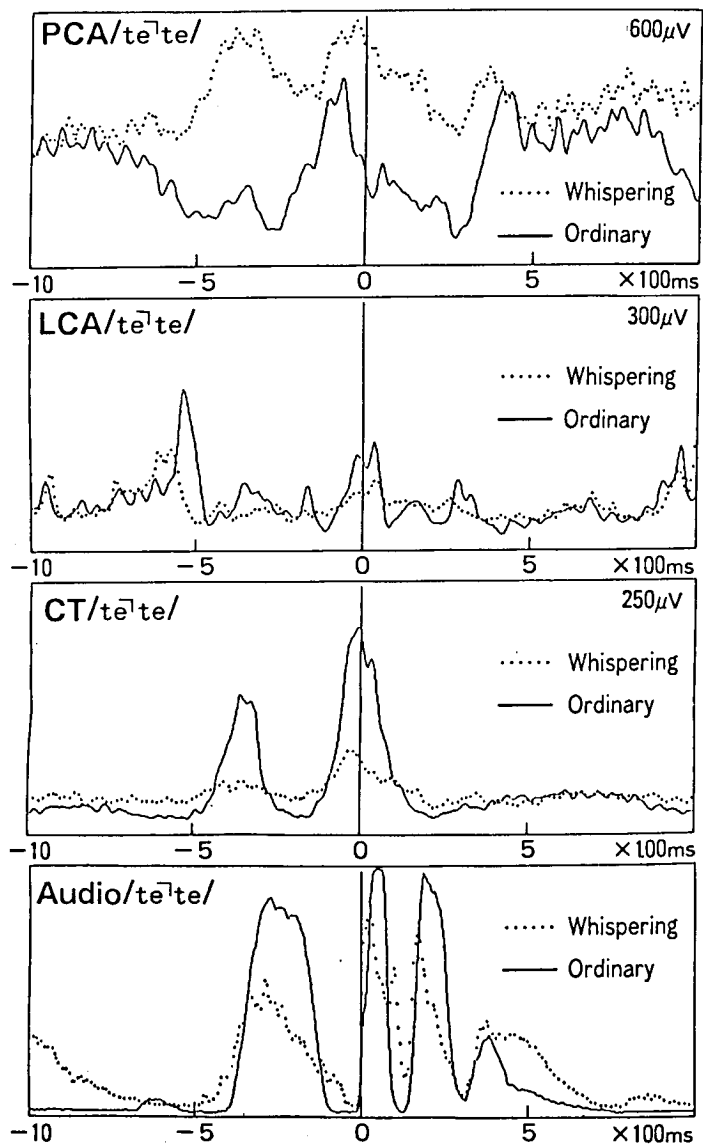
PCA is supposed to be related to the first syllable of the carrier sentence /i:/. This observation is contradictory to the previous finding, that is, in whispering PCA becomes active for "voiceless" segments and less active for "voiced" segments. To explain this controversial phenomenon, we have to consider that this segment /i:/ is located at the utterance initial. Probably, this particular phonetic condition may cause a different laryngeal gesture. We can speculate that for /i:/ at the utterance initial, since the adductor muscles become highly active, PCA should be prove this speculation, a fiberoptic study is mandatory.

4. Conclusion

For the production of whispering, PCA activity level becomes higher to prevent vocal fold vibration. In general, the intrinsic laryngeal muscles including PCA still have segmental activity patterns which are compatible to laryngeal gestures observed in ordinary phonation. At the very initial of the utterance, laryngeal gesture for whispering may special.

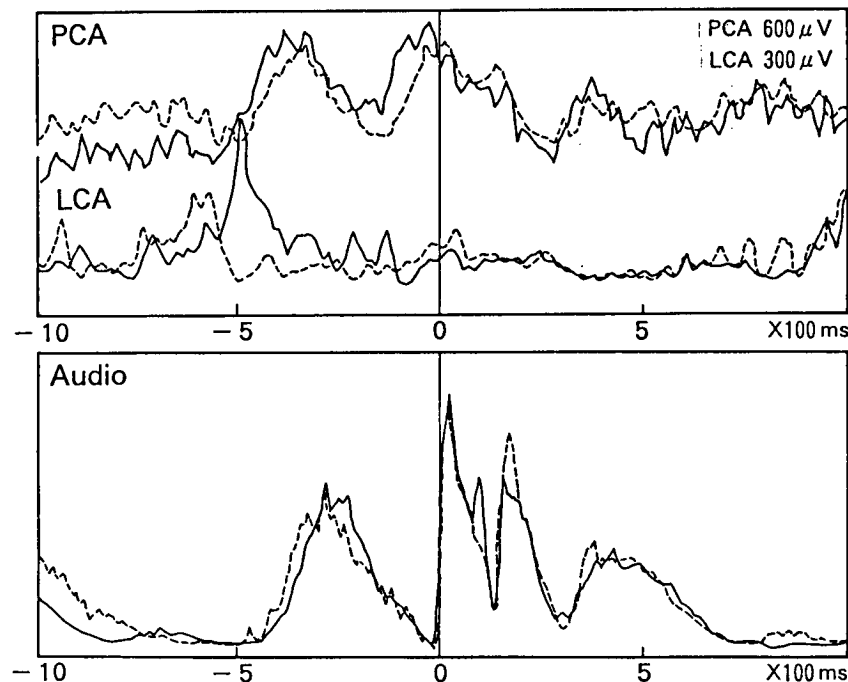
(This study was partially supported by Grant-in-Aid for Scientific Research (A) No.01440071 from the Japan Ministry of Education, Science and Culture.)

Figure1



Subj. K. T.

Figure2 Subj. K.T. { — / $te/$ / $te/$
- - - / $de/$ / $de/$



References

- [1] LUNCSIGER, R (1965), "Whispering and aspiration", Voice-Speech-Language. Wadsworth Publishing Company.
- [2] TARTTER, V.C. (1989) "What's in a whisper?", Journal of the Acoustical Society of America, 86:1678-1683.
- [3] FAABORG-ANDERSEN, K. (1957) "Electromyographic investigation of intrinsic laryngeal muscles in humans", Acta Physiologica Scandinavica, 41, Suppl. 140, 1-149.
- [4] SAWASHIMA, S. and HIROSE, H. (1973) "Laryngeal Gestures in Speech Production", In MacNaeilage, P. (Ed.), The Production of Speech. Springer-Verlag.
- [5] HIROSE, H. (1977) "Electromyography of the larynx and other speech organs." In M. Sawashima and F.S. Cooper, (Ed.), Dynamic Aspects of Speech Production. Univ. of Tokyo Press.
- [6] WEITSMAN, R., SAWASHIMA, M., HIROSE, H. and USHJIMA, T. (1976) "Devoiced and Whispered vowels in Japanese", Ann. Bull. RILP, 10:61-79.

THE INTRINSIC FUNDAMENTAL FREQUENCY OF VOWELS AND THE ACTIVITY IN THE CRICOTHYROID MUSCLE

N-J. Dyhr

Department of General and Applied Linguistics,
University of Copenhagen, Denmark.

ABSTRACT

The relationship between EMG activity from the CT muscle and IF_0 of Danish vowels was investigated. The results show a positive correlation: The CT activity rise in a vowel with a higher IF_0 starts earlier and has a higher overall amplitude than in a vowel with a lower IF_0 . A subset of data was compared to data from an identical material recorded earlier from the same subjects. Both sets show the positive correlation described above. The following conclusions are drawn: The CT muscle is an important factor in controlling IF_0 , and the reproducibility of EMG data is high.

1. INTRODUCTION

It is well known that the fundamental frequency (F_0) of vowels is correlated with vowel quality: other things being equal, high vowels tend to have a higher F_0 than low vowels. This phenomenon, which is known as intrinsic fundamental frequency (IF_0), has been reported in numerous investigations. Various theories have been advanced to explain IF_0 ; for a detailed survey see Silverman [6]. The generally accepted theory is based on the assumption that an increased tongue pull during the articulation of high vowels may give rise to an increased vertical tension in the vocal folds, Ohala [3].

Most theories have explained IF_0 as a passive influence from articulation. Very recent data,

however, point in the direction of a nervous control, more specifically that the cricothyroid (CT) muscle seems to be an important factor in the control of IF_0 . Honda [2] and Vilkmán et al. [7] found a positive correlation between peak values of CT activity and IF_0 . Dyhr [1] observed an earlier and steeper CT activity rise in high vowels than in low vowels. It was impossible to determine whether the CT activity was the result of a synergistic relationship between CT and other muscles, or whether it was a planned control of IF_0 .

The purpose of the present investigation is: 1) To look into the relationship between CT activity and IF_0 in a large set of Danish vowels, in both stressed and unstressed positions. 2) To compare a subset of data to data from an identical material recorded at an earlier session.

2. METHODS

2.1 Subjects and Material

The subjects were one female and four males, all phoneticians and native speakers of Danish.

The material consisted of the following Danish vowels: [i e ɛ u o ɔ ɔ] inserted in nonsense words of the type fvfvf'V, with identical vowels in each syllable, and [i: æ: u:] inserted in natural words of the type C'V:lɔ. The test words were embedded in carrier sentences and were read in four different randomizations.

2.2 Recordings

The electromyographic (EMG) recordings took place at the Dept. of Clinical Neurophysiology, Copenhagen University Hospital, Denmark, using Disa Electromedical Equipment. The EMG signals were collected from pars recta of the CT muscle, via concentric bipolar needle electrodes (Dantec Electronics, type 13L51, 20 mm). The insertion was performed by Dr. S. Fex, Lund University Hospital, Sweden, and was made percutaneously. The correct electrode position was controlled by a series of tests such as swallowing, high/low and gliding tones, and glottal closure. The acoustic signal was registered by an accelerometer (Brüel & Kjær, type 4375), taped directly to the skin above the larynx. The EMG and acoustic signals were recorded on an FM tape recorder, 30 ips. The recordings were monitored continuously via an oscilloscope.

2.3 Data Analysis

The EMG signals were rectified and integrated, with an integration time of 25 ms. The filtering was done in accordance with the results from Rischel & Hutters [5]. The F_0 analysis was carried out on an F-J. Electronics Fundamental Frequency Meter. Physiological and acoustic signals were sampled and averaged. Data were sampled through an eight-channel multiplexed analog-to-digital converter controlled by a real time clock. The sampling took place with a 1.25 sec. window at a sampling frequency of 200 Hz. Data and results were displayed on a graphic terminal.

3. RESULTS

The results are based on visual inspection of average EMG and F_0 curves. The average is based on at least six repetitions. The individual curves were carefully examined before averaging, and none of them were in disagreement with the final results shown in

the average curves.

In order to correlate a muscle action with the resultant effect the muscle activity must be shifted forward in time compared to the actual event. This is because it takes some time for the muscle to contract after being innervated, primarily due to electrochemical transmission and inertia. In the present data this so-called time lag was measured from the EMG activity peak to the corresponding F_0 peak on the individual curves before averaging. The time lag varied between 60-110 ms. In the following description and discussion compensation has been made for this time lag.

3.1 IF_0

For all subjects there is a correlation between vowel height and IF_0 , such that a higher vowel has a higher IF_0 than a lower vowel. This difference is most prominent among stressed vowels and least prominent among vowels in second pretonic position. The order of the vowels from high to low IF_0 is identical to data reported earlier for Danish, Reinholt Petersen [4]. In stressed position two subjects have [u] clearly higher than [i], whereas the rest of the subjects hardly have any difference between the two.

3.2 CT Muscle Activity

Two subjects show a positive correlation between CT activity and IF_0 for each step in vowel height in both stressed and unstressed vowels. Two subjects have the same results in stressed vowels, but for the unstressed vowels there is a positive correlation only between high and low vowels. The fifth subject shows a positive correlation in stressed vowels only, and only when the vowels are separated by two or more steps in vowel height. The observed IF_0 difference between [i] and [u] is clearly reflected in the CT activity. The general picture is that the CT

activity rise related to a vowel with a higher IF_0 starts earlier and has a higher overall amplitude than in a vowel with a lower IF_0 , see figure 1.

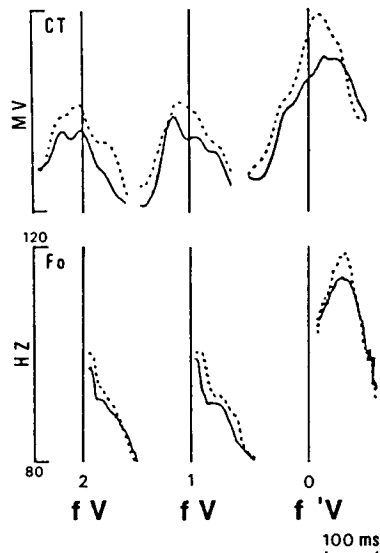


Figure 1. Comparison of extracts from superimposed average CT and F_0 curves from sentences containing the nonsense words [fifif'i] (broken lines) and [fefef'e] (solid lines) from the same subject. The line-up points (2,1,0) are the onset of the second pretonic (2), the first pretonic (1), and the stressed vowel (0). Notice that the IF_0 differences between [i] [e] are reflected in the CT curves in such a way that the CT activity rises earlier and has a higher overall amplitude in [i] than in [e].

Data from [i: æ: u:] are compared to data from an identical material recorded at an earlier session by the same subjects, Dyhr [1]. However, the EMG activity was then recorded from the pars obliqua of the CT muscle by means of bipolar hooked-wire

electrodes. Both sets of data show the positive correlation between CT activity and IF_0 described above, see figure 2.

4 DISCUSSION

The results indicate that the CT muscle plays an important part in the control of IF_0 in vowels. Such a control can be explained in two ways: It may be the by-product of a synergistic relationship between the CT muscle and other larynx muscles and/or the muscles responsible for shaping the vocal tract for the different vowels, or it may be a specific, planned control of IF_0 . A possible explanation of a planned control could be that IF_0 differences are important to speech perception, Silverman [6]. The results also imply that the pars recta and pars obliqua of the CT muscle are physiologically identical even if they are anatomically separated (at the moment further investigation is being carried out on this matter). Apart from some discomfort while swallowing, the use of bipolar concentric needle electrodes was a success. The insertion and adjustment of electrode position was easy. The fixation of the needle electrodes was surprisingly stable, and the EMG signals were less problematic than the ones collected via hooked-wire electrodes and were consequently less complicated to filtrate. Even with different registration methods the reproducibility of the EMG data was high. This implies that EMG is also a reliable tool when applied to tiny muscles as found in the speech apparatus.

5 CONCLUSION

The following conclusions are drawn: 1) The CT muscle is an important factor in controlling the IF_0 in vowels. Whether this is the result of a synergistic relationship between larynx and other muscles or a planned control of IF_0 is impossible to determine from the present data. 2) The

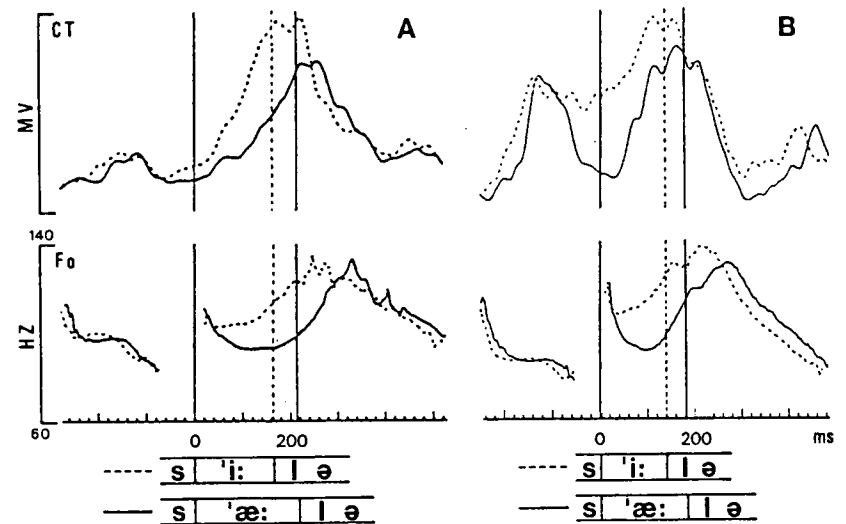


Figure 2. Comparison of superimposed average CT and F_0 curves from sentences containing the natural words [s'i:lə] (broken lines) and [s'æ:lə] (solid lines) from the same subject. The line-up point (0) is the onset of the stressed vowels, the offsets are marked with vertical broken/solid lines. A: The EMG signal is recorded from pars obliqua of the CT muscle via bipolar hooked-wire electrodes. B: The EMG signal is recorded from pars recta of the CT muscle via bipolar concentric needle electrodes. Notice that the IF_0 differences are reflected in the CT curves in such a way that the CT activity rise related to the high vowel starts earlier and has a higher overall amplitude than when related to the low vowel.

reproducibility of EMG data is high, even when different registration methods are used.

5. REFERENCES

- [1] DYHR, N.-J. (1991), "The activity of the cricothyroid muscle, and the intrinsic fundamental frequency in Danish vowels". forthc. *Phonetica*, 47.
- [2] HONDA, K. (1983), "Variability analysis of laryngeal muscle activity", in Titze, I. R. Scherer, R. C. (eds.): *Vocal fold physiology*, Denver: The Denver Center for the Performing Arts. 127-137.
- [3] GHALA, J. J. (1973), "Explanation for the intrinsic pitch of vowels", *Monthly Internal Memorandum, Phonology Lab. Univ. of Calif., Berkeley*, 12, 9-24.

- [4] REINHOLT PETERSEN, N. (1978), "Intrinsic fundamental frequency of Danish vowels", *Journal of Phonetics*, 6, 177-190.
- [5] RISCHER, J. HUTTERS, B. (1980), "Filtering of EMG signals", *Ann. Rep. Inst. Phonetics, Univ. Copenhagen*, 14, 285-309.
- [6] SILVERMAN, K. (1984), "What causes vowels to have intrinsic fundamental frequency?" *Cambridge Papers in Phonetics and Experimental Linguistics*, 3, 1-15.
- [7] VILKMAN, E. AALTONEN, O., RAIMO, I. ARAJÄRVI, P. OKSANEN, H., (1989), "Articulatory hyoid-laryngeal changes vs. cricothyroid muscle activity in the control of intrinsic F_0 of vowels", *Journal of Phonetics*, 17, 193-203.

LARYNGEAL AND ORAL GESTURES IN ENGLISH /P, T, K/

André M. Cooper

The University of Michigan, Ann Arbor, Michigan and
Haskins Laboratories, New Haven, CT, USA

ABSTRACT

The present results support recent findings showing that VOT is shorter for /p/ than for the two lingual stops /t, k/ and that VOT for lingual stops are generally equivalent. Further, the results offer no support for a compensatory relationship between closure duration and VOT and show that the laryngeal devoicing gesture differs for stops produced at different places of articulation, thus ruling out several articulation-based explanations for place-related differences in VOT. Finally, the results suggest that the timing of glottal adduction relative to oral release most nearly accounts for observed differences in VOT.

1. Introduction

A number of researchers have found that VOT increases as the place of articulation of a stop progresses from the front to the back of the vocal tract [4, 6, 8]. One possible explanation for this finding is based on the assumption that the devoicing gesture (i.e., the opening and closing of the glottis for devoicing) is invariant while supralaryngeal gestures get progressively shorter the further back a stop is articulated [7, 8]. Other proposed explanations refer to automatic aerodynamic or mechanical consequences or to perceptual requirements associated with stops produced at different places of articulation [3, 5].

Results from a number of recent studies of both American [1] and British [2] English, however, have cast doubt on the conventional view of place-related differences in VOT and their explanations. These findings indicate that VOT for labial stops is shorter than for lingual stops, while VOT for /t, k/

tend not to differ from one another. Indeed, most earlier studies reporting place-related differences in VOT show smaller VOT differences between /t/ and /k/ than between /p/ and /t, k/, a difference that may not have been statistically significant [2]. Thus, explanations of VOT differences which crucially refer to a stop's place of articulation cannot account for the data from these recent findings.

The purpose of this study is three-fold: (1) to examine place-related differences in closure duration and VOT in different word positions and under different stress conditions to test whether there is a compensatory relationship between the two; (2) to determine whether there is a single invariant devoicing gesture for all stops across different places of articulation; (3) to explore the role of oral-laryngeal timing with respect to VOT.

2. Methods

Two male speakers of English, ES and KM, spoke the nonsense words /pipip, titit, kikik/ with primary stress either on the initial or the final syllable in the carrier phrase "say _ again". Both acoustic and transillumination signals were collected simultaneously. Since the two speakers sometimes exhibited different articulatory patterns, separate statistical analyses were performed for each.

3. Results and Discussion

3.1. Acoustics

3.1.1. Closure duration

Separate ANOVAs for each word position indicate that stops produced at different places of articulation differ in closure duration in both initial and medial positions for both speakers (fig.

1). For ES, individual protected t-tests indicate that closure duration is significantly longer for /p/ than for /t, k/ but that there is no significant difference between closure durations for /t/ versus /k/. Similarly, for KM, closure duration is longest for /p/, and although mean closure duration is consistently longer for /t/ than for /k/, the effect is only significant in medial stressed position.

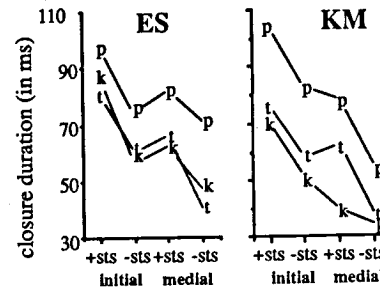


Fig. 1. Closure duration results for ES are presented on the left and results for KM on the right. The letter corresponding to the stop category is plotted in the graph. +sts = stressed and -sts = unstressed.

3.1.2. VOT

In general, the well-documented place-related VOT pattern for English voiceless stops is exhibited for each stress category in both word-initial and word-medial positions for both speakers (fig. 2). Separate one-way ANOVAs confirm that stops produced at different places of articulation significantly differ in VOT for both speakers. For ES, VOT is significantly longer for /t/ than for /p/, and significantly longer for /k/ than for /t/. The only exception is that VOTs for medial unstressed /t, k/ are not significantly different from one another, although they manifest the same rank order as the other groups.

Results for KM differ somewhat from those for ES. Like ES, VOT for /p/ is significantly shorter than that for /t, k/ for each level of stress within each word position. Unlike ES, however, there is no significant difference in VOT for /t, k/ even though there is a tendency for mean VOT for /k/ to be slightly longer than for /t/.

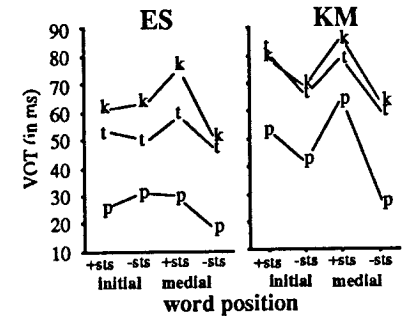


Fig. 2. VOT results presented as in fig. 1.

The acoustic data offer no support for VOT as a function of closure duration across an invariant devoicing gesture. Since closure duration for /t, k/ is equivalent for both speakers, one might expect VOT to be equivalent for both speakers under the invariant devoicing gesture proposal. However, ES shows VOT differences between /t, k/, while KM does not. It is, nevertheless, still possible that there is an invariant devoicing gesture for all stops. The devoicing gesture may simply be shifted in time with respect to oral closure for /t, k/. We examine these possibilities in the following sections.

3.2. Transillumination

3.2.1 Devoicing Gesture Duration (DGD)

ANOVAs show that there is also a significant effect of place of articulation on DGD for both speakers (fig. 3). However, a clear pattern of results does not emerge unless DGDs for lingual stops are considered as a group separate from labial stops. For ES, DGDs for /t/ are significantly longer than those for /k/ regardless of stress or word position. For KM, DGDs for /t, k/ only differ significantly from one another in medial stressed syllables, although mean DGD for /t/ is longer than for /k/ for each condition.

DGDs for labial stops in general appear to differ from DGDs for lingual stops. In word-initial position, differences between stressed and unstressed DGDs for labial stops are small and non-significant, but are comparatively large and significant for the lingual stops. Within stress categories mean DGD for

initial stressed /p/ is somewhat similar to that for medial stressed /p/, while DGDs tend to be longer in initial than in medial position regardless of stress for lingual stops. The aforementioned patterns are especially evident for KM.

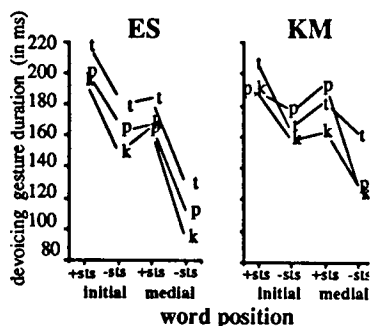


Fig. 3. DGD results presented as in fig. 1.

3.2.2. Peak Glottal Magnitude (PGM)

PGM (i.e., the greatest distance between the vocal folds during the devoicing gesture) results are similar to DGD results for both speakers in that there is a significant effect of place of articulation on PGM for both speakers and in that labial stops behave somewhat differently than lingual stops (fig. 4). For ES, PGM is always significantly greater for /t/ than for /k/. For KM, PGM is only significantly greater for /t/ than for /k/ in medial unstressed syllables; otherwise, they are equivalent in magnitude.

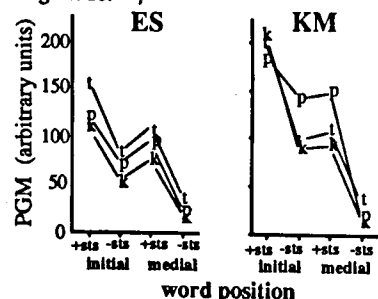


Fig. 4. PGM results presented as in fig. 1.

The present results show that there is no single invariant devoicing gesture for stops across place of articulation. Indeed, it seems that the devoicing gesture may be influenced by both the supralaryngeal

constriction location as well as the primary supralaryngeal articulator. Specifically, it appears that devoicing gestures are generally sensitive to whether the supralaryngeal constriction is produced with the lips or the tongue since the data suggest that stress and word position have different effects on labial versus lingual stops.

It seems possible that VOT differences among the lingual stops arise from differences in the duration and magnitude of the devoicing gesture since ES shows consistent differences for /t/ versus /k/ for both DGD, PGM and VOT, and since KM shows no difference in DGD, PGM or VOT for /t/ versus /k/. However, such a relationship seems especially doubtful since one might expect larger devoicing gestures to give rise to longer VOTs, whereas just the opposite result obtains for ES. In any case, these findings suggest that the timing of oral and laryngeal gestures must play a crucial role in VOT since variations in neither oral nor laryngeal gestures alone can account for the observed VOT patterns.

3.3. Interarticulator Timing

The coordination of laryngeal and supralaryngeal gestures has been intimately linked with VOT [4]. Here we consider the coordination between two pairs of articulatory events associated with the beginning and the end of voiceless stop-related gestures (namely, the interval from oral closure to the onset of glottal opening and the interval from oral release to the onset of glottal adduction) in order to determine whether the relationship between either of these events covaries with VOT.

3.3.1. Closure to Onset of Glottal Opening (C-OGO)

There is a significant effect of place of articulation on C-OGO for both speakers (fig. 5). For ES, OGO always occurs significantly later for /k/ than for /p, t/, but only occurs significantly earlier for /t/ than for /p/ in initial stressed and medial unstressed syllables. For KM, C-OGOs for the lingual stops are not significantly different from one another in any word position or for any stress category. There is no clear pattern for labial stops.

The C-OGO results closely mirror the patterns found for DGD and PGM suggesting that the larger the devoicing gesture, the earlier it begins relative to closure. When considering the labial stops in conjunction with the lingual stops, it becomes even more clear that the onset of the devoicing gesture does not simply shift in time relative to oral gestures to achieve a specific VOT. Rather, C-OGO is related to the size of the devoicing gesture. In fact, even the mean C-OGO data closely follow the same rank order as for DGD and PGM.

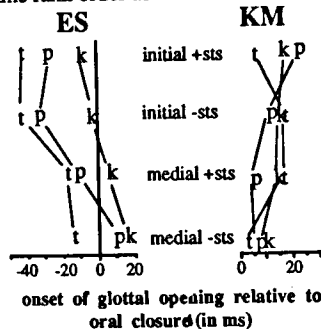


Fig. 5. Results for the interval from oral closure (at 0ms) to the OGO (the points plotted) presented as in fig. 1.

3.3.2. Release to Onset of Glottal Adduction (R-OGA)

There is a significant effect of place of articulation on R-OGA for both speakers (fig. 6). For ES, mean OGA relative to release occurs earliest for /p/, latest for /k/, and intermediate for /t/; this effect is significant except in medial unstressed position where R-OGAs for /t, k/ are not significantly different from one another.

Like ES, the OGA for KM always occurs significantly earlier for /p/ than for /t, k/ in both word positions and for both stress categories. Unlike ES, however, OGAs for /t/ only occur significantly earlier than for /k/ in medial stressed position; otherwise, R-OGAs for /t, k/ do not differ significantly.

R-OGA results are practically identical to the corresponding VOT results. Specifically, the earlier the OGA, the shorter the VOT for all stop categories (cf. fig. 2). Thus it appears that R-OGA is responsible for differences in VOT, and not variations in closure duration plus an invariant devoicing

gesture, or differences in the size of the devoicing gesture.

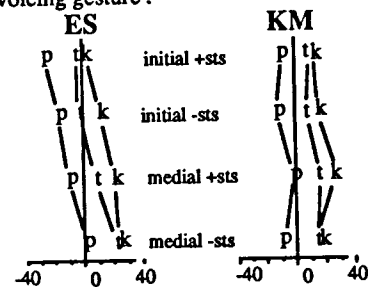


Fig. 6. Results for the interval from oral release (at 0ms) to the OGA (the points plotted) presented as in fig. 1.

Finally, it is important to note that the OGA is an active gesture rather than a passive aeromechanical consequence of oral release. Since both oral release and the OGA are controlled by muscular forces, it follows that oral release and the OGA are actively timed relative to one another. Thus it is possible that VOT differences are simply a byproduct of, rather than the motivation for, the timing of OGA relative to oral release.

REFERENCES

- [1] Crystal and House A., (1988) Segmental durations in connected speech signals: Current results. *JASA*, 83(4): 1553-1573.
- [2] Docherty, G. J. (1989) *An experimental phonetic study of the timing of voicing in English obstruents*. University of Edinburgh unpublished Ph.D. Dissertation.
- [3] Fischer-Jorgensen, E. (1980) Temporal relations in Danish tautosyllabic CV sequences with stop consonants. *Annual Report Institute of Phonetics University of Copenhagen*, 14:207-261.
- [4] Lisker, L. & Abramson, A. (1967) Some effects of context on voice onset time in English stops. *Language and Speech*, 10:1-28.
- [5] Stevens, K. N., & Klatt, D. H. (1974) The role of formant transitions in the voice-voiceless distinction for stops. *JASA*, 55: 653-659.
- [6] Weismer, G. (1979) Sensitivity of voice onset measures to certain segmental features in speech production. *Journal of Phonetics*, 7: 194-204.
- [7] Weismer, G. (1980) Control of the voicing distinction for intervocalic stops and fricatives: some data and theoretical considerations. *Journal of Phonetics*, 8: 427-438.
- [8] Zue, V. (1976) *Acoustic characteristics of stop consonants*. Indiana University Linguistics Club.

Work supported by NIH Grant DC 00121 to Haskins Laboratories and the University of Michigan.

DEVOICING OF JAPANESE /u/: AN ELECTROMYOGRAPHIC STUDY

Zyun'ichi B. Simada

Department of Physiology, School of Medicine, Kitasato University
Sagamihara, Kanagawa, Japan

Satoshi Horiguchi*, Seiji Niimi, Hajime Hirose

Research Institute of Logopedics and Phoniatrics, University of Tokyo
Bunkyo, Tokyo, Japan

ABSTRACT

Both cricothyroid and sternohyoid muscle activity was examined in a speaker of Tokyo Japanese with respect to devoicing of the vowel /u/. Sternohyoid activity, always with the cricothyroid activity suppressed, was related to implementation of a syllable of "stop consonant + u". However, this peculiar pattern of activity disappeared for the devoiced /u/ after an affricate or fricative. These results suggest that /u/ tends to be reduced or deleted in the nonplosive environment.

1. INTRODUCTION

Japanese has often been cited as an example of a language having voiceless or devoiced vowels [3]. They would sound unnatural if they were pronounced as voiced. For example, in two-mora words, the close vowel /i/ or /u/ must be devoiced when it is unaccented and occurs between voiceless obstruents. But the vowel devoicing in these cases is said to result from distinct articulatory processes. The difference between devoicing processes was first pointed out by Sakuma [4], and clarified in part by Han spectrographically [1]. Kawakami [2] summarizes as follows: the /pi, pu, ki, ku, cyu, syu/ followed by a voiceless consonant are simply devoiced, but the /ci, cu, si, su, hi, hu/ syllables usually do not manifest any voiceless vowels. Accordingly, the event as called "vowel devoicing" collectively is assumed to fall

*Current affiliation: Department of Neurootology, Tokyo Metropolitan Neurological Hospital, Hucyuu, Tokyo, Japan.

into two classes: (1) a devoiceable vowel is weakened or becomes voiceless when it appears after a stop consonant, whereas (2) the vowel usually is deleted and reduced to a mere consonantal lengthening after an affricate or fricative consonant (the /h/ in /hi/ or /hu/ is generally pronounced as fricative).

It seems difficult to distinguish acoustically between the simple devoicing and the entire deletion. We have tried to look into the physiological mechanism underlying this difference and report in this paper on a speaker who had a different control over the cricothyroid (CTh) and sternohyoid (SH) muscles between the classes 1 and 2.

2. METHOD

2.1. Subject

A male speaker of Tokyo Japanese, who has lived from the teens in a city on the outskirts of Tokyo, served as subject. The subject was one of the authors.

2.2. Speech Material

The electromyographic (EMG) experiment was carried out twice on different days, but here we will discuss only the data concerning the vowel /u/ obtained from Experiment 1. The words tested were of a form of /Cuki/ where C was /k, c, s, or zero consonant/ (/c/ stands for [ts]). Thus the following words

kuki^ 'stalk'
cuki^ 'moon'
suki 'plow'
uki 'float'

were tested in the frame sentence *ii ... no yo^o desu* '(it) looks like a nice ...'. These words are all ordinary, and the /u/ after a consonant is devoiceable. The

words *kuki* and *cuki* are accented on the second syllable, but we asked the subject to utter them with no accent (The syllable immediately preceding the particle *no* can be unaccented; the symbol ^ indicates that the syllable marked with it has an accent. For detailed Japanese phonology, see Vance [6]). Moreover, the voiced counterparts

kugi 'nail'
cugi 'patch'
sugi 'Japanese cedar'
ugi (nonsense word)

were included as a control in the sentence list.

2.3. Data Recording and Analysis

EMG activity was derived from the CTh and SH muscles by using bipolar

hooked-wire electrodes. The subject was asked to produce a total of 13 to 14 repetitions of each sentence, and recorded on a PCM tape recorder together with the EMG signals. All signals were digitized with 12-bit resolution and stored on an NEC desktop computer. On digitization, the EMG signals, after being full-wave rectified, integrated over a period of 5 ms, and finally low-pass filtered, were sampled at a rate of 1 kHz. The audio signal was sampled at a rate of 5 kHz.

The data files of interest were transferred to a Hewlett-Packard computer to calculate an ensemble average of the integrated EMG signals and to determine the line-up point necessary for averaging.

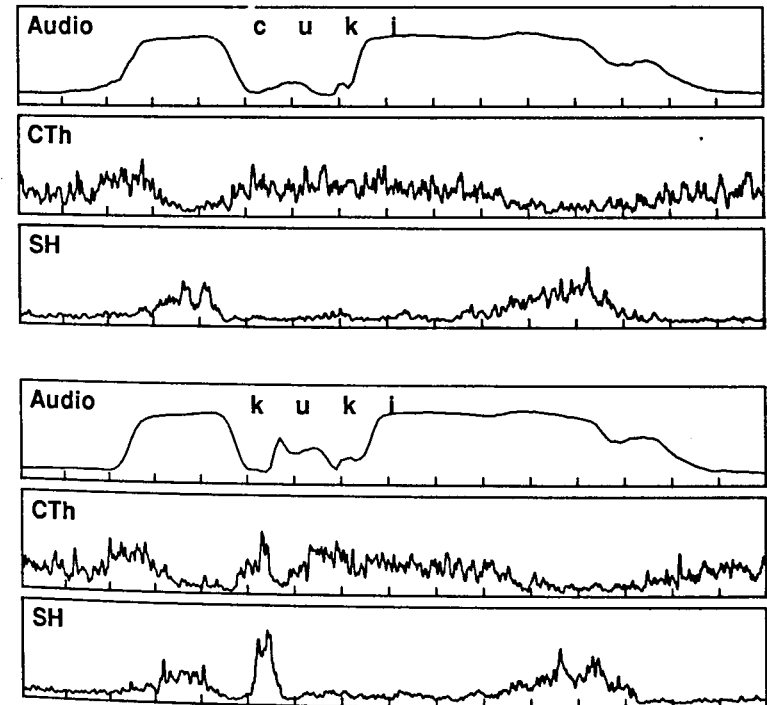


Figure 1. Cricothyroid (CTh) and sternohyoid (SH) activity for the test words *cuki* (upper panel) and *kuki* (lower panel). The time axis is marked off at every 100-ms interval.

3. RESULTS

3.1. CTh and SH Activity

We were able to classify the CTh activity into two patterns depending on whether or not a short suppression was present in its activity. One was a pattern for the words *uki*, *ugi*, *cuki*, and *suki* which showed no apparent CTh suppression; the other pattern was seen for the words *kuki*, *kugi*, *cugi*, and *sugi* which showed a clear suppression in the CTh activation. Figure 1 illustrates an example of the results. It compares the CTh and SH EMG curves for the class 2 word *cuki* (upper panel) and the class 1 word *kuki* (lower panel), respectively. The EMG curves are shown here in a form of ensemble average of twelve tokens obtained from the same utterance type. Each curve is aligned with reference to the line-up point which indicates the instant of release of the second-mora [k] of the words.

For the affricate *cuki*, the CTh begins to increase its activity before the first-mora /cu/, and it remains rather active during the voiceless period of /cu/ and until the occurrence of an accent in yo^o. This activity is typical of Tokyo Japanese and has been viewed as contributing to manifestation of the fundamental tone of a phrase with no unaccented syllables [5]. In contrast to *cuki*, we can see a short suppression of CTh activity for the plosive *kuki*. The suppression occurs immediately after the release of articulatory closure of the first-mora /k/.

It is noteworthy here that this CTh suppression is concurrent with a burst of SH activity which continues for some 90 ms. The SH activity is overlapped with that of the CTh. But a closer inspection reveals that there is a gap of timing between their activity; the SH is activated about 20 ms later than the CTh, and does not cease its activity sooner after the CTh becomes rapidly inactive. The same pattern of activity in both muscles was observed regularly for the tokens of *kugi*, *cugi*, and *sugi*.

However, we could not see the SH activity in question at all for the other three words *uki*, *ugi*, and *suki*, and neither could we identify any CTh activity accompanied by an evident suppressive phase.

3.2. Acoustic Findings

The /u/ of *kuki* was not entirely devoiced in all the tokens; there were observable minute oscillations of the vocal folds with the exception of two tokens. On the other hand, we could not recognize any vocal oscillations for the words *cuki* and *suki* from their acoustic waveforms.

4. DISCUSSION

The characteristic pattern of activity observed from both CTh and SH shares a common syllabic structure of "obstruent + u". This was not exceptional to the word *kuki*, either. Consequently, we assume that an accentual phrase calls for SH activity when it begins with this type of syllable, and this SH activity is linked with a short suppression of CTh activity. Our assumption can be supplemented by other findings: the words *uki* and *ugi*, which do not begin with an obstruent consonant, never showed any such SH activity, and in addition, the SH basically behaved reciprocally against the CTh in the present speaker. In any case we cannot account for the devoicing process of class 1 by the peculiar EMG pattern we observed.

On the other hand, the SH activity under discussion disappeared for the words *cuki* and *suki*, which correspond to the class 2 devoicing. This was in sharp contrast to the voiced counterparts *cugi* and *sugi* where the vowel /u/ was not devoiced. Following our assumption above, this contrast is due to a kind of neural mechanism as follows: the vowel /u/ does not manifest itself entirely and the resultant sequence of consonants, such as [ck] or [sk], does not require any more SH contraction; this vanishment of the SH burst in turn releases the CTh from a suppressive phase.

5. SUMMARY

Our EMG findings from the cricothyroid and sternohyoid muscles suggest that the devoiceable /u/ after an affricate or fricative consonant is not merely devoiced, but rather tends to be reduced or deleted. This tendency is quite consistent with Han's observation [1]. Finally, we feel that we should make an additional measurement of laryngeal or tongue movements to clarify the

difference between the processes of vowel devoicing.

We are grateful to Takashi Katakura for technical assistance in transmitting data files to the Hewlett-Packard computer.

6. REFERENCES

- [1] HAN, M. S. (1962), "Unvoicing of vowels in Japanese", *The Study of Sounds*, 10, 81-100.
- [2] KAWAKAMI, S. (1977), "*Nihongo onsei gaiseicu* [An overview of Japanese speech sounds]", Tokyo: Oohuusyua.
- [3] LADEFOGED, P., MADDIESON, I. (1990), "Vowels of the world's languages", *Journal of Phonetics*, 18, 93-122.
- [4] SAKUMA, K. (1929), "*Nihon onseigaku* [Japanese phonetics]", corrected edition, Tokyo: Kyoobunsysa.
- [5] SIMADA, Z. B., HORIGUCHI, S., NIIMI, S., HIROSE, H. (1989), "Tookyoo no oncyoo to kanren sita kootookin no kacudoo [Laryngeal muscle activity related to tone patterns of Tokyo Japanese]", *IEICE Technical Report, Speech SP88-163*, 88, 59-64.
- [6] VANCE, T. J. (1987), "*A n introduction to Japanese phonology*", Albany: State University of New York Press.

IS SUBGLOTTAL PRESSURE A CONTRIBUTING FACTOR TO THE INTRINSIC F0 PHENOMENON?

E. Viikman¹, I. Raimo² and O. Aaltonen²

¹Department of Otolaryngology and Phoniatrics, University of Oulu

²Phonetics Department, University of Turku

ABSTRACT

The hypothesis that compensation for lower loudness of high vowels (/i, u/) in speech might contribute to the higher intrinsic F0 of these vowels in comparison with low vowels (/a, æ/) was tested. F0, intensity and subglottal (oral) pressure were measured in two tasks. In the first the subjects (n=2) produced the test word /pV:ppV/ (V=/i, u, æ, a/) embedded in a carrier phrase. The pressure measurements showed highest pressure values for the vowel /u/ for both two subjects. In the second task the subjects read a /pV:ppV/ word list and tried to maintain the same SPL of the long vowel through different vowels. The results showed that compensation for the SPL differences between vowels produced greater intrinsic F0 variation than in normal speech. However, the subglottal pressure differences were too small to explain the differences in the F0 values.

1. INTRODUCTION

The intrinsic F0 of vowels, a vowel-specific variation of F0 in comparable contexts, is a well-known phenomenon. The physiological background of this phenomenon remains partly unclear. Our earlier studies suggested that one important factor in this respect is the vowel-specific activity of the cricothyroid muscle activity. It does not, however, exclusively explain the vowel intrinsic F0 variation. Changes in the vertical tension of the vocal folds has been found to be one additional factor in the production of the intrinsic F0 phenomenon. Acoustical explanations have been rejected [1, 2].

Vowel intrinsic F0 variation has been reported to be present even in esophageal speech [3]. This might imply a

sub(pseudo)glottal pressure-dependent control mechanism.

Subglottal pressure can affect the fundamental frequency in normal voice production [e.g. 4]. The present study is aimed at testing the hypothesis that compensation for lower loudness of high vowels (/i, u/) in speech might contribute to the higher intrinsic F0 of these vowels in comparison with low vowels (/a, æ/).

2. SUBJECTS AND METHODS

The subjects were two male native speakers (IR, OA) of Finnish without any known voice problems.

In the first task the subjects produced the test words in randomized order /pV:ppV/ (V=/i, u, æ, a/) embedded in a carrier phrase (/sano 'pV:ppV ta:s; "Say /pV:ppV/ again!") (n=25 for each vowel). In the second task the subjects tried to maintain the same sound pressure level (SPL) of the long vowel through different vowels by monitoring the display of an SPL meter (B & K 2209). Due to difficulties in adjusting SPL adequately the carrier phrase could not be used in the second experiment. /pV:ppV/ words were read in the following order: V=/a, i, u, æ/ (n=30 for each vowel).

The acoustical samples were recorded using a microphone (JVC MD 247) (distance 30 cm) and a tape recorder (JVC CD 1635 MARK II). F0 peak values of the vowels of the first stressed syllable were analysed using a microcomputer-based analysis program (ISA). The subglottal pressure was estimated from the intraoral pressures (F-J Manophone) during /p/-consonants obtained from a tube (diameter about 3 mm) placed between the lips [a.m. Löfqvist et al. 5]. SPL peak values (F-J Intensity Meter) and subglottal (oral)

pressure values were measured from calibrated plotted recordings (Siemens Oscillomink L). Pressure values were measured at two separate points: the peak for 1) the first /p/(point a) and 2) for /pp/(point b).

The results obtained are represented by arithmetic means (X) and standard deviations (SD). Statistical tests were carried out using Student's t-test.

3. RESULTS

The results for both experimental conditions are shown in Figs. 1 (subject IR) and 2. (subject OA).

The results of the first experiment with a carrier phrase showed a normal vowel intrinsic F0 pattern for both subjects (/i,u/ > /æ, a/). Also the SPL values obtained showed expected patterns (/i, u/ < /æ, a/). The pressure values for the first measuring point (a in Figs.) showed significant vowel-specificity only for subject IR. In this case the pressure for the vowel /a/ tended to be lowest. However, the second measuring point (b in Figs.) showed statistically significantly higher pressure values for the vowel /u/ (IR: p=0.49 kPa; OA: p=0.60 kPa) than for other vowels /i, æ, a/ (respectively, IR: p=0.40 kPa, 0.39 kPa, 0.40 kPa; OA: p=0.55 kPa, 0.55 kPa, 0.55 kPa).

The pressure values in the second experiment showed more vowel-specificity for both subjects than in the first "normal" condition. The subglottal pressures measured at point b for the vowel /u/ (IR: 5.8 kPa; OA: 7.3 kPa) were significantly higher than for the vowels /i, æ, a/ (IR: 4.4 kPa, 3.3 kPa, 3.0 kPa; OA: 6.9 kPa, 5.8 kPa, 6.1 kPa, respectively).

As can be seen in Figs. 1 and 2 the equalization of the SPL level between the vowels was not a simple task. From the point of view of the present study, however, the fact that the SPL pattern could be changed (/i, u/ > /æ, a/) is important. As compared to the first part of the study the range of intrinsic F0 variation grew in the second part. This was exclusively due to a drop in F0 values of /æ/ and /a/. For both subjects the F0 of vowels /i/ and /u/ did not change significantly even though both pressure and intensity values for these vowels were significantly higher.

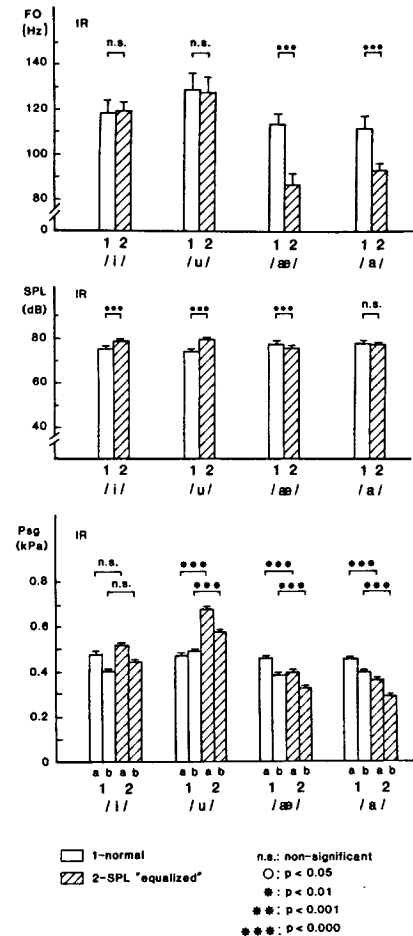


Figure 1. Average (X±SD) F0, SPL and subglottal pressure (Psg) values for four vowels (subject IR).

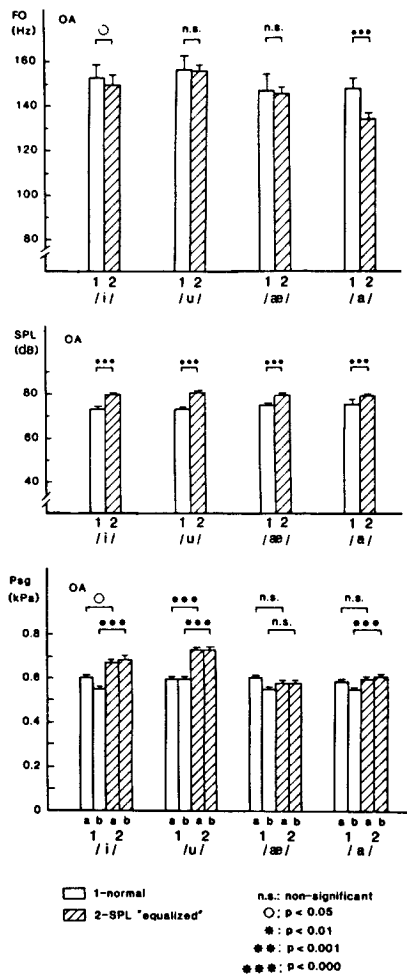


Figure 2. Average ($\bar{X} \pm SD$) F0, SPL and subglottal pressure (p_{sg}) values for four vowels (subject OA).

4. DISCUSSION

The subglottal pressure for the highest vowel /u/ was significantly higher than that for the other vowels for both subjects in the experiment in which test words were embedded in a carrier phrase. This might imply a compensation for the low loudness of the vowel /u/. However, the estimated pressure difference between the average values for /u/ and /i, æ, a/ was only 0.05 kPa. From physiological studies it is known that in low chest register phonation a pressure rise of 0.1 kPa causes an F0 rise of 5 Hz (see e.g. [4] for a review). Thus, it seems that only a few hertz of intrinsic F0 variation could be explained by pressure differences.

In the second part of the present study in which the SPL of the long vowel of the word /pV:ppV) was kept as constant as possible the pressure differences were greater (approximately 0.1 kPa). However, even in this case the F0 difference between the vowels /u/ and /æ, a/ cannot be explained on this basis. It is obvious that a laryngeal contribution is necessary (c.f. [1]).

Usually F0 and intensity are known to be closely correlated (e.g. [4]). However, in the second condition of the present study the higher intensity and pressure values co-occurred with lower F0 than in the first part. Two tentative explanations can be suggested. Firstly, the intrinsic F0 phenomenon is under keen cortical control and the intrinsic F0 of /i/ and /u/ represent the "correct" values. Now that the SPL was not allowed to change deliberately the situation was unnatural from the point of view of the low vowels /æ/ and /a/, which caused a reduction in laryngeal activity and, consequently, a drop in F0. The second possibility is that the finding was caused by the difference in the test tasks. I.e. the reading style of a natural sentence is produced with a higher F0 than the list of separate and equally stressed words. In this case the F0 of /æ/ and /a/ would reflect the normal values of the task. Correspondingly the F0 of /i/ and /u/ would reflect the increased effort needed to reach the SPL in question. Thus the same F0 values for /i/ and /u/ in the two test conditions would be coincidental. Further studies are needed to distinguish between these two possibilities suggested by this preliminary study.

It can be concluded that equalization of the output SPL of vowels has an influence

on the vowel intrinsic F0 variation. This is also reflected in the subglottal pressure level, but the difference in the F0 patterns can not be explained on the basis of subglottal pressure differences alone. A laryngeal contribution is necessary.

5. REFERENCES

- [1] VILKMAN, E., AALTONEN, O., RAIMO, I., ARAJÄRVI, P., OKSANEN, H. (1989), "Articulatory hyoid-laryngeal changes vs. cricothyroid muscle activity in the control of intrinsic F0 of vowels." *J. Phonetics*, 17, 193-203.
- [2] VILKMAN, E., AALTONEN, O., LAINE, I., RAIMO, I. (1991), "Intrinsic pitch of vowels - a complicated problem with an obvious solution". *Proc. of the 6th Vocal fold physiology Conference*, Stockholm 1989, Raven Press, in press.
- [3] GANDOUR, J., WEINBERG, B. (1981), "On the relationship between vowel height and fundamental frequency: evidence from esophageal speech." *Phonetica*, 37, 11-22.
- [4] TITZE, I.R. (1989), "On the relation between subglottal pressure and fundamental frequency in phonation." *J. Acoust. Soc. Am.*, 85, 901-906.
- [5] LÖFQVIST, A., CARLBORG, B., KITZING, P. (1982), "Initial validation of an indirect measure of subglottal pressure during vowels." *J. Acoust. Soc. Am.* 72, 633-635.

MODELLING OF SPEECH MOTOR CONTROL AND ARTICULATORY TRAJECTORIES

P. Perrier, R. Laboissière & L. Eck.

Institut de la Communication Parlée - URA CNRS n° 368
INPG-ENSERG, Grenoble, France.

ABSTRACT

In order to study motor control strategies in speech production, we propose to simulate the dynamical behaviour of speech articulators with a stiffness commanded distributed second order model, where the set agonist/antagonist is commanded as a whole. For [CVCV] utterances, inversion from the jaw movements to the corresponding stiffness commands is proposed using a guided algorithm of error backpropagation. We focused the analysis of our results on the ability of our model to predict target undershoot, and to detect hypo- and hyper-articulation strategies used by two speakers.

1. INTRODUCTION

The so-called *Equilibrium Point Hypothesis*, introduced by Asatryan & Feldman ([2]), suggests that skilled movements correspond to shifts in the equilibrium state of the motor system. In this framework, two major, quite different kinds of modeling have been developed: the λ model proposed by Feldman and his co-workers (see [2], [6] and [7]) and α model, first proposed by Bizzi [3] (see also [4]). According to the former the commanded variable is the *recruitment threshold* of the used muscle, whereas in the latter the muscle stiffness (corresponding to the muscle activation) is controlled. Both approaches yielded very appealing results for jaw or multi-joint arm movements ([4] and [7]). But in his clarification article [6] Feldman develops his argumentation against Bizzi's α model: this latter can actually neither explain movements occurring with a constant muscle activation level nor the absence of movement for a certain kind of variation of this activation; moreover stiffness cannot be centrally commanded since, due to afferent signals, this variable depends on the length of the muscle and then varies during the movement.

From our point of view, if the *agonist and*

antagonist muscles are considered as a whole, the concept of equilibrium point defined as an *equilibrium between agonist and antagonist stiffnesses* is functionally very appealing. Thus, in spite of everything, we proposed [10] to use a stiffness commanded distributed second order model for the set agonist/antagonist muscles considered as a specifically commanded whole; the advantage of this global modeling lies in the fact that it overcomes the main critics of Feldman against the stiffness model (see further). In order to test the validity of such an approach, we propose here to confront our model with data on jaw movements in the production of CVCV sequences. Inversion using an error back propagation algorithm is studied in order to infer the stiffness commands which allow the generation of suitable trajectories. The results are analysed in regard to the control strategies proposed by this technique for certain kinds of speech production.

2. OUR SECOND ORDER MODEL

According to the kinematics characteristics of skilled movements presented by Nelson [9], our model [10] (see fig.1) consists in a couple of springs, one for the agonist set and another for the antagonist one; these springs are linked by a material point, whose mass (m) is normalized to 1. The displacements of this point correspond to shifts from an equilibrium point of the system to another. This latter is called *target* of the movement. The successive targets (or equilibrium points) are determined by the ratio (η) between the stiffness (k_1 and k_2) of the two springs. These mechanical targets are directly linked to the underlying phonetic targets of the sequence: each vowel and each consonant correspond to a specific value of the ratio η .

Both springs have the same rest length x_0 . When the equilibrium point of the system is shifted, an unidirectional movement of

the material point occur. Let x be the spatial variable in the direction of the movement; the dynamical equation describing the system is then:

$$\frac{\partial^2 x}{\partial t^2} = -f \cdot \frac{\partial x}{\partial t} - (k_1 + k_2) \cdot x - (k_1 - k_2) \cdot x_0 \quad (1)$$

It is, of course, quite easy to notice from this equation that, given f , the values of k_1 and k_2 determine completely the trajectory. Because we consider the agonist and antagonist sets as a whole, we propose to command the model with two variables which act simultaneously on these sets:

- the *stiffness ratio* η which determine the equilibrium position of the target;
- the *cocontraction* K , corresponding to the global activation $k_1 + k_2$ of the set agonist/antagonist.

The major critics of Feldman [6] against the stiffness model don't apply to our approach: it is actually obvious that movement can occur without modification of the cocontraction level (with a reciprocal variation of k_1 and k_2), and that this level can be modified without change in the resulting stiffness ratio η and therefore without movement. Moreover the cocontraction level is not dependent on the length of each spring and can therefore be centrally commanded. This holds if we suppose a symmetrical modeling of the agonist and antagonist sets, which would induce a reciprocal lengthening/shortening on each of them.

The commands η and K vary theoretically by step between targets. However, in order to propose more realistic variations of the commands, we have smoothed the abrupt transitions of these signals by filtering them with a critical second order filter ($\tau=80$ ms). The duration of each step is explicitly commanded.

3. THE INVERSION APPROACH

Equation (1) describes the dynamics of the model, where K and η are the inputs and x is the output. The goal of the inversion procedure is to infer the time-varying functions $K(t)$ and $\eta(t)$ that generate the actual jaw displacement $x(t)$. Since equation (1) does not have constant coefficients, it is quite hard to derive an analytical solution to the general inversion problem. We applied then an iterative optimization procedure, essentially a gradient-descent technique, where we minimize a cost functional given by the squared error between the actual and the

model output signals integrated over the time interval of interest. We carry this optimization over the space of possible functions $K(t)$ and $\eta(t)$, with the constraints described in section 2.

The gradient of the cost functional can be obtained using the calculus of variations, but for a discrete version of (1) (see [5]) we obtain a formulation close to the *error backpropagation through time* [11]. Without truncation in time, this method give an exact gradient and the error cost tends asymptotically to a local minimum through the iterations. With good guesses for the initial state and some interactive control during the process (e.g. alternating the optimization of the duration and amplitude of the commands) we get reasonable results, like those shown in the following section.

4. RESULTS - DISCUSSION

4.1. Description of the corpus

The corpus consists of the utterance [zɛzzɔ] in Tunisian Arabic (what means: "he rewarded"). It is pronounced within a carrier sentence at two different speech rates ("normal" and "as fast as possible") and by two different native speakers. The movements of the jaw are considered here to be pertinent enough for a reliable description of the production strategies. They were recorded with a mandibular kinesiograph (K5AR), and sampled at 160 Hz (for more details see [1]).

4.2. Undershoot phenomenon in the inter-consonantal vowel

In the following we denote by $\Delta\eta$ the amplitude of the variation of η , and by $\Delta\tau$ the temporal percentage of the vocalic command within the total duration of the vocalic plus consonantal commands.

We tried to fit the output of our model to the jaw data from one speaker at the two rates. First of all, *the level of cocontraction K of the model was held constant*. Fig.2 shows the corresponding results :

- at normal rate, the spatial positions corresponding to the mechanical equilibrium points (called *ideal targets*) are reached for both consonants [z] and [zz], but a slight undershoot occurs for the vowel [ɛ]; $\Delta\eta$ is 0.49, and $\Delta\tau$ is 33%.
- at fast rate, the ideal targets are reached for the consonants, and we observe a very clear undershoot for the vowel; $\Delta\eta$ is 0.62, and $\Delta\tau$ is 34% .

At first glance, these results are satisfying:

through our inversion, the well-known *vocalic reduction* phenomenon due to speech rate increasing (see [8]) stands out. However the underlying command strategy here proposed seems to be unrealistic: $\Delta\eta$ increases in the case of vocalic reduction. This would mean that the speaker point to a further target to minimize the undershoot!!!

We adopt then the same fitting approach but with simultaneous optimization of K and η . The results (Fig.3) are more satisfying:

- at normal rate, K remains approximately the same as above for the consonant, but increases for the vowel production; all ideal targets are reached, $\Delta\eta$ is 0.41 and $\Delta\tau$ is 41%; it seems then that in order to prevent any influence of the consonantal context on the vowel, the speaker makes a particular effort for the vocalic articulation, corresponding to the increase of K.

- at fast rate, we observe a clear undershoot in the production of the vowel; we notice a reduction of the vocalic duration ($\Delta\tau=33\%$) and a decrease of the cocontraction level for the vowel production. The vocalic reduction could thus be explained through a credible strategy: the instruction "speak as fast as possible" induces in the speaker a *decrease in his articulation effort*, corresponding to a decrease of the cocontraction level for the vowel production. From this point of view this second inversion is very interesting. However we observe again an increase of $\Delta\eta$, whose value is 0.53. This can be explained by the fact that too many parameters (K, η , and the durations of each command step) have to be optimized at the same time for this simulation. In order to get a better inversion, we have to propose constraints on the respective evolutions of these parameters; for example, the constraint " $\Delta\eta$ must be the same for normal and fast rates" would consolidate the above assumed strategy of our speaker for vocalic reduction.

4.3. Hypo- and hyperspeech strategies

Our further point is to compare the production strategies used by two different native speakers for the same utterance. Fig.4 depicts the results of the inversion in the same conditions as just above. For both normal and fast rates, all ideal targets are reached for our second speaker: the cocontraction level increases strongly for the vowel production, and particularly at

fast rate; this speaker seems to increase his articulation effort when speech rate increases. This assumption corresponds to an audible characteristic of the speech signal: at fast rate this speaker cries out!!! We think so that our model provides a good tool to detect the phenomenon of hypo/hyperarticulation, as proposed by Lindblom ([8]), from the articulatory data. At fast rate, the first speaker tends to hypoarticulate whereas the second one tends to hyperarticulate.

5. CONCLUSION

By means of an error backpropagation technique we were able to fit available data on jaw movement to the output of a model consisting of agonist/antagonist pair of springs. The controlled variables in this model are the stiffnesses of the springs taken as a whole. In spite of Feldman's interesting criticisms against stiffness control for skilled movements, we showed that our model can explain known phenomena in speech production, namely vowel reduction and hypo/hyperarticulation strategies.

ACKNOWLEDGEMENTS

To Christian ABRY for his incitation on working in this direction, and to Jomaa MOUNIR for the data collection on jaw movements. The second author was supported by a scholarship from the National Research Council (CNPq), Brazil, under FIAS-CNPq file number 92.0208/88.6. He is also with the Aeronautics Technology Institute (ITA), São José dos Campos, Brazil.

REFERENCES

[1] ABRY C., PERRIER P. & JOMAA M. (1990), "Premières modélisations du timing des pics de vitesse de la mandibule" *Proceedings of the 18th J.E.P., Société Française d'Acoustique*, 99-102.
 [2] ASATRYAN D.G & FELDMAN A.G. (1984), "Functional tuning of the nervous system with control of movement or maintenance of a steady posture. I. Mecanographic analysis of the work of the joint on execution of a postural task." *Biophysics*, 10, 925-935.
 [3] BIZZI E. (1980), "Central and peripheral mechanism in motor control," in *Tutorials in motor behavior*, G.E. Stelmach & J. Requin (eds.), Amsterdam: North-Holland, 131-144.
 [4] COOKE J.D. (1980), "The organization of simple skilled movements," in *Tutorials in motor behavior*, G.E. Stelmach & J. Requin (eds.), Amsterdam: North-Holland,
 [5] ECK L. (1990), "Modélisation des gestes articulatoires," Mémoire de DEA, Institut National Polytechnique de Grenoble, France.

[6] FELDMAN A.G. (1986), "Once more on the equilibrium-point hypothesis (λ model) for motor control," *Journal of Motor Behavior*, 18, 1, 17-54.
 [7] FLANAGAN J.R., OSTRY D.J. & FELDMAN A.G. (1990), "Control of human jaw and multi-joint arm movements," in *Cerebral Control of Speech and Limb Movements*, G. Hammond (ed.), London: Springer-Verlag.
 [8] LINDBLOM B. (1990), "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Modelling*, W.J. Hardcastle & A. Marchal (eds.), London: Kluwer Academic Publishers.
 [9] NELSON W.L. (1983), "Physical principles for economics of skilled movements," *Biol. Cybern.*, 1-9.
 [10] PERRIER, P., ABRY, C. & KELLER E. (1989), "Vers une modélisation des mouvements du dos de la langue," *J. Acoustique*, 2, 69-77.
 [11] WILLIAMS R.J. & ZIPSER D. (to appear), "Gradient-based learning algorithms for recurrent connectionist networks" in *Backpropagation: Theory, Architectures and Applications*, Y. Chauvin & D.E. Rumelhart (Eds.), Hillsdale, NJ: Lawrence Erlbaum.

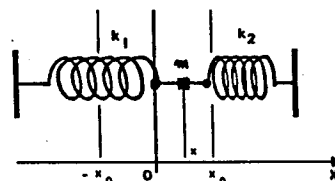


Figure 1: The distributed second order model.

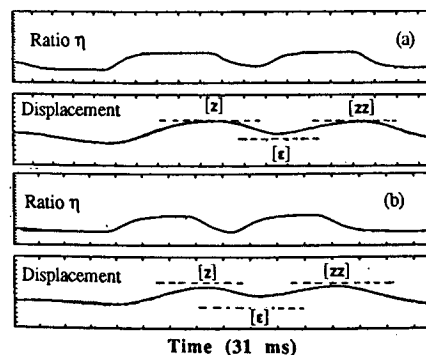


Figure 2: Stiffness commands and jaw displacement obtained for a constant cocontraction level; as regards the displacement, at the scale of the figure, there is no perceptible differences between model output and data; the dotted segments on the displacement curve correspond to the different inferred ideal targets (see text); (a)=normal rate, (b)=fast rate.

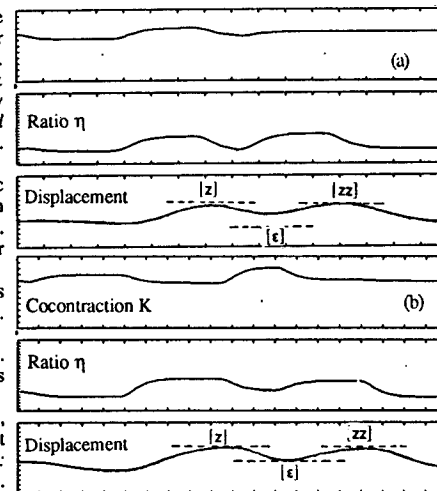


Figure 3: Stiffness commands and jaw displacement obtained for the first speaker; for comments see figure 2.

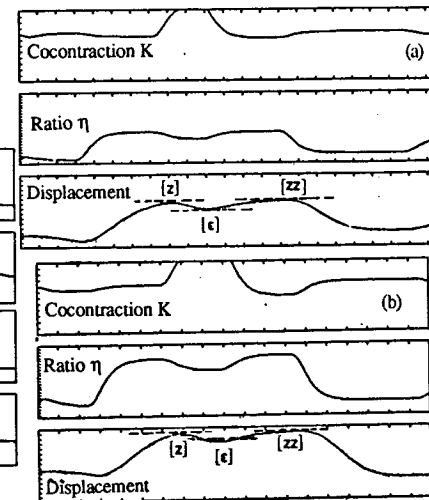


Figure 4: Stiffness commands and jaw displacement obtained for the second speaker; for comments see figure 2.

TRYING TO DETERMINE PLACE OF ARTICULATION OF PLOSIVES WITH A VOCAL TRACT MODEL

ALAIN SOQUET, MARCO SAERENS, AND PAUL JOSPA¹

Institut de Phonétique and IRIDIA
 Université Libre de Bruxelles - CP 110
 50, av. Franklin Roosevelt 1050 Bruxelles - Belgium

Abstract

We have recently used the Distinctive Regions and Modes theory ([12]), coupled with a neural controller ([15]), to produce an acoustic-articulatory inversion of a vocal tract model ([17]). This paper presents results on the possible detection of the place of articulation of plosives on the basis of this inversion scheme.

1 Introduction

Mrayati, Carré & Guérin ([12]; see also [3]) have recently presented a theory of speech production based on distinctive modes and on distinctive spatial regions along the vocal tract. This theory provides a framework for articulatory speech synthesis ([13]). It supplies relationships between the variations of the first three formants and the cross sectional areas of eight vocal tract regions of the model. Previous work has shown that such a-priori qualitative knowledge can be used to control and invert non-linear physical processes with a neural network ([15]). In this work, the relationships between the cross sectional areas of the regions and the formant variations are used to provide an acoustic-articulatory inversion of a vocal tract. Acoustic-articulatory inversion is a one-to-many nonlinear problem. It is usually managed by generating articulatory vectors in the articulatory space, and computing the corresponding acoustic parameters. Then, a look-up table can be constructed, providing the relationships between acoustic parameters and articulatory vectors ([11], [1], [7]).

A previous paper ([17]) has shown that a network is able to learn to invert the process, for the eleven French oral vowels. The addition of a constraint on the average volume

of the vocal tract allows the system to provide more realistic vocal tract shapes, and clearly improves the convergence rate of the network. These results have been extended to a 30-sections vocal tract by introducing a continuity constraint, and the inversion has been generalized to the vowel space ([18]).

In this paper, the inversion scheme is used to provide an articulatory gesture in the neighbourhood of plosives. This gesture is then analysed to locate the candidates for place of articulation.

Bailly et al. [2] are currently studying a similar, but more ambitious problem: They use Jordan's approach ([5]) to control Maeda's articulatory model ([9]).

2 The Vocal Tract Model

Vocal tract shapes are generated in the framework of the so-called Distinctive Regions and Modes theory ([12], [3]). The model involves an acoustical tube closed at one end (glottis), and open at the other (lips) (Figure 1). This model is based on the study of acoustical properties of vocal tract shapes, compared to those of a neutral uniform tube. For the three formants model, eight regions of different length (the distinctive regions) can be defined. Varying the mean cross sectional area of each of these regions induces specific and quasi monotonic formant variations. The eight regions will be denoted as -A, -B, -C, -D, and D, C, B, A.

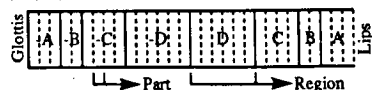


Figure 1: Vocal tract divided in 30 parts and 8 different regions.

The vocal tract is divided into thirty parts of equal length. Each part belongs to one region and has the qualitative behaviour of this region (See Figure 1). The cross sectional areas for the first region (-A) are scaled from 0.8 cm^2 to 3.0 cm^2 , and the remaining ones from 0.5 cm^2 to 15.0 cm^2 . The effective length of the acoustic tube has been set to 19 cm .

3 The Neural Controller

A neural network is used to provide the cross sectional areas to the vocal tract model, when the first three target formants are given as input (Figure 2). Standard back-propagation cannot be used directly for the controller because the optimal control parameters are not known.

Therefore, we use a specialized learning scheme based on an approximation of the back-propagated error that allows adaptive control with the neural network ([16]).

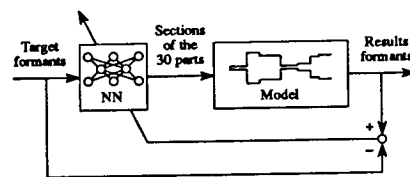


Figure 2: Architecture of the system.

To perform the inversion (see Figure 2), the following three steps are iterated until the vocal tract model produces the target formants:

1. The neural network is given target formant values.
2. The outputs of the network supply the vocal tract cross sectional areas, and the corresponding formant values are computed (we use an algorithm developed by [8]).
3. The difference between these values and the desired formants are used to correct the connection weights of the network with the modified version of back-propagation algorithm.

By training, the neural network learns to supply the correct cross sectional areas for the production of the target formant values.

The controller is a network with three layers (one hidden layer). Every unit of each layer is connected with the units of the adjacent

layers. There are three input units (corresponding to the first three formant values), ten hidden units, and thirty output units (corresponding to the thirty parts of the vocal tract).

The error used for the back-propagation algorithm in the neural network is composed of three terms: The difference between the actual and the target formants, a constraint on the average volume of the vocal tract, and a continuity constraint:

$$E = \sum_v \left[\sum_{i=1}^3 (F_i^v - F_i^{vd})^2 + k_1 \left[\left(\sum_{i=1}^{30} LA_i^v \right) - V_0 \right]^2 + k_2 \sum_{i=1}^{29} (A_i^v - A_{i+1}^v)^2 \right] \quad (1)$$

where the F_i^v are the formant values computed through the transfer function of the tube ([8]), the F_i^{vd} are the target formants, L is the length of a part, the A_i^v are the corresponding areas supplied by the network, $k_1 = 5 \cdot 10^{-5}$, $k_2 = 2 \cdot 10^{-3}$, and the average volume $V_0 = 85 \text{ cm}^3$.

This way, the network approximates the nonlinear mapping from the acoustic parameters (the three first formants) to the articulatory space (the cross sectional areas). The net provides one possible solution to this problem and, since it is a one-to-many problem, constraints are introduced in order to reduce the number of possible solutions. Hence, we observe that the different mapping obtained with different initial weights are quite similar.

4 Experiment

The network is first trained on the 11 French oral vowels (we use values published by [10]), then, the training set is generalized to the whole vowel space (see [18]).

After this training, the network approximates the nonlinear mapping from the acoustic parameters (the three first formant values) to the articulatory space (the cross sectional areas). The net is used to provide vocal tract shapes in the neighbourhood of consonant plosives. These shapes are then used to locate a possible constriction place. This allows us to establish whether there is a correlation between this constriction place and

¹ The following text presents research results of the Belgian National incentive-program for fundamental research in artificial intelligence initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. We would like to thank René Carré for helpful discussions, and Philip Miller for his help in the draft of the English text.

the real place of articulation of the plosive. Indeed, Mrayati, Carré & Guérin ([14]) claim that the different regions of the acoustic tube correspond to precise places in the vocal tract. For instance, labials are associated with the region A, dentals with the region B, and palato-velars with the regions C and D.

It is well known that important cues for identification of place of articulation of plosives are located in the formant transitions ([4]) and burst spectrum ([19]). In this work, we only take into account the formant values to realize the inversion.

Speakers. Two Belgian male subjects, native speakers of the French spoken in the Brussels area, and with university education were employed.

Recording procedure. The VCV items were recorded with a Studer A310 tape recorder in an anechoic room through a Neumann U88 microphone. They were sampled at 20 kHz with the Macspeech Lab software on a Macintosh II.

Items. The speakers were asked to produce VCV items, C being one of the six plosives [p, t, k, b, d, g] and V one of the five vowels [a, e, i, y, u]. There were $5 \times 6 \times 5 = 150$ items for each speaker. The sequence consisted of three blocks of 50 items in random order.

Acoustic analysis. The formant values were manually extracted with the Macspeech Lab software, at two different locations, for both adjacent vowels: The middle of the stable portion of the vowel (t_1) and the end of the vocalic transition (t_2). We were unable to extract these values for 4 items, the transitions being not detectable. The formants are provided as input to the network, which associates vocal tract configurations. Two different cues are computed on the vocal tube, a static cue, which simply corresponds to the sections at t_2 , and a dynamic cue, which is:

$$I_i(t_2) = \frac{A_i(t_2) - A_i(t_1)}{A_i(t_2) - A_{i \min}} \quad (2)$$

The place of articulation of the plosive is determined on the basis of the constriction deduced from the two different cues and the two adjacent vowels. The final decision is taken by a vote of the different knowledge sources.

Confusion matrix is presented in Table 1. We obtain 72.0% identification of classified places and 21.6% of ambiguous cases. Table 2 shows that the dynamic cue is more reliable than the static one, but provides more ambiguity.

There is indeed a correlation between the place of articulation of the plosive and the constriction of the tube. However, a more detailed analysis of the results shows a great influence of the context on the behaviour of the tube. This is not surprising provided that cues for place of articulation are known to be context-sensitive ([20]; [6]). For instance, for context [i], the constriction of the vowel is palato-velar, and remains during the transition. In this particular case, the dynamic cue is much more reliable.

Table 1: Total results for the 296 VCV items.

produced identified	labial	dental	velar
labial	75	13	24
dental	5	47	0
velar	6	17	45
ambiguous	12	23	29

Table 2: Results for static cue (upper table) and dynamic cue (lower table).

produced identified	labial	dental	velar
labial	57	10	28
dental	1	27	0
velar	6	15	32
ambiguous	34	48	38

produced identified	labial	dental	velar
labial	51	8	11
dental	4	26	1
velar	3	8	35
ambiguous	40	58	51

5 Conclusion

Results show a correlation between the region of constriction of the acoustic tube and the place of articulation of the plosive. Nevertheless, we observe a strong variability with the vocalic context, which is not surprising given the simplicity of the defined cues. The acoustic tube has a complex dynamic behaviour, which cannot be accounted for by introducing such simple articulatory cues. The definition of context-dependent cues could achieve more accurate results.

Bibliography

- [1] B.S. Atal, J.J. Chang, M.V. Mathews, and J.W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. *J. Acoust. Soc. Am.*, 63(5):1535-1555, 1978.
- [2] G. Bailly, M. Bach, R. Laboissière, and M. Olesen. Generation of articulatory trajectories using sequential networks. In *Proc. of the ESCA Workshop on Speech Synthesis*, pages 67-70, Autrans, 1990.
- [3] R. Carré and M. Mrayati. New concept in acoustic - articulatory - phonetic relations. Perspectives and applications. In *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, pages 231-234, Glasgow, 1989.
- [4] F. Cooper, P. Delattre, A. Liberman, J. Borst and L. Gerstman. Some experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Am.*, 24(6):597-606, 1952
- [5] M. I. Jordan. Learning to control an unstable system with forward modeling. In *Neural Information Processing Systems 2*, pages 324-331. Touretzky, D S, 1990.
- [6] D. Kewley-Port. Measurement of formant transitions in naturally produced stop consonant-vowel syllables". *J. Acoust. Soc. Am.*, 72(2):379-389, 1982.
- [7] N.J. Larar, J. Schroeter, and M.M. Sondhi. Vector quantisation of the articulatory space. *IEEE Transactions on Acoustics Speech and Signal Processing*, 36(12):1912-1918, 1988.
- [8] J. Liljencrants and G. Fant. Computer program for vt-resonance frequency calculations. Technical Report 4/1975, Stockholm: Speech Transmission

Laboratory - Quarterly Progress and Status Report, 1975.

- [9] S. Maeda. Compensatory articulation during speech: Evidence from the analysis and the synthesis of vocal-tract shapes using an articulatory model. In *NATO Meeting*, Bonnace, 1989.
- [10] R. Majid, L. J. Boë, and Perrier P. Fonctions de sensibilité, modèle articuloire et voyelles du français. In *Actes des 15èmes Journées d'Etude sur la Parole*, pages 59-63, Aix en Provence, 1986.
- [11] P. Mermelstein. Determination of the vocal-tract shapes from measured formant frequencies. *J. Acoust. Soc. Am.*, 41(5):1283-1294, 1967.
- [12] M. Mrayati, R. Carré, and B. Guérin. Distinctive regions and modes: A new theory of speech production. *Speech Communication*, (7):257-286, 1988.
- [13] M. Mrayati and R. Carré. Speech synthesis based on a vocal tract region theory. In *Proc. European Conf. Speech Communication and Technology*, pages 172-175, Paris, 1989.
- [14] M. Mrayati, R. Carré, and B. Guérin. Un nouveau modèle acoustique de production de la parole. In *13th International Congress on Acoustics*, pages 373-376, Yugoslavia, 1989.
- [15] M. Sauerens and A. Soquet. A neural controller. In *Proc. First IEE Int. Conf. Artificial Neural Networks*, pages 211-215, London, 1989.
- [16] M. Sauerens and A. Soquet. Neural controller based on back-propagation algorithm. *IEE Proceedings-F*, 138(1):55-62, February 1991.
- [17] A. Soquet, M. Sauerens, and P. Jospa. Acoustic-articulatory inversion based on a neural network controller of a vocal tract model. In *Proc. of the ESCA Workshop on Speech Synthesis*, pages 71-74, Autrans, 1990.
- [18] A. Soquet, M. Sauerens, and P. Jospa. Acoustic-articulatory inversion based on a neural network controller of a vocal tract model: Further results. In *Proc. of the ICANN*, Helsinki, 1991.
- [19] K.N. Stevens, and S.E. Blumstein. Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.*, 64(5):1358-1368, 1978.
- [20] K. Suomi. The vowel-dependence of gross spectral cues to place of articulation of stop consonants in CV syllables. *Journal of Phonetics*, (13):267-285, 1985.

COMPLEX NATURE OF THE SEEMINGLY SIMPLE VOCAL FOLD CYCLE

J. Pešák

Neurological Clinic, Medical Faculty, Olomouc
Czechoslovakia

ABSTRACT

Vocal fold vibration by phonation is currently viewed as a passive, myoelastic-aerodynamic process [1] of simple opening and closing of the glottal chink at fundamental frequency. However, vibration recorded directly from the thyroid cartilage could prove this seemingly simple vocal fold cycle to be more complex and associated with a probably reflex event.

1. INTRODUCTION

Laryngeal anatomy does seem to be simple at first sight, see Fig. 1. As such it could

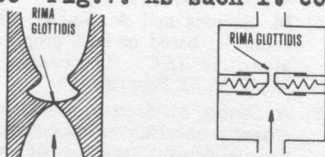


Fig. 1 Simplified laryngeal section

in the past do with the substitute model of a kind of Ewald's pipe, see Fig. 2.

2. LARYNX BY PHONATION

Phonation is measured by methods focusing on the behaviour of the proper glottal chink opened by the exhaled air stream as seen in Fig. 3, large arrow. The lateral opening along the axis y , see arrows, is well evident in the electroglotte -

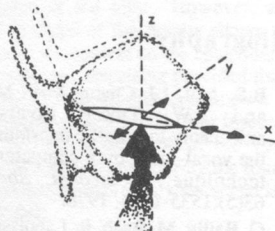


Fig. 3 Idealized glottal chink

graphic recording (EGG) of Fig. 4 below. This is matched by actualized frame sectors in the upper part of Fig. 4

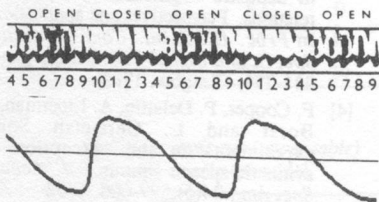
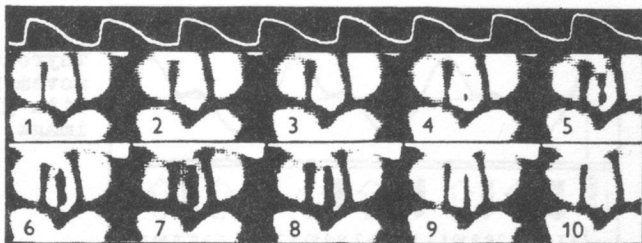


Fig. 4 Simultaneous recordings of EGG signal and glottal chink image

according to Hirose (Fig. 5) [2]. Recordings belonging to one vocal fold cycle are numbered 1 to 10. The course of EGG impedance changes informs about the way the chink is opened or closed. Other information can be obtained from a simultaneous recording of thyroid cartilage vibration. An acceleration recorder can be placed on the thyroid cartilage,

Fig.5 Glottal chink image according to Hirose



with an example in Fig.6. [3]

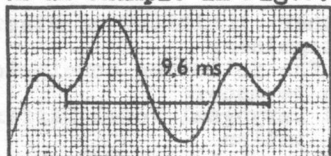


Fig.6 Course of thyroid cartilage amplitude

The current conception would expect an uncomplicated process corresponding to simple opening or closing of the chink. The measured course, however, describes a complex event with two oscillations within one cycle.

3. ACCOUNT OF THE COMPLEX EVENT

Let us inspect the larynx more closely, noticing the two ligaments joining the arytenoid with the thyroid cartilage. The upper one, ligamentum ventriculare, probably has a centring role, the ligamentum vocale playing the part on an oscillator for the thyroid cartilage as a resonator [4], [5]. A simplified description of laryngeal activity in the course of the four basic phases of the vocal fold cycle can be derived from Fig.7. The first phase is preceded by the mentioned setting up of phonatory position. The symbolic section through the thyroid cartilage passes from the respiratory to the centred position. Now the thyroid cartilage can vibrate around this new centred position. During the first phase, the expired air stream

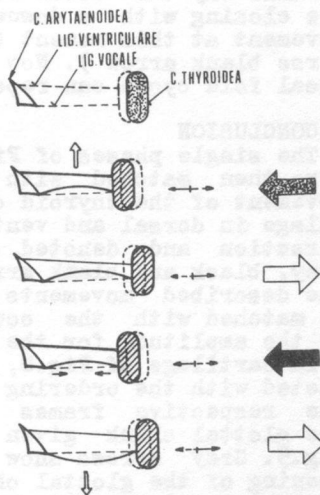


Fig.7 Four phases of the vocal fold cycle

stretches the vocal ligament pulling the thyroid cartilage inwards. Blank arrow indicates upward movement of the vocal folds. During the second phase, due to its own elasticity, the thyroid cartilage returns back to its equilibrium at once to overshoot outwards. During the third phase, the musculus vocalis probably contracts to attract the thyroid cartilage. During the last, fourth phase, the thyroid cartilage, again due to its own elasticity, will first return back at once to overshoot, taking away with it the vocal ligament and giving it an impulse to a downward movement.

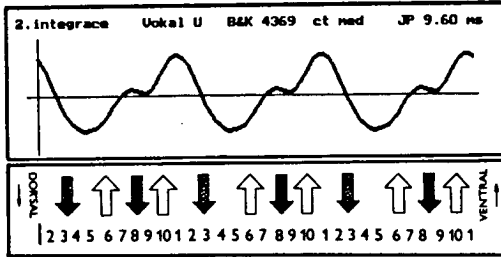


Fig.8
Thyroid cartilage
movement matched
with glottal chink
image

In this way the vocal folds are closing with a downward movement at this moment (inverse blank arrow). Now the vocal fold cycle can repeat.

4. CONCLUSION

The single phases of Fig.7 were then matched with the movement of the thyroid cartilage in dorsal and ventral direction and denoted by grey, black and blank arrows. The described movements can be matched with the course of the amplitude of Fig.8, completed with the ordering of the respective frames for the glottal chink given in Fig.5. Grey arrows show the opening of the glottal chink by the air stream, blank arrows the elastic backward movement of the thyroid cartilage, and black arrows the presupposed presence of a neuroreflex event, whose role it probably is to close the glottal chink before its subsequent opening. Compared with the situation in Fig.4, we thus obtain new information. What we now have is not only information on the progress of the opening and closing, but also on the way the glottal chink is being closed. So far, this process is accounted for by reference to Bernoulli's effect. The presence of neuromuscular junction is supported by the results obtained in subjects suffering from some organic lesions of the nervous system, in whom this event was

either inhibited or missing altogether.

The following conclusions can be made:

4.1 Laryngeal Vowel Differentiation

The complex event recorded straight on the thyroid cartilage is of vowel differentiated nature [2].

4.2 Laryngeal Diagnostic

Investigation of laryngeal vibrations offers diagnostic utilization in some organic lesions of the nervous system.

4.3 Study of Voice

The described complex event, that can be observed during speech and two octaves of a modal voice, can be used in the study of voice production.

5. REFERENCES

- [1] van den BERG, Jw. (1958), "Myoelastic - aerodynamic theory of voice production", Journ. Speech Hear. Res. 1: 227 - 244.
- [2] HIROSE, H. (1988), "High-Speed Digital Imaging of Vocal Fold Vibration", Acta Otolaryngol (Stockh). Suppl. 458.
- [3] PEŠÁK, J. (1990), "Differentiated Laryngeal Phonation", Folia Phoniatr. 42: 296-301.
- [4] PEŠÁK, J. (1990), "Complex Mechanism of Laryngeal Phonation: A. Description of Activity", Folia Phoniatr. 42: 201 - 207.
- [5] PEŠÁK, J. (1990), "Complex Mechanism of Laryngeal Phonation: B. Analogue Pattern of the Larynx", Folia Phoniatr. 42: 208 - 212.

AN INVESTIGATION OF LOCUS EQUATIONS AS A SOURCE OF RELATIONAL INVARIANCE FOR THE STOP PLACE DIMENSION

H. Sussman

University of Texas at Austin, Texas, U.S.A.

Locus equations, straight line regression fits to data points formed by plotting onsets of F2 transitions along the Y-axis & corresponding midvowel target frequencies along the X-axis, were generated across 20 speakers using speech tokens /b(V)t/, /d(V)t/, & /g(V)t/ with 10 vowel contexts. Slopes & y-intercepts were significantly different across stop place & correctly predicted stop categorization. Locus equations provide a higher-order category-level metric capable of capturing relational invariance for place of articulation.

1. INTRODUCTION

The coarticulatory nature of speech has led to the theoretical impasse known as the "invariance problem"- i.e., perceptual constancy despite physical variation in the signal. Phonetic segments are realized in an overlapped, dynamic, & context-sensitive fashion, while conceptualizations in the abstract depict them as discrete, static, & context-independent. The elusive quest for invariance, the search for stable acoustic cues that isomorphically encode the phonetic segment has been ongoing since the early 1950's. The 'limus test' for invariance has been place of articulation for stop consonants. The pur-

pose of this research was to offer a refocused conceptualization of a traditional candidate for acoustic invariance - the F2 transition as its onset & trajectory vary with the following coarticulated vowel. A formal metric that succinctly captures the relative changes occurring in F2 transitions will be investigated as a potential, higher-order cue invariantly signalling place of articulation in voiced stops. This metric was initially formulated by Lindblom [3] & termed "locus equations." Despite a surface similarity to the "virtual locus" concept [1], Lindblom's metric was not intended to formalize a fixed & context-independent acoustic correlate of stop place, but rather to illustrate the context-dependence existing between onset frequency of F2 and its location in the vowel nucleus of the syllable. The regression plots showed extreme linearity and tight clustering of data points. Moreover, the slope & y-intercept differed as a function of stop place for Swedish /b/, /d/, & /g/ followed by 8 vowels. One purpose of the present study was to determine if American English stop + vowel syllables would also show the extreme linearity & orderliness exhibited by Lind-

blom's data. Another rationale was that a higher-order linguistic abstraction could be used to investigate the invariance issue. All previous studies have examined acoustic cues derived at & characterizing the single phonetic segment. The locus equation metric is derived over & characterizes an entire stop place category. A nontrivial aspect of the invariance dilemma might very well relate to the proper level of abstractness of the linguistic elements for which the invariant acoustic properties are sought

2. PROCEDURE

2.1 Subjects

Twenty subjects, 10 male & 10 female were used, ranging in age from 18 to 46. Varied dialects of American English were spoken.

2.2 Stimulus materials

Subjects were asked to produce CVC syllables in a carrier phrase format "Say CVC again." Words were typed on a list in five randomized orderings. Initial stops were /b/, /d/, and /g/ followed by 10 medial vowels contexts /i, I, e, . ae, a, o, . , u/. The final consonant was always /t/. Thus, there were 10 /b(v)t/ tokens, 10 /d(v)t/ tokens, and 10 /g(v)t/ tokens, each repeated five times yielding a total of 150 utterances per subject.

2.3 Instrumentation

Each speaker's productions were recorded in a soundproof booth using a high quality microphone & cassette tape recorder. The recorded signal was sampled at 10kHz & digitized using an Apple MacIntosh II computer with MacAdios II hardware. The MacSpeech Lab II package was used for all display, editing, measurement & playback rou-

tines. F2 measures were obtained from three sources: (1) direct on-screen wideband spectrograms; (2) LPC spectra; (3) wide & narrowband FFTs.

2.4 Data Sample Points

The two formant measurement points were F2 onset defined as the frequency value of F2 at the first glottal pulse following the release burst and F2 vowel defined as the frequency of F2 at the midvowel nucleus. F2 vowel measurement points were not fixed in time: if F2 was "steady-state" a midpoint on the formant was taken; if F2 was diagonally rising or falling a midpoint position was similarly used; if F2 was 'parabolic' a minima/maxima point was taken for F2 vowel.

3. RESULTS

Sixty locus equation scatterplots were generated. Extremely tight clustering of points about the regression line were found throughout all speakers, regardless of gender. Collapsing across repetitions & subjects, group mean locus equation plots are presented in Figure 1 for initial labial, alveolar, and velar stop place. It can be seen that male & female coordinates lie along the same linear function with female values lying further out each axis. Labial /b/ had the steepest slope (.91) followed by /g/ (slope = .79, and then alveolar /d/ (slope = .54). An ANOVA on both slope & y-intercept parameters revealed significant main effects for place of articulation ($p < .001$).

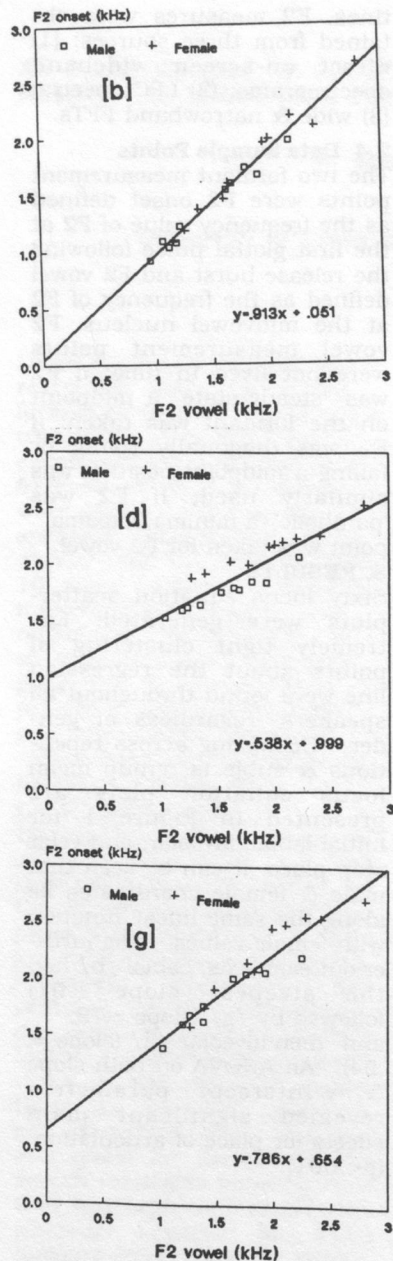


Fig. 1: Group mean locus equation scatterplots for /b/, /d/, & /g/.

To summarize up to this point plotting F2 onsets obtained at the CV boundary in relation to midvowel 'target' frequencies (F2 vowel) yields linear relationships that systematically vary with stop place. Thus, despite the extreme context-dependency of the coarticulated CV gesture a relational form of invariance is captured that is independent of the following vowel. At the single CV level no absolute signal invariance is present; only when the higher-order stop + vowel phoneme category is represented acoustically does a relational-type of invariance first emerge.

3.1 Discriminant Analyses

To test the categorical classification success of token-level versus category-level predictor variables a set of linear discriminant function analyses were run. At the single token level F2 onset & F2 vowel frequencies were used as predictors for place of articulation (chance = 33.3%). Percent correct classification rates were 83.1, 79.4, and 67.9 for labials, alveolars, & velars respectively. When the 60 derived slopes & y-intercepts (3 stop place locus equation functions per subject X 20 speakers) were used as predictors 100% correct classification was obtained.

3.2 CV "Prototypes"

Figure 2 shows canonical group mean regression lines for each stop place category. Velar /g/ is shown broken down into a more accurate subgrouping of allophonic variants of /g/ preceding front vowels (palatal -/g/p) & /g/ preceding back vowels (velar -/g/v). These mean regression

Prototypical "CV" Locus Equations

(n = 10 male + 10 female speakers)

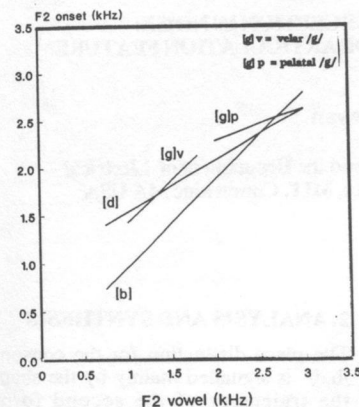


Fig. 2 Locus equation "prototype" functions for /b/, /d/, /g/-velar, & /g/-palatal place of articulation.

male & 10 female speakers are currently conceptualized as representing "prototypes" for CVs, & as such may contain the theoretical framework for understanding & studying the auditory representation of dynamically coarticulated CVs.

MALE GRAND MEAN (n=10)

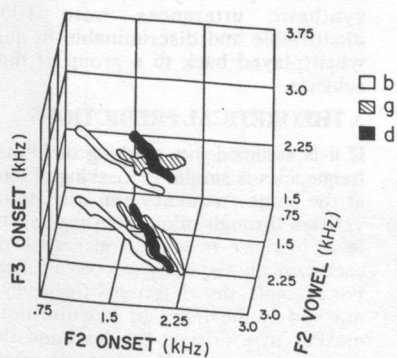


Fig. 3 3-D scatterplot of F2 onset, F2vowel, & F3onset frequencies for /b/, /d/, & /g/ followed by 10 medial vowels.

3.3 3-D Acoustic Space

Adding F3 onset (Hz) to the locus equation parameters provided a 3-D perspective of acoustic/phonetic space. Fig.3 illustrates a typical scatterplot averaged across the 10 male speakers. Distinct & minimally overlapping 'cloud' & respective 'shadow' representations were consistently found for both individual & group data as a function of the 3 stop place categories.

4. DISCUSSION

The data of this study demonstrate acoustic orderliness for stop place categories emerging from both locus equation & 3-D scatterplots of F2 & F3 data. A context-independent phonemic class descriptor, & hence a logical alternative to gestural-related invariance notions [2] has been demonstrated. The acoustic signal, despite coarticulatory complexity, contains a systematic set of correlational attributes of formant information capable of coding place of articulation without recourse to a gestural-level recoding of the signal. Stop place categories are sufficiently contrastive in their physical instantiation in the speech sound wave to permit direct auditory decoding.

5. REFERENCES

- [1] DELATTRE, P., LIBERMAN, A., & COOPER, F. (1955), "Acoustic loci and transitional cues for consonants," *JASA*, 27, 769-773.
- [2] LIBERMAN, A., & MATTINGLY, I. (1985), "The motor theory of speech perception revised," *Cognition*, 21, 1-36.
- [3] LINDBLOM, B. (1963), "On vowel reduction," *STL-No.29*. Stockholm, Sweden.

MODELLING SPEECH PERCEPTION IN NOISE: A CASE STUDY OF THE PLACE OF ARTICULATION FEATURE

Abeer Alwan

Research Laboratory of Electronics and the Department of Electrical
Engineering and Computer Science, MIT, Cambridge MA USA

ABSTRACT

In this study, perceptual confusions of the place of articulation feature for syllable-initial /b,d/ stop consonants in noise are examined. Experimental data are compared with a model, based on auditory masking theory, that estimates the level and spectrum of noise needed to mask each formant peak. Results show good agreement between the experiments and the theoretical model, and indicates that F2 transition is essential in signalling the place distinction for these consonants.

1. INTRODUCTION

The goal of the study is to develop procedures for predicting the perceptual confusions of speech sounds in noise. The prediction is based on the following premise: if the acoustic attributes that signal a particular phonetic contrast are known, then, based on auditory masking theory, it should be possible to calculate the level and spectrum of noise that will mask these acoustic attributes, and hence will lead to confusions in listener responses to that phonetic contrast. The methodology here is threefold: 1) quantifying acoustic correlates of phonetic features in naturally-spoken utterances and using the results to synthesize these utterances, 2) using masking theory to predict the level and spectrum of the noise which will mask these acoustic correlates, and 3) performing a series of perceptual experiments, using synthetic stimuli, to evaluate the theoretical predictions.

The feature chosen for the present phase of the study is place of articulation for the stop consonants /b,d/ in CV syllables with the vowels /a/ and /e/.

2. ANALYSIS AND SYNTHESIS

The place distinction for the consonants /b,d/ is signalled mainly by the shape of the trajectory of the second formant frequency (F2) and by the spectral shape of the burst. The F2 transition is thought to carry most of the place information for syllable-initial stop consonants [2]. In the /Ca/ case, the F2 trajectory falls into the vowel for the alveolar /d/ and rises for the labial /b/. With the vowel /e/, the F2 trajectory rises for /b/ and is almost flat for /d/. Figures 1 and 2 show schematized spectrograms of synthetic burstless utterances of /ba, da/ and of /be, de/, respectively, illustrating the differences in the F2 trajectories. These synthetic utterances, which are used later in perceptual experiments, are generated using KLSYN88 [4]. The choice of parameters is based on analyses of natural utterances spoken by a male speaker. The synthetic utterances were 100% identifiable and discriminable in quiet when played back to a group of three subjects.

3. THEORETICAL PREDICTIONS

If it is assumed that masking of formant frequencies is similar to masking of tones at the same frequency (an assumption verified through pilot experiments), the level of noise needed to just mask out each formant frequency can be calculated. For example, the i th formant frequency is masked if the level of a white-noise masker in a critical-band around that formant frequency (Nc_i) is 4 dB greater than the amplitude (in dB) of the formant frequency (A_i) [5]. Nc_i is the rms level of the noise (in dB) estimated from the DFT spectrum and corrected by 10 log

(ratio of the analysis-filter bandwidth to the critical bandwidth the i th formant frequency). That is, the condition is that $Nc_i \geq A_i + 4$. Calculations can then be made to determine the time interval over which each formant frequency is masked. Figure 3 illustrates these computations for a white-noise masker at a particular level for which F2 transition in the synthetic /da/ stimulus is partially masked. In this case, F1 is never masked ($A_1 + 4 > Nc_1$), F3 is always masked ($A_3 + 4 < Nc_3$) and only the first 10 ms of F2 is masked. Note that the spectral peak of F2 changes by about 10 dB during the transition period. This is in accordance with observations of amplitude changes in natural speech. The computations are done every pitch period.

4. EXPERIMENTS

The goal of these experiments is to examine the perceptual importance of the F2 trajectory in signalling the place-of-articulation feature distinction for the plosives /b,d/ in syllable-initial position with the vowels /a/ and /e/. Nonsense syllables were used to make sure that lexical effects, such as word frequency, do not bias subjects' responses.

4.1 Stimuli and Experimental Design

Synthetic utterances of /ba, da, be, de/ were attenuated, mixed with white noise, randomized, repeated 10 times and presented to subjects in identification tests. There were 13 stimuli with different signal-noise ratios (SNR) for each utterance. The SNR was varied by changing the signal level in 1 dB steps while keeping the noise level constant. The presentation level, as determined by the peak in the vowel, was 66 to 79 dB SPL.

4.2 Subjects

Four subjects participated in the /Ca/ experiments and three subjects participated in the /Ce/ experiments. Two of the subjects were students at MIT. None had any known speech or hearing problems. Training periods, lasting between 1/2 h to 1 h depending on the subject, preceded each listening session.

4.3 Results

4.3.1 /Ca/ case

The results of these experiments show that the /ba/ stimuli were perceived correctly at all noise levels used in the experiment. Figure 4 shows the results of the experiments for each subject individually for the /da/ stimuli. The responses are plotted as a function of the SNR in a critical band of F2 in the steady-state portion of the vowel. These identification functions show an abrupt shift from /da/ to /ba/. The average threshold for the subjects occurs at the stimulus where 23 ms of the F2 transition is masked. The total duration of the F2 transition is 40 ms.

4.3.2 /Ce/ case

The results of these experiments show that the /de/ stimuli were identified correctly at all noise levels used in the experiment even when F2 is completely masked. Figure 5 shows the results of the experiments for each subject individually for the /be/ stimuli. The responses are plotted as a function of the SNR in a critical band of F2 in the steady-state portion of the vowel. These identification functions show a shift in perception from /be/ to /de/. However, the identification functions show individual differences in listeners' responses. It is interesting to note that these differences are similar to those found in the listeners' masked thresholds of pure tones in independent tests.

5. Discussion

The results of this study show that the shape of the F2 trajectory is essential in identifying the place of articulation for the consonants /b/ and /d/ preceding the vowels /a/ and /e/. The labial feature is signalled by a flat trajectory when preceding /a/ and a rising trajectory preceding /e/. If noise masks most of the F2 transition such that only the steady-state part of the transition is free of masking, then /de/ is perceived. The feature alveolar, on the other hand, is signalled by a flat trajectory preceding /e/ and a falling trajectory preceding /a/. If noise masks out most of the F2 transition for /da/ such that the movement of F2 is minimal, then the stimulus is perceived as /ba/. This result is in agreement with results of other researchers [1][3] who

observed that the first 20 ms or so of the F2 transition carries important place information for /d/. Their observations were based on perceptual experiments conducted in quiet.

Other experiments examining the perceptual role of stop bursts are underway. Preliminary results indicate that in the /Ca/ case and in the presence of white noise, the burst is masked at very low SNR and, hence, does not play a significant perceptual role. We plan to pursue this approach further in investigating other phonetic contrasts in noise such as manner of articulation and voicing and to test the model under 'shaped' noise conditions.

6. References

[1] Blumstein, S., and Stevens, K.N. (1980). "Perceptual invariance and onset spectra for stop consonants in different environments," *J. Acoust. Soc. Am.*, 67, 648-662.

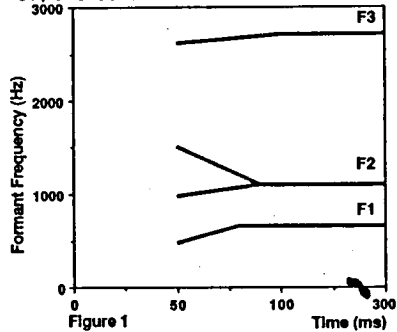


Figure 1

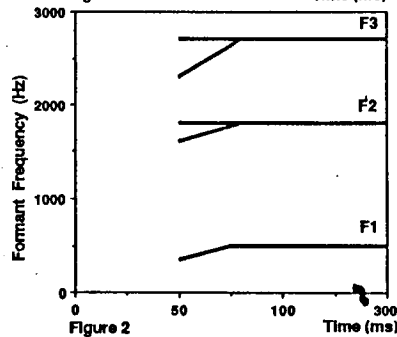


Figure 2

Schematized trajectories for the first three formant frequencies for synthetic /ba/ (solid line) and /da/ (dashed line) utterances in Fig.1 and of /de/ (solid line) and /be/ (dashed line) utterances in Fig.2.

[2] Delattre, P.C., Liberman, A.M., and Cooper, F.S. (1955). "Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. Am.*, 27, 769-773.

[3] Kewley-Port, D. (1983). "Time-varying features as correlates of place of articulation in stop consonants" *J. Acoust. Soc. Am.*, 73, 322-335.

[4] Klatt, D. H. and Klatt L.C. (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, 87, 820-857.

[5] Moore, B. (1982). *An Introduction to the Psychology of Hearing*. Academic Press, London.

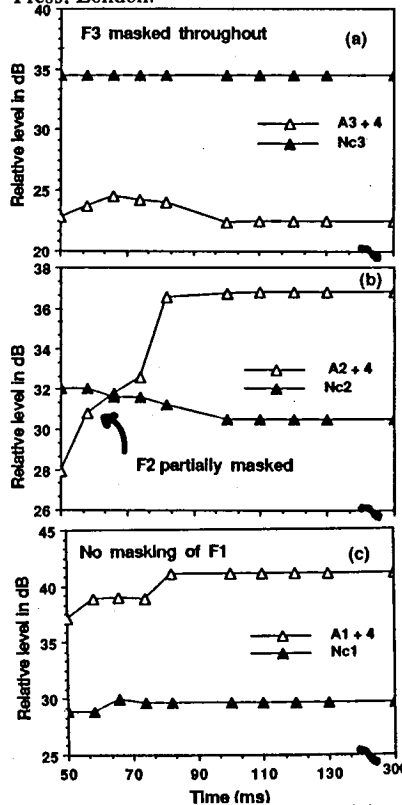


Figure 3. Plots of the relative levels of the amplitudes of the first three formant frequencies plus 4 dB (re 0.0002 bar) along with the noise levels (Nc_i) in the corresponding critical bands. Masking occurs when Nc_i is at least as high as $A_i + 4$. The data are for a /da/ stimulus.

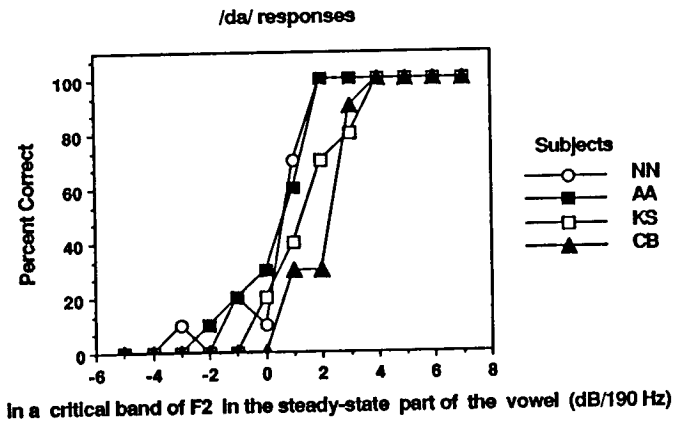


Figure 4. Plots of percent correct versus the signal-to-noise ratio in a critical band of F2 in the steady-state part of the vowel. The critical band in this case is 190 Hz. The plots are for the /da/ stimuli.

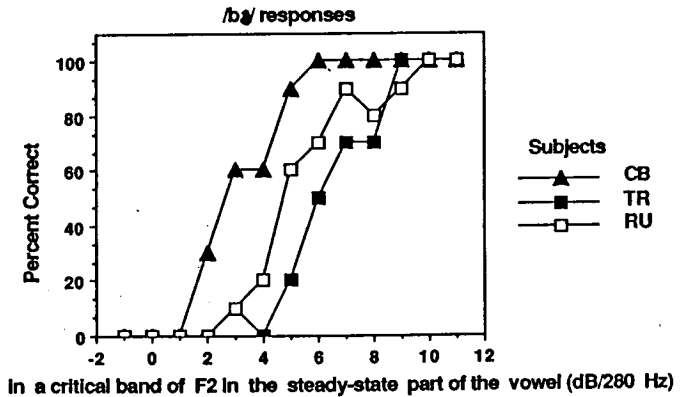


Figure 5. Plots of percent correct versus the signal-to-noise ratio in a critical band of F2 in the steady-state part of the vowel. The critical band in this case is 280 Hz. The plots are for the /be/ stimuli.

MECHANISMS OF VOWEL PERCEPTION:
EVIDENCE FROM STEP VOWELS

F. GOODING

Dept. of Linguistics, Univ. of Wales, Bangor

ABSTRACT

"Step tones" were constructed in which a series of equal intensity harmonics substituted for the upper formants of synthetic English front vowels. The number of HF harmonics, and thus the lower "edge" of the HF step was varied, along with the relative level of the HF step. Since it was already known from previous studies that step tones were perceived as vowels, the present experiments were designed to explore systematically the effects of edge frequency and level in order to determine the role of these attributes in determining vowel quality.

1. INTRODUCTION

Earlier work by the author has provided evidence on vowel perception that cannot be accounted for by traditional formant frequency based theories. First, experiments with two-formant vowels [3] demonstrated that continuous changes in phonetic quality can be achieved by altering relative formant amplitudes. Secondly, a wide range of highly recognizable vowel qualities can be elicited by stimuli with no formant peaks [4,5]. The stimuli used in the latter experiments had auditorily flat spectra: a single LF series of loudness-equalized harmonics ('step') eliciting back vowels,

and both low and high frequency steps together eliciting central and front vowels. For a given LF step, phonetic quality was dependent upon both the LF edge of the HF step and the relative LF/HF step amplitudes. In general, the findings with single step tones [4] were similar to those with single formant vowels, in that a large range of satisfactory back vowels were produced, while the range and naturalness of the front vowels were considerably less (indeed, it has been found that [i] was the only front vowel that could be elicited by single formant stimuli [2]).

The results with two-step stimuli [5] showed that highly identifiable, natural sounding vowel qualities, including the troublesome front vowels, can be elicited by such stimuli. In addition, the relative amplitude of the low and high frequency steps in these stimuli contributed to vowel quality in a fashion analogous to that of the formant amplitudes in two formant vowels of [3].

2. PRESENT STUDY

2.1 Aim and Rationale

The primary aim of the present study was to provide more evidence that would help to choose between alternative explanations of the earlier data. To this end, it was decided to examine more closely the role of the frequency of the lower edge of the HF step in tones with energy

in both the F1 and upper formant regions, and any possible interaction of edge frequency with level. Theories to accommodate the earlier data must account for the following: (1) the lack of formant peaks (2) the "edge effect" -- vowel quality dependence on the edge frequencies of the LF and HF steps (3) amplitude dependence -- quality dependence on the relative LF/HF level.

It would seem that two types of theory could account for the edge effect (restricting our attention here to the LF edge of the HF step): (a) the edge hypothesis (EH) -- the lower edge frequency is extracted and used directly. (b) a "center of gravity" (CoG) hypothesis (CGH) --- something akin to the local CoG of the whole HF step (or upper formant region in natural vowels) is taken. Changing LF edge would have the effect of changing the CoG as well. CGH would predict that quality would depend on both the HF as well as the LF edge of the HF step (though to reduce the number of stimuli only the LF edge was manipulated in these experiments. This can still distinguish between the competing hypotheses). CGH would predict that identifications would take place by matching the CoG of the HF step (roughly the mid point on a pitch scale for a flat series of harmonics) against the CoG of the upper formant region in the S's internal reference vowel (perhaps roughly equivalent to F2 prime). In short, EH would predict best identifications with edge frequencies at or slightly below F2, while CGH would predict identifications with edges well above F2. As a guide, midpoints for the HF steps ranged from 2554 Hz to 3057 Hz when measured on the ERB-rate scale [7]. To account for the amplitude effect, it would seem that a mechanism involving some form global spectral balance, or

perhaps CoG, is implicated. In the case of the latter, it would involve operation over a distance of greater than the 3.5 Bark limit originally suggested by Chistovich and her colleagues [1].

2.1 Stimuli

The stimuli were produced by digital harmonic synthesis with a sampling rate of 10 KHz, and LP filtered at 4 KHz (filter cutoff rate 180 dB/Octave). All had a fundamental frequency of 125 Hz, and duration was 300 ms, with 20 ms linear onsets and offsets. The stimuli were modified from those in [5] in that in the F1 region a formant appropriate to one of four RP front vowel was substituted for the LF step of loudness-equalized harmonics. This was done in order to reduce the strong sensation of nasality that accompanied some of the earlier stimuli. It was assumed that this was caused by the apparent broad bandwidth of the F1 region, known to be associated with the secondary feature of nasality. Four F1 values, 272, 380, 525, and 713, were used, appropriate to the RP phonemes /i/, /I/, / / and /ae/, respectively [6].

The LF edge of the HF step varied from 1750 Hz to 2500 Hz, and the HF limit was fixed at 3750 Hz. This value was chosen because HF energy above this frequency added a fricative-like or whistling sound to some stimuli, which, while distracting, was clearly heard as separate from the vowel. This might be seen as casting doubt on the CGH, since altering the HF edge is thus shown not to effect vowel quality. However, it can be claimed that since the energy above ca. 3750 Hz is not integrated into the vowel percept, this does not constitute a real test of that theory.

HF amplitudes varied in 10 dB steps from 0 to -40 dB for the earlier stimulus sets and from -10 dB to -30 dB for the final

set.

2.3 Procedure

A matching experiment (not reported on here) and an identification experiment were carried out. This was computer controlled, the stimuli being presented on-line and responses entered via the keyboard. A total of 22 Ss listened to four different randomized blocks of the stimulus set (ultimately 39 stimuli, though some subjects heard supersets of 52 and 86 stimuli). Randomizations were different for each S. Ss could listen to the stimulus as often as they wished by pressing a key. For each stimulus, Ss were asked to enter a score representing the English (RP) vowel it most resembled, identified by 13 key words shown at the top of the screen and identified by number. They also entered a confidence score (0-9) for their choice.

3. RESULTS

3.1 Edge Effect

The results, interpreted through the use of stimulus and response profiles, clearly support the EH and contradict the CGH. The stimuli most identified as the front vowels were in all cases those with edge frequency close to the F2 of the natural vowel, as predicted by the EH. /i/ unsurprisingly proved to be the most identifiable vowel. /ae/ was the least identifiable, with stimuli designed to elicit it achieving only a 24% score for the first 10 Ss. This is probably due to the lack of the RP value for most Ss in the own speech, in favor of North and Midlands [a]. These stimuli were hence dropped from the final set for the last 12 Ss.

3.2 Amplitude Effect

For a given F1 value and a given response category, identification scores were not always a monotonic function of HF level (though scores for /i/ with F1 = 272 most closely approximate this), but rather showed evi-

dence of a trading relation between edge frequency and level. This needs to be examined more closely, but if verified, it could be taken as evidence for the operation of a global spectral balance or CoG mechanism.

With the exception of stimuli eliciting the most /i/ responses (F1 at 272 and edge frequency at 2250 and 2500 Hz), virtually all responses for stimuli with levels of -40 dB were back or central vowels. The /i/ stimuli, by contrast, achieved scores over 60% at -40 dB, compared to 90% at -10 dB. Below -20 dB there was a clear shift to central and back vowel responses. The main point is that the responses for the 4 different levels were significantly different, indicating phonetic change with level. This is borne out by the judgements of two professional phoneticians to the whole range of stimuli.

Table 1 shows a brief summary of the pooled results with rounded scores. Note that except for /e/ (which is phonetically [eɪ] in RP) the F1 of the stimuli corresponded to the F1 of the response vowel. No stimuli were designed to elicit /e/, but the stimulus eliciting the most /e/ responses had an F1 of 380 Hz (appropriate for /I/).

TABLE 1
Summary of Pooled results

Vowel	F2	Edge Freq	Score	Level
/i/	2361	2500	90%	-10 dB
		2250	83%	-10 dB
/I/	2085	2000	42%	-10 dB
/e/	(2000)	2125	39%	-10 dB
/ /	1943	2000	47%	-10 dB

4. DISCUSSION

The implication of the finding of a spectral edge feature in synthetic step vowels for the

perception of natural vowels is that F2 in front vowels must serve as marker of the edge of the upper formant region. This feature appears to be used in conjunction with global amplitude information. The evidence reported here is not consistent with a local CoG mechanism operating over upper formant region.

5. REFERENCES

- [1]CHISTOVICH, L.A. and LUBLIN-SKAYA, V.V. (1979), "The 'centre of gravity' effect and the critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli", *Hearing Res.* 1, 185-195.
- [2]DELATTRE, P., et al. (1952), "An experimental study of the acoustic determinants of vowel color", *Word*, 8, 195-210.
- [3]GOODING, F., (1986), "On the role of formant amplitudes in vowel perception", *IEE Conf. Pub.* 258.2, 287-291.
- [4]_____ (1986b), "Formantless vowels and theories of vowel perception" *JASA* 80:Sup.1, S126
- [5]_____ (1988), "Formantless vowels: the next step", in WA Ainsworth and JN Holmes (eds.) *Speech '88. Proc. 7th FASE Symposium*, Edinburgh, 747-754.
- [6]HENTON, C., (1983), "Changes in the vowels of Received Pronunciation", *J. Phon.*, 11, 353-371.
- [7]MOORE, B., and GLASBERG, B., (1986), "The role of frequency selectivity in the perception of loudness, pitch and time", in B.Moore, ed., *Frequency Selectivity in Hearing*, Academic, 251-308.

TWO PROCESSING MECHANISMS IN RHYTHM PERCEPTION

Morio Kohno* Asako Kashiwagi** Toshihiro Kashiwagi***

* Kobe City University of Foreign Studies, Kobe, Japan.

** Kyoritsu Rehabilitation Hospital, Hyogo, Japan.

*** Kyowakai Hospital, Osaka, Japan

ABSTRACT

In Experiment I, it was found that the left hand of a patient with infarction involving the forebrain commissural fibers (S-1) could not follow the slow rhythms of 500 and 1000ms IBIs, but it could follow the rapid stimuli of 250ms IBI. The right hand of S-1, however, could synchronize his tapping with all rhythms as well as normal adults (S-2). Negative autocorrelations were detected among the adjacent IBIs in slow response beats by S-2 and by the right hand of S-1, but these correlations were never found in the rapid response movements (250ms) of any subjects. This means that normal adults use ongoing, analytic processing for slow rhythm but holistic processing for rapid rhythm. Evidence was found that the left hand of S-1 can use only the holistic approach, not only for the rapid rhythm but also for the slow rhythm, and that this is the very reason why it cannot follow the slow tempos. Experiment II was performed to show that the above two processings are qualitatively different from each other, and Experiment III & IV show that the holistic approach is more tenable to memorize a nonsense succession of approximately seven syllables than is the analytic processing.

1. PURPOSE OF THE PRESENT STUDY

There are no established ideas about the universal timing fundamentals among phoneticians.

This paper is to propose the universal timing measure on the basis of neuropsychological experiments on

sound sequence processing using as subjects a patient with infarction in corpus callosum, children with age variety from 1 year and 4 months to 9 years of age as well as normal adults.

2. MECHANISM OF RHYTHM PROCESSING

2.1 Experiment 1

Subjects: Three kinds of subjects were prepared.

Subject 1 (S-1, henceforth) was a patient (male) with infarction involving the forebrain commissural fibers. He was a 56-year-old right-handed public official having education of 16 years. He developed a sudden paresis on the right limb in December 1983, and was admitted to a local hospital. Diagnosis was made as having cerebral thrombosis, but no pathological lesion was recognized by the CT scan. On July 14, 1985, he was found not awoken in bed by a family and brought to another local hospital. His CT and MRI scans at the time of this study shows lesions situated in the posterior half of the genu and the whole truncus of the corpus callosum as well as in a posterior superior portion of the left medial frontal lobe, and in the left medial temporo-occipital lobes. Small low density spots are also found in the bilateral basal ganglia. S-1 was normal in his words and consciousness, and although he was disoriented in time and space due to an amnesia, general intelligence was not grossly impaired. Pathological reflexes and ankle clonus were negative. Muscle tone was

normal. He was not ataxic. He had neither gross paresis nor definite sensory loss, but he demonstrated a small step arteriosclerotic gait[2].

Other subjects were a healthy 55-year-old woman (right-handed) (S-2, henceforth) and seven young children with ages ranging from 1 year and 4 months old(1:4) to 9 years old(9:0). These children were all righthanded and had no known abnormalities.

Method: Three kinds of rhythm whose inter-beat intervals (IBI, henceforth) were 250, 500, 1000ms were prepared by the use of a metronome, SEIKO Rhythm Trainer SQM-348, and then they were demonstrated to the above-mentioned three kinds of subjects. They were all requested to tap the table simultaneously in time with the above rhythms using the knuckles of third fingers, first of their right hands and then of their left hands. When the children's tapping was measured, their mothers accompanied them together with the experimenter and explained how to carry out their tasks giving some examples and let them practice beforehand. The 1:4 child's tapping, however, was recorded in her house by her mother letting the child use a toy which made sound. The mother understood the aim of this experiment very well.

All the tapping records were analyzed by the YOKOKAWA Electro-magnetic Oscillograph 2901 connected with the Amplifier 3125.

Results: Table 1 shows the tapping behavior of the normal adult (S-2) and Table 2, of the patient (S-1). These two tables tell us that the right hand of S-1 moves very normally as well as the both hands of S-2, which show the standard movements of normal adults. In other words, the means of IBIs (\bar{x}) and the values of S.D. of the both subjects do not differ so much. S-1's left hand, however, moves very differently from not only the both hands of the normal adult (S-2) but from his own right hand. S-1's left hand can manage to follow the rapid stimuli of 250ms IBIs, but it can never follow any slow rhythms with 500 and 1000ms IBIs. The values of the S.D. spread

very widely. The right and the left hands of S-1 seem to move separately. S-1 often said to the authors, "It (=the left hand) moves out of my own will."

Table 1 Tapping by a normal adult (female, 55 years old, right hander)

used hand	target tempos (A)	inter-beat intervals				
		n.	\bar{x}	SD	SD/A	r
right	1000ms	63	1022.7	52.1	13.0	-0.29
	500ms	55	512.7	22.9	11.5	-0.21
	250ms	99	257.5	10.8	10.6	+0.45
left	1000ms	57	1017.7	54.9	13.7	-0.10
	500ms	71	515.3	22.2	11.1	-0.12
	250ms	94	257.0	11.0	11.1	+0.04

Table 2 Tapping by a patient with infarction in the corpus callosum (male, 57 years old, right hander)

used hand	target tempos (A)	inter-beat intervals				
		n.	\bar{x}	SD	SD/A	r
right	1000ms	27	1020.1	46.5	11.6	-0.52
	500ms	46	506.3	31.6	15.8	-0.25
	250ms	56	261.9	27.1	27.1	+0.19
left	1000ms	62	673.3	285.7	71.4	+0.36
	500ms	99	476.4	198.9	99.5	+0.07
	250ms	51	266.6	36.0	36.0	+0.13

Table 3, which gives us the whole view of the tapping behavior of the children, shows that movements of the children older than 4 years old (S-3, henceforth) are remarkably different from the movements of children younger than four (S-4, henceforth), and we should notice that the former's behavior is very much like the behaviors of the normal adult (S-2) and of the patient's (S-1's) right hand and the latter's one is exactly the same as the movement of the patient's left hand. In other words, younger children (S-4), just like S-1's left hand, can synchronize their tapping with the fast rhythm of 250ms IBI, but they cannot follow the slow tempos of 500 and 1000ms IBIs. In this connection, we should notice that, according to Krashen, the lateralization of the brain must be established at about the age of five[6].

Table 3 Tapping by the right hand of children
(All are right handers)

age	target tempo (A)	Inter-beat intervals				
		n.	x	SD	SD/A	r
9:0 (female)	1000ms	25	1003.8	78.0	19.0	-0.15
	500ms	80	508.7	25.7	12.9	-0.82
	250ms	76	253.4	20.8	20.8	+0.13
5:10 (male)	1000ms	13	1058.5	122.2	30.8	-0.38
	500ms	45	499.2	52.0	28.0	-0.33
	250ms	43	301.3	31.5	31.5	+0.01
4:2 (male)	1000ms	58	637.2	340.8	85.2	-0.02
	500ms	30	479.0	31.2	15.6	-0.29
	250ms	18	319.8	24.0	24.0	+0.57
3:9 (female)	1000ms	84	581.4	333.2	83.3	+0.34
	500ms	78	448.9	112.8	56.3	+0.31
	250ms	87	284.5	25.1	25.1	+0.41
3:4 (male)	1000ms	50	801.4	354.7	88.7	-0.03
	500ms	78	431.7	109.0	64.5	+0.32
	250ms	53	276.3	31.2	31.2	+0.38
2:8 (male)	1000ms	22	471.7	301.9	75.5	+0.32
	500ms	50	478.8	81.0	40.5	+0.80
	250ms	38	348.8	30.2	30.2	+0.04
1:4 (female)	1000ms	11	708.5	218.4	54.5	
	500ms	18	408.1	158.7	79.4	

(The left hand movements are the same as the right hand movements.)

Autocorrelation among adjacent IBIs (r) was then calculated for all the response data of Tables 1, 2 and 3 (see the right column of each table) and negative correlations (r=0.2~0.6) were found among the adjacent IBIs in the slow response beats (fitted to 500 and 1000 ms) by S-2 (Table 1), S-3 (Table 3) and by the right hand of S-1 (Table 2), but we could not find any negative autocorrelations in any of the responses of the children younger than 4 years old (S-4) and the patient's (S-1's) left hands who produce only rapid responses even to slow stimuli like 500 or 1000 ms. None of the subjects' response beats to the rapid 250ms stimulus, on the other hand, produced any negative correlations.

All these mean that, as suggested by Hibi who did his experiment by letting subjects say "pa-pa-pa--" instead of tapping[1], the normal adults including the children older than 4 years old usually use ongoing, analytic (prediction-testing) processing to the slow rhythm, but holistic processing to the rapid rhythm. The right hand of S-1 can

properly process both the slow and rapid rhythms correctly using either the analytic or the holistic approach, while his left hand cannot, nor can both hands of S-4.

The above-mentioned facts suggest that following the rapid tempos and following the slow ones are neuropsychologically different from each other --- the former is holistic and the latter is analytic.

3. MECHANISM OF ECHOIC MEMORY Experiment 2

In order to verify the above-mentioned hypothesis of two mechanisms in perception from another viewpoint, the author held the following experiment, which also made clear the mechanism of echoic memory.

Method: Nonsense words of seven CV (consonant and vowel) syllables such as 'ga ta ku da do pe ki' were prepared. When these words were recorded, the speaker (a Japanese female in her twenties) read these words fitting each syllable to each beat produced by the metronome with beat intervals of 250, 500, and 1000ms, and then these intervals were adjusted so as to be rigidly spaced in these intervals by the use of ILS (DEC Micro Computer PDP 11/73). By the use of this method 6 nonsense words were made per each of the rhythmic patterns of 250, 500, 1000ms IBIs.

The materials were then given, in a language laboratory, to Japanese students (25 in number) majoring English at a women's college and before they tried to write down the nonsense words on the answer sheet, they were requested to do simple multiplication of two digit numbers (e.g. 16×75) immediately after they had listened to each of the nonsense words, and then soon to write them as well as possible. Marking was done giving two points to each correct answer, that is, correct recall not only in reproduction of the syllable but in its position in the words. As the initially presented nonsense word was used for exercises, the full mark was 70 (7×5×2) for each category of syllable intervals.

Results: Table 4 shows the result of the experiment.

Table 4

IBIs among syllables	the state of reproduction		
	n	\bar{x}	S.D.
1000ms	25	38.4	13.2
500ms	25	38.7	12.0
250ms	25	45.4	11.4

n=number of subjects

250>500 t=2.01 p<0.05

250>1000 t=1.99 p<0.05

These tables show that the syllables connected closely with short IBIs, which are processed holistically, bring about significantly longer retention than the slow-tempo-syllable-sequences which are processed analytically. The analytic processing of the nonsense words and the work of multiplication may be both cognitive, and therefore the retention of the nonsense words was interfered by the calculation. The holistic processing of the closely connected syllable sequence, on the other hand, will be neuropsychologically different from the work of multiplication, and it was never disturbed by the calculation.

The memory span of holistically processed syllable sequences, however, is not so large. Miller suggests that it might be 7±2[7]. Kohno and Tsushima confirm this number by the fact that the babbling and the one word utterances (in total, 2448) by a child of age one and half never continue over 7 syllables in succession[5].

4. Conclusion

Holistic processing copes with fast rhythmic condition of less than about 300ms IBIs, analytic one with slow tempos of more than about 400ms, and the tempos between 300 and 400ms IBIs may be the threshold area which belongs neither to holistic nor analytic ones (cf. [1]). Whether or not the tempos in this area belong to holistic or analytic will depend on individuals (Kohno & Ishikawa, forthcoming paper). Holistic processing, which is neuropsychologically different from analytic one, will never be disturbed by the lat-

ter, which plays an important role of 'analysis by synthesis' to get the whole meaning of discourse.

REFERENCES:

- [1] Hibi, S. "Rhythm Perception in Repetitive Sound Sequence." *Journal of the Acoustical Society of Japan* 4-5, 83-95. 1983.
- [2] Kashiwagi, A., T. Kashiwagi, T. Nishikawa and J. Okuda. "Hemispheric Asymmetry of Processing Temporal Aspects of Repetitive Movement in Two Patients with Infarction Involving the Corpus Callosum." *Neuropsychologia* 27-6, 799-809
- [3] Kohno, M. "Effect of Pausing in Listening Comprehension." T. Konishi ed. *Studies in Grammar and Language* 392-405, Kenkyusha, Tokyo. 1981.
- [4] Kohno, M. Two processing Mechanisms in Rhythm Perception and Listening Comprehension -- A report of Research in 1989 by Grant-in-Aid for Scientific Research on Priority Areas. The Ministry of Education, Science and Culture, Japan. 1990.
- [5] Kohno, M. and T. Tsushima. "Rhythmic Phenomena in a Child's Babbling and One-word Sentences." *The bulletin No.191*, 6-13, The Phonetic Society of Japan.
- [6] Krashen, S. D. "Lateralization. Language Learning and the Critical Period." *Language Learning* 23-1, 63-74. 1973.
- [7] Miller, G. "The Magical Number Seven, Plus or Minus Two." *Psychological Review* 63-2, 81-97. 1956.

**LEXICAL STRESS IN A 'STRESSLESS' LANGUAGE:
JUDGMENTS BY TELUGU-AND ENGLISH-SPEAKING LINGUISTS**

L. Lisker and Bh. Krishnamurti

Haskins Laboratories, New Haven, USA;
University of Hyderabad, Hyderabad, India

ABSTRACT

The properties of speech referred to by the terms *stress accent prominence sonority* have never been defined physically with enough precision for descriptions of language in those terms to be validated on the basis of "hard" data. The basis for saying that English *bellow* is trochaic and *below* is iambic is simply that a consensus of "competent observers" finds this so. What constitutes competence in stress perception is not clear, but a minimum requirement is some degree of consistency in judging native forms. This paper reports, for linguistically trained listeners with and without command of a language without contrastive stress, Telugu, just how consistently the location of primary stress or prominence is reported.

1. INTRODUCTION

Phonetic-phonological discussions of a language sometimes refer to a property X, and sometimes to a perceived property X. Thus, in the phonology of the English lexicon, the word *believe* may be said to have an initial voiced element or one perceived to be voiced. The distinction (supposing one to be intended) implies that there is a viable non-perceptual, i.e. physical definition of voicing. Some properties ascribed to speech are not of this kind; thus, to say that the second syllable of *believe* is perceived as stressed is not different from simply asserting it to be stressed, since there seems to be no reliable independent acoustic basis for defining the feature of stress [4]. Whereas we can point to a mismatch between voicing and perceived voicing, we cannot similarly claim perceived *believe* is "really" something else.

Most of the literature on stress takes for granted the stress status of a linguistic sample, and addresses itself to the search for its physiological and/or acoustic correlates. No doubt nearly all linguists who speak English natively would agree on the stressing of *believe*, and the truth of the assertion about its stressing is entirely a matter of the degree to which it is accepted by the community of "competent observers." Linguistic opinion on that "community" is not clear. Is training in phonetics and phonology enough, or, as per Jones [2] and Chomsky and Halle [1], must one also have native control of the language? Aside from the fact that few of us explicitly disqualify ourselves as judges of languages not our own, the second requirement would render questionable, if not entirely nugatory, a large part of the literature on stress, including that dealing in general characterizations of languages. Presumably a minimum requirement is a certain degree of consistency in judging native forms, as well as agreement with other observers presumed to be equally or better "qualified." So while it is probably true that stress rules for English are largely the work of English-speaking linguists, stress in other languages has certainly not been pursued exclusively by those with native command. Hyman [3], for example, surveying accounts of some 444 languages, raised no question of observers' competence by this criterion.

In Hyman's survey languages are classified into those with contrastive stress and those with either a fixed or otherwise "predictable" pattern over the word. Of several Indian languages included in the survey a number, including Telugu, are described as

having "dominant initial stress." The basis for this assessment of Telugu seemed doubtful to us (although it has been reported that Telugu speakers tend to stress the initial syllables of English words [5]), and the present research was undertaken to establish a proper empirical basis for a description of stress in the language. Another purpose of the study was to compare the consistency of stress judgments rendered by observers of roughly similar levels of linguistic sophistication, but with very different levels of competence in the language under examination, in order to test the proposition that a difference in language command plays a considerable role in determining consistency of stress perception.

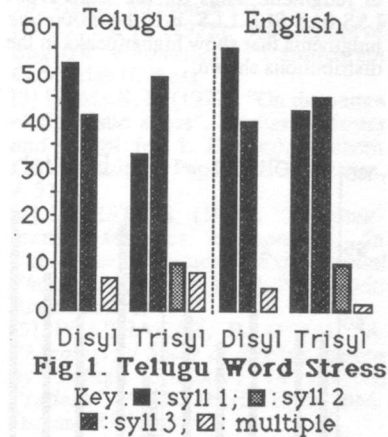
2. PROCEDURE

A list of randomly ordered di- and trisyllabic Telugu words was recorded by a single native speaker of the language. Each word was pronounced with a pitch fall on the final syllable. There were thirty disyllables and twenty trisyllables in the list, one token per word. Since in Telugu both vowels and consonants have distinctive length, and since vowel length or syllable "weight" are known to be factors in other accentual systems, words chosen included various combinations of short and long vowels (and light and heavy syllables). Two groups of listeners were tested: ten Telugu-speaking graduate students in linguistics with training in phonetics, all with a knowledge of English; 2) fifteen English-speaking linguists, none with any previous exposure to Telugu. Listeners were provided with the words in standard broad transcription, and were asked to respond to the recorded words, as these were presented over a loudspeaker in a reasonably quiet room, by marking the location of primary stress, with the option of selecting more than one as "equally stressed." Two responses per word were elicited from each test subject.

3. RESULTS

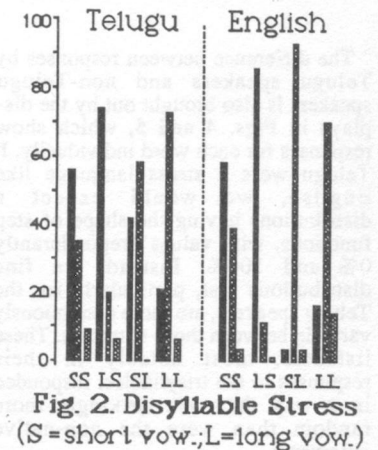
The responses by our Telugu and English speakers are tabulated in several ways in Figs. 1-5. From Fig. 1 it

seems clear that there was no very large difference between the two listener groups, and that neither showed any strong preference for selecting initial syllables as the bearers of primary stress. If anything, the penultimate syllable was somewhat more often chosen as the locus of primary stress.



Disyll Trisyll Disyll Trisyll
Fig. 1. Telugu Word Stress

Key: ■: syll 1; ■: syll 2
■: syll 3; ▨: multiple



SS LS SL LL SS LS SL LL
Fig. 2. Disyllable Stress
(S=short vow.; L=long vow.)

When stimuli are grouped on the basis of their vowel composition, as per Figs. 2 and 3, it is again clear that there is no great difference between the groups. Disyllables with no long vowel are somewhat more often reported as trochaic, but in trisyllables with only short vowels it is the second syllable that is most often judged to be stressed. Moreover, the English speakers show a somewhat greater degree of consistency of judgment. Thus for the word types LSS SLS SLL LLS, it is the *nonnative* judgments that show higher peaks in the distributions shown.

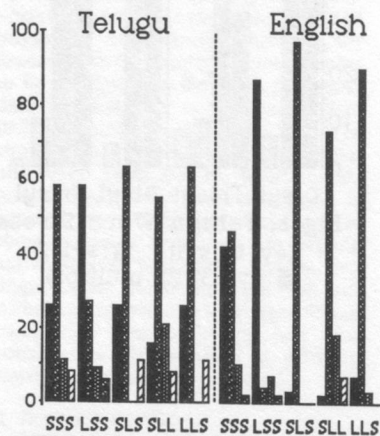


Fig. 3. Trisyllable Stress

The difference between responses by Telugu speakers and non-Telugu speakers is also brought out by the displays in Figs. 4 and 5, which show responses for each word individually. If Telugu were a stress language like English, we would expect a distributions having the shape of step functions, with values preponderantly 0% and 100%. Instead, we find distributions that, particularly for the Telugu speakers, are more continuously variable between those extremes. These listeners, most notably in their responses to the trisyllables, responded in a way that was strikingly more random than were the non-native judgments.

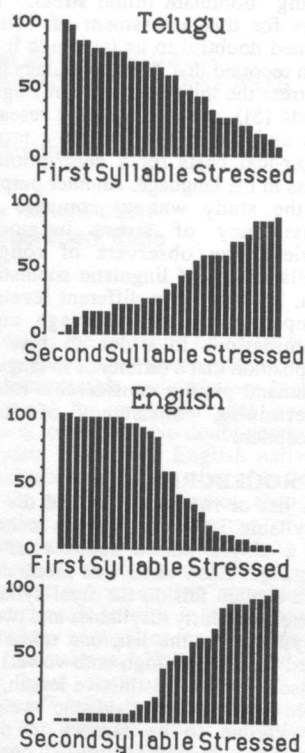


Fig. 4. Judgments of 30 Disyllables

The original motivation for the exercises just described was to characterize stress or prominence patterns for Telugu, presupposing such to be universal features of speech, in order then to proceed to a search for acoustic properties marking these patterns. To some extent both kinds of listeners showed a preference for locating stress on the last long vowel of a word, and that would seem to answer our original question. But the notable failure of the Telugu listeners to respond as categorically as the English speakers calls for explanation, and there is little enough that can be mustered to account for this aspect of our data. We hesitate to conclude that the greater consistency of judgments by the non-Telugu speakers means that they are a "better" guide to stress in the language, for that would

be to assume that we already know what we are trying to find out: the truth of the matter.

4. ACKNOWLEDGMENTS

This work was supported by the American Institute of Indian Studies and by NIH Grant HD-01994 to Haskins Laboratories.

5. REFERENCES

- [1] CHOMSKY, N. and HALLE, M. (1968), "The Sound Pattern of English", The Hague: Mouton & Co.
- [2] JONES, D. (1972), "An Outline of English Phonetics", Cambridge: Cambridge U. Press.
- [3] HYMAN, L. (1977), "On the nature of linguistic stress", *Studies in Stress and Accent* (ed. L. Hyman), Southern Cal. Occasional Papers in Linguistics, 4, 37-82.
- [4] LEHISTE, I. (1976), "Suprasegmental features of speech", In *Contemporary Issues in Experimental Phonetics* (ed. N.J. Lass), New York: Academic Press.
- [5] PRABHAKAR, B. (1974), "A Phonological Study of English Spoken by Telugu Speakers in Andhra Pradesh", Ph. D. Thesis. Hyderabad: Osmania University.

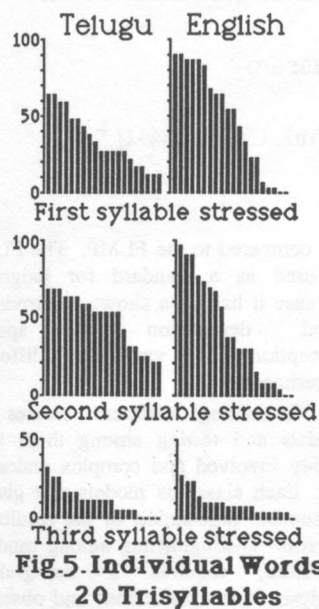


Fig. 5. Individual Words 20 Trisyllables

CONNECTIONIST MODELS OF SPEECH PERCEPTION

Dominic W. Massaro

University of California, Santa Cruz, CA 95064 U.S.A.

ABSTRACT

Interactive-activation and feed-forward connectionist models are evaluated, tested, and compared to a process model, the Fuzzy Logical Model of Perception (FLMP). Empirical results indicate that while several sources of information simultaneously influence speech perception, the representation of each source remains independent of other sources. This independence is strong evidence against interactive activation in speech perception. Although some feed-forward models with input and output layers bear some similarity to the FLMP, there is evidence against the additive integration that is assumed by feed-forward models.

1. INTRODUCTION

At the Eleventh Congress of Phonetic Sciences, I described the Fuzzy Logical Model of Perception (FLMP), an information processing model of speech perception [2]. The FLMP has been shown to provide a good description of speech perception in a variety of different experiments. The model accounts for the evaluation and integration of multiple sources of information in speech perception. These sources of information include acoustic, visible, and electrotactile sources of bottom-up stimulus input, as well as top-down sources of phonological, syntactic, and semantic context. In the present paper, several classes of connectionist models

are compared to the FLMP. The FLMP is used as a standard for judgment because it has been shown to provide a good description of speech perception in a variety of different experiments.

Evaluating different classes of models and testing among them is a highly involved and complex endeavor [6]. Each class has models that give a reasonable description of the results of interest. Distinguishing among models, therefore, requires a fine-grained analysis of the predictions and observations to determine quantitative differences in the accuracy of the models. Preference for one class of models is also influenced by factors other than just goodness of fit between experiment and theory. Some models are too powerful and thus not falsifiable. With enough hidden units, for example, connectionist models can predict too many different results [3]. Models should also help us understand the phenomena of interest. For example, parameters of a model might provide illuminating dependent measures of the information available in speech perception and the processing of that information. Finally, one should take into account the parsimony of a model. Certainly, a model should contain fewer parameters than the number of data points that it predicts. Models which can provide a good fit to the data with relatively few parameters should be preferred.

2. INTERACTIVE ACTIVATION

In interactive activation models, layers of units are connected in hierarchical fashion with two-way connections among units both within a layer and between layers. For example, the TRACE model of speech perception has feature, phoneme, and word layers. There are excitatory two-way connections between pairs of units from different layers and inhibitory two-way connections between pairs of units within the same layer. Thus, interactive activation is based on the assumption that the activation of a higher layer eventually modifies the activation and information representation at a lower layer [7].

How might interactive activation and the TRACE model be formulated to predict the results of bimodal speech perception? Given audible and visible speech, for example, separate sets of feature units would be associated with the two different information sources. Figure 1 gives a schematic representation of the auditory feature, visual feature, and phoneme layers and the connections between units within and between these layers. The two layers of feature units would both be connected to the phoneme layer. Following the logic of interactive activation, there would be two-way excitatory connections between the feature and phoneme layers (as in the TRACE model). Presentation of auditory speech would activate some units within the auditory feature layer. These activated units in turn would activate certain phoneme units, which would in turn activate units at both feature layers, and so on during the period of interactive activation. Activated units would also inhibit other units within the same layer.

If auditory and visual units interact, as assumed by interactive activation, then presentation of a syllable in one modality should influence processing of

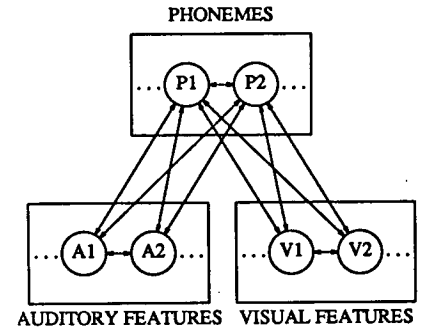


Figure 1. Illustration of the TRACE model applied to bimodal speech perception. Two input layers contain auditory and visual feature units, respectively. The third layer contains phoneme units. There are positive connections between two units from different layers and negative connection between two units within the same layer.

the syllable in the other modality. If interactive activation does not occur, on the other hand, the contribution of visible speech should be independent of the contribution of audible speech. Independence means that the representation of the visible speech should not be modified by the representation of the audible speech. The results from several different experiments in several different tasks indicate that interactive activation does not occur in bimodal speech perception [2, 8]. More generally, there is now a substantial body of evidence against interactive activation in speech perception [4, 5].

3. FEED-FORWARD MODELS

In contrast to interactive activation, feed-forward models assume that activation feeds only forward. Two-layer models have an input layer connected to an output layer, as illustrated in Figure 2.

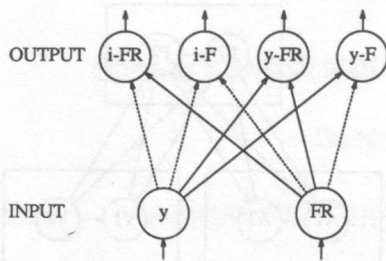


Figure 2. Illustration of a connectionist model (CMP) of speech perception of four words in Mandarin Chinese. The formant structure y and (F_0) contour FR input units are connected to the four response units corresponding to the four word alternatives. Solid arrows indicate connections with weight 1, and dashed arrows indicate connections with weight -1.

The feed-forward model illustrated in Figure 2 is tested against the results of an experimental study of the identification of Mandarin Chinese words [6]. There were four possible responses in the experiment. The experimental task was a graded factorial design with seven levels of each of two factors. The factors were the formant structure of the vowel in the monosyllabic words and the fundamental frequency (F_0) contour (tone) during the vowel. Mandarin Chinese is a tone language and both of these sources of information function to distinguish among different words. The formant structure was varied to make a continuum of vowel sounds between /i/ and /y/. (The phoneme /y/ is articulated in the same manner as /i/, except with the lips rounded.) The F_0 contour varied between falling-rising to falling during the vowel. Six native Chinese speakers participated for four days, giving a total number of 48 responses to each of the 49 test stimuli. The subjects identified each

of the 49 test stimuli as one of the four words.

Figure 3 gives the observed results and the predictions of the FLMP and the connectionist model (CMP). As can be seen in the figure, the CMP fails catastrophically primarily because it cannot predict a probability of a response greater than .5. The FLMP, on the other hand, captures the results reasonably well. The success of the FLMP is due to the multiplicative integration of the two sources of information. A perfect match of a stimulus with a given response alternative on just one source does not necessarily mean that this alternative should qualify as a reasonably good alternative. The linear integration in the CMP, however, guarantees that a perfect match of a response alternative with just a single source of information will be significantly activated even if the other source of information mismatches the response alternative completely.

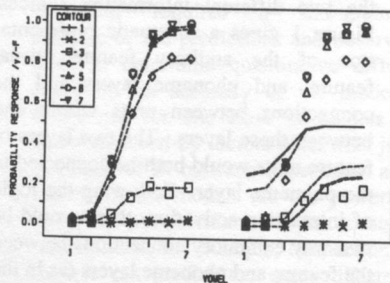


Figure 3. Observed (points) and predicted (lines) probability of /y/-falling responses for the Chinese word identification study [6]. The left panel gives the predictions for the FLMP and the right panel gives the predictions for the CMP.

Admittedly, we have falsified a very restricted implementation of the class of feed-forward connectionist models. However, we are only willing to test models that are falsifiable. Three-layer models, for example,

assume that the input units are connected to a layer of "hidden" units that are connected to an output layer of units. In a theoretical and analytical report, I have shown that models with hidden units are superpowerful—that is, they can predict many types of results and even results that do not occur [3]. Because these models can predict many results—not just those that are empirically observed, this superpower might be better described as flabbiness. Therefore, one cannot reasonably propose feed-forward models with hidden units as testable models of speech perception. These models are not reasonable because they are not falsifiable. In one case, for example, the model is essentially assuming more than it is predicting [1], and the good performance by the model in this situation should not be surprising.

In summary, there is evidence against interactive activation models, while feed-forward models with hidden units are not falsifiable. Feed-forward models with input and output units can be shown to be mathematically equivalent to the FLMP in situations with just two responses [6]. With a larger number of responses, the FLMP provides a more adequate description of the results than does this feed-forward model.

4. REFERENCES

- [1] Landauer, T. K.; Kamm, C. A.; & Singhal, S. (1987). Network to recognize speech sounds. *Proceedings of the Cognitive Science Society*, 531-536. Hillsdale, NJ: Lawrence Erlbaum Associates.
- [2] Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

- [3] Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language*, 27, 213-234.
- [4] Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, 21, 398-421.
- [5] Massaro, D. W., & Cohen, M. M. (in press). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology*.
- [6] Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, 97, 225-252.
- [7] McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.
- [8] Roberts, M., & Summerfield, Q. U. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics*, 30, 309-314.

V. ACKNOWLEDGMENT

The research reported in this paper and the writing of the paper were supported, in part, by grants from the Public Health Service (PHS R01 NS 20314), the National Science Foundation (BNS 8812728), a James McKeen Cattell Fellowship, and the graduate division of the University of California, Santa Cruz. The author would like to thank Michael M. Cohen for eclectic assistance.

PERCEPTION OF SPECTRALLY COMPRESSED SPEECH

R.R.HURTIG

Dept. of Speech Pathology & Audiology
University of Iowa, Iowa City, Iowa, USA

ABSTRACT

The effect of spectral compression and frequency transposition on the perception of vowels and simple sentences is examined. A computational algorithm which can accomplish spectral compression and frequency transposition is presented. Analysis of confusion data suggests that spectral shape rather than absolute formant frequency differences is critical to vowel identification. The limits of frequency transposition appear to be determined by critical bark differences.

1. INTRODUCTION

The perception of speech, and in particular vowels, is in many respects conditioned by both static and dynamic cues in the speech spectrum. The centrality of the first and second formants in vowel identification is generally accepted. At one level it has been suggested that the detection of spectral prominences (formants) provides the necessary and sufficient acoustic cues to the vocal tract configuration which is associated with a particular vowel production. This position must be tempered somewhat by the "speaker normalization effect" and the range of formant values for speakers of very different ages. Recent accounts of vowel perception [5,6] suggest that an extension of the "center of gravity" hypothesis [1] may provide the basis for the detection of vowels based on the notion that a given vowel can be represented in terms of the

extent to which formant frequency differences are greater than or less than 3 Bark. Thus it may be a categorical difference rather than an absolute difference in formant frequencies that specifies a particular vowel.

This leads to a prediction that only a subset of the signal processing schemes should preserve vowel identification and that other transformations should lead to a decrement in speech perception. Specifically, any processing of the speech spectrum which preserves the Bark difference values distinguishing vowels will result in adequate speech perception. By contrast any transformation which alters the distinctions marked by the Bark difference values of the natural stimuli should render the speech unintelligible and perhaps alter the signal sufficiently to make it lose its speech quality.

Such an interpretation holds that within limits it is the spectral shape (i.e., slope) which is the invariant cue to vowel identity. Thus the acoustic cues are not necessarily formant values associated with normal vocal tract configurations but are rather values which can be associated with a scaled vocal tract. To test this general hypothesis an algorithm for processing the speech spectrum was selected which would maintain the spectral shape but which would allow a scaling of the spectrum (frequency

compression) and a transposition of the spectrum (frequency shift).

2. PROCEDURE

2.1 TFT-Algorithm

A computational algorithm [2] was used to achieve spectral compression and frequency transposition of natural speech samples. The TFT (Time-to-frequency-to-time domain) algorithm performs its manipulations of the signal in the frequency domain. The algorithm operates on successive windows of n -samples. First an n -point FFT is computed. The resultant complex array is padded with the spectrum of a Hamming window. An IFFT is then computed on the padded array. The resultant real array is then trimmed to the length of the original input window. The size of the pad relative to the size of the input window determines the degree of compression to be achieved. The placement of the pad relative to the complex array determines the degree of frequency transposition. For a pad of a length equal to the input window a frequency compression to 50% of the original spectrum is achieved; a pad three times the length of the input window yields a compression to 25% of the original. Positioning the pad following the complex array yields compression with no frequency transposition of the resultant spectrum. By contrast a leading pad would achieve both compression and an upward shift in the frequency of the resultant spectrum. A partitioning of the pad into a leading and following component can yield varying degrees of frequency transposition based on the proportion of the pad in the leading position.

2.2 Stimuli

Multiple tokens of natural vowels [i, I, c, æ, a, ʌ, ɔ, o, u, u] in the hVd context as well as sentences from short narrative passages were digitized and subjected to the TFT algorithm. These stimuli were compressed to 50% or

25% of the original and were also frequency transposed in 5 steps from 150 Hz to 2500 Hz.

2.3 Presentation and testing of vowel tokens.

Listeners (N=7) were given the opportunity to explore a set of vowels at a given compression and frequency transposition using a listener paced exploratory learning task in which the listener is free to select the specific token to be played out. The listeners explored a given set of stimuli for a minimum of fifteen minutes after which the computer presented the tokens in a random order in a closed set recognition paradigm.

2.4 Presentation and testing of the sentence tokens.

The sentence tokens were presented in either a random order or in short connected discourses. Listeners (N=9) were required to write down as much of each sentence as they could. In addition some listeners were presented sentences with varying degrees of frequency transposition for open set recognition and judgments of speech quality and intelligibility.

3. RESULTS

3.1 Vowels.

The correct identification of vowels was 44% in the 50% compression condition and 81% in the 50% compression with a 150 Hz transposition condition. These scores represent performance which is significantly better than chance performance. All listeners showed a marked improvement in the 150 Hz transposition condition. An Information Transfer Analysis [3] reveals the relative information transferred for the first three formants, for each of the conditions. It should be noted that in both conditions there were comparable transfer rates for each of the format cues.

Table 1.
Percent Relative
Information Transferred

	F1	F2	F3
50%	26	33	27
50%(+150Hz)	73	73	74

The best subject's performance was 64% in the 50% compression condition and 86% in the condition with 150 Hz transposition. Table 2 indicates near perfect transfer of formant information in the transposed condition.

Table 2.
Best Subject Percent
Relative Information Transferred

	F1	F2	F3
50%	53	51	44
50%(+150Hz)	92	92	93

3.2 Sentences.

Tracking was assessed in terms of the percentage of words correctly identified. The tracking of speech compressed to 50% is fairly easy to do and requires effectively no listening experience. Tracking 50% compressed sentences with a 150 Hz frequency transposition was better than tracking untransposed sentences. Likewise tracking of 50% compressed speech was better than the tracking of 25% compressed speech. As expected, providing a context improves tracking.

Table 3.
Percent Correct Tracking
No Context Context

50%	87	99
50%(+150Hz)	97	100
25%(+150Hz)	32	67

The subjective quality of the resultant frequency compressed speech varies with the degree of frequency transposition. The most natural sounding compressed tokens are those

with a minimal amount of frequency transposition (from about 150-500 Hz). As the frequency transposition exceeds 500 Hz the intelligibility and naturalness of the tokens deteriorates.

4.0 DISCUSSION

The ability to identify vowels and to track running speech which has been frequency compressed argues that speech perception is the consequence of a process which is sensitive to spectral shape rather than specific spectral prominences which can be associated with formants produced by natural vocal tracts. The limitation on this ability appears to be based on the critical bandwidth of the auditory system's filters [7]. A linear transposition of the spectrum in the frequency domain will result in a greater probability of formant peaks falling within a single critical band. At the extreme the signal loses any resemblance to speech and has a cricket-like quality. This effect is easily predicted from the Bark difference scores which result from linear frequency transpositions. Using the formant values from the TI data base [5] one can calculate the bark difference values for both natural vowels as well as those with 50% and 25% compression and various amounts of linear frequency transposition. (As Figure 1 illustrates the natural distinctions are basically maintained with transposition of less than 650 Hz.)

These data are consistent with the results obtained with frequency transposition hearing aids which fail to yield results better than those obtained with simple filter circuits. The tinny percept is perhaps similar to unpleasant percept reported by cochlear implant users. This may be due to typical electrode insertion which stimulates fibers associated with higher frequencies and correspondingly internal filters with wider critical bandwidths.

	TI	A: F2-F1						
		50%	50% 150	50% 350	50% 650	50% 1250	50% 2500	25%
heed								
hid								
head								
had								
herd								
huhd								
hahd								
hod								
hood								
huwd								

	TI	B: F3-F2						
		50%	50% 150	50% 350	50% 650	50% 1250	50% 2500	25%
heed								
hid								
head								
had								
herd								
huhd								
hahd								
hod								
hood								
huwd								

Figure 1.

Bark Difference values for vowels in the TI Data Base and for compressed and frequency transposed versions. (+ indicates a difference of <3 Bark)

5. REFERENCES

- [1] Chistovich, L.A. & Lublinskaya, V.V. (1979), "The 'center of gravity' effect in vowel spectra and critical distance between formants: Psychoacoustical study of the perception of vowel-like stimuli.", *Hear.Res.*, Vol.1, 185-195.
- [2] Hurtig, R.R. (1989), "TFT, An algorithm for the spectral compression of natural speech signals.", *J.Acoust.Soc.Am.*, Suppl.1, Vol. 85, S44.
- [3] Nabelek, A.K. & Letowski, T.R. (1985) "Vowel confusions of hearing-impaired listeners under reverberant and nonreverberant conditions.", *JSHD*, Vol. 50, 126-131.
- [4] Peterson, G.E. & Barney, H.L. (1952), "Control methods used in the study of vowels.", *J.Acoust.Soc.Am.*, Vol. 24, 175-184.
- [5] Syrdal, A.K. (1985), "Aspects of a model of the auditory representation of American English Vowels.", *Speech Commun.*, Vol. 4, 121-125.
- [6] Syrdal, A.K. & Gopal, H.S. (1986), "A perceptual model of vowel recognition based on the auditory representation of American English vowels.", *J.Acoust.Soc.Am.*, Vol.79(4), 1086-1100.
- [7] Zwicker, E. & Terhardt, E. (1980), "Analytic expressions for critical-band rate and critical bandwidth as a function of frequency.", *J.Acoust.Soc.Am.*, Vol.68(5), 1523-1525.

A MODEL OF OPTIMAL TONAL FEATURE PERCEPTION

D. House

Department of Linguistics and Phonetics
Lund University, Lund, Sweden.

ABSTRACT

This paper postulates an optimal perceptual model for the tonal features *High*, *Low*, *Falling* and *Rising* based on proposed perceptual constraints of the auditory system in processing tonal movement in speech. These constraints involve two critical issues concerning the perception of tonal movement: namely the relationship between perception and the timing of tonal movement in terms of segment boundaries, and the perceptual primacy of level features over movement contour features.

1. INTRODUCTION

A classic descriptive problem in tone feature analysis concerns the division of tones into level tones and contour tones [1]. The principal question relating to this problem is whether tonal features can best be described in terms of both levels (e.g. *High*, *Low*, *Mid*) and contours (e.g. *Falling*, *Rising*) or as only levels and combinations of levels (e.g. *High + Low* instead of *Falling*).

This paper attempts to answer the question in terms of an optimal perceptual model of tonal movement categorization. The model is based on a series of speech perception experiments where tonal contours were varied in relation to segmental boundaries (see House [6] for a full account of the experiments). Results of the experiments indicate that actual pitch movement is optimally perceived as movement when it occurs during spectrally stable portions of vowels. Pitch movement occurring through areas of spectral discontinuities with rapidly shifting intensity and spectral information (roughly corresponding to segment boundaries) tends to be recoded in terms

of pitch levels. In the model, therefore, tonal movement through areas of rapid spectral change is optimally categorized as level features while constraint conditions must be fulfilled for tonal movement to be optimally perceived as contour features.

2. THE MODEL

2.1. Description and Constraints

The model proposed here is an optimal perceptual model for the tonal features *High*, *Low*, *Falling* and *Rising*. According to the model, tonal movement through areas of rapid spectral change will be optimally categorized as level features, a falling movement as the feature *Low* and a rising movement as the feature *High*. These basic level features *High* and *Low* are then perceptually associated with the vowel following the rapid spectral changes. This perceptual recoding of movement into level features is also seen as a perceptual primacy of level features over movement contour features where level features can also be manifested without tonal movement.

For the movement contour features *Falling* and *Rising* to be optimally perceived, three constraint conditions must be fulfilled. First of all, the falling or rising movement must take place through a zone of relative spectral stability during the vowel. Second, the beginning of the movement must be synchronized in relationship to vowel onset so that the beginning of the fall or rise coincides with an area of decreasing new spectral information following the rapid spectral changes associated with the transitions from consonant to vowel. This enables pitch extraction of a relative pitch frequency (high before falling and low before rising) to which the perception of pitch movement direction can be

calibrated. Finally, the model proposes a duration constraint which requires a vowel duration greater than 100 milliseconds to optimize movement feature perception.

100 milliseconds is an ad hoc value chosen to illustrate the durational component of the model. Relative vowel durations vary with speech tempo and speaking style, but the basic tenet is that longer vowel duration is associated with movement features while shorter duration is associated with level features. Based upon the effect of duration on the experimental results reported in [6], 100 milliseconds is a reasonable quantification of a duration constraint.

There may also be differences between rises and falls in their influence on production and perception of vowel durations [10]. These differences could have implications for the duration constraint in the model. However, for the purpose of simplicity in the model the duration constraint is the same for both rises and falls.

When the three constraint conditions are met, tonal movement is optimally categorized in terms of the movement contour features *Falling* and *Rising*. By implication, when the conditions are not met, tonal movement is then optimally categorized in terms of the level features *High* and *Low*.

Finally, the model assumes a tonal movement of 3 to 8 semitones per 100 milliseconds. Although the size of tonal movement needed for the optimal perception of movement contour features in the context of this model has yet to be tested experimentally, this range corresponds to that used in the

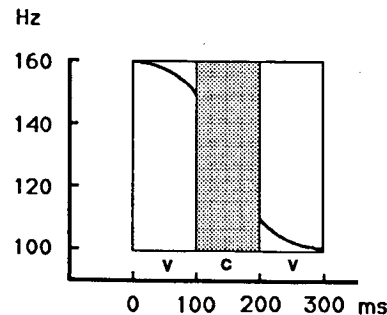


Figure 1. Illustration of the model applied to a falling tonal movement through a VCV context. The model predicts perception of the tonal features *High + Low* in the two vowels.

experiments in [6] and in many other experiments and models (e.g. [5], [13] & [15]).

2.2. Illustrations

Illustrations of the model as applied to a prototypical falling fundamental frequency contour in different segmental contexts are shown in Figures 1-4.

In Figure 1, with a VCV context and most of the tonal movement occurring during the consonant, the model would predict recoding in terms of the level features *High* and *Low*. In this example, none of the three movement feature conditions is met.

Figure 2 presents a CVC context in which only one of the three movement feature conditions is met. Although the falling movement does occur during spectral stability, the beginning of the fall occurs during spectral change and is not synchronized with the area of decreasing new spectral information following vowel onset. Finally duration does not exceed 100 milliseconds. Here, the model would predict categorization in terms of the level feature *Low*.

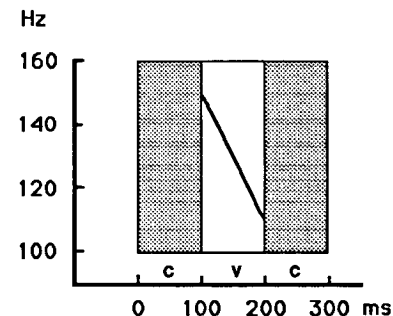


Figure 2. Illustration of the model applied to a falling tonal movement through a CVC context. The model predicts perception of the tonal feature *Low* in the vowel.

In Figure 3, a VC context is presented. Here, all three movement conditions are met. The model would therefore predict optimal coding in terms of the movement feature *Falling*.

Finally, in Figure 4, a CV context is presented. In this example, two of the three movement conditions are met. Tonal movement occurs during spectral stability in the vowel and vowel duration exceeds 100 milliseconds, but the beginning of the tonal movement is not synchronized with the end of maximum new spectral

information after vowel onset. Therefore the model would predict optimal recoding in terms of the level feature *Low*.

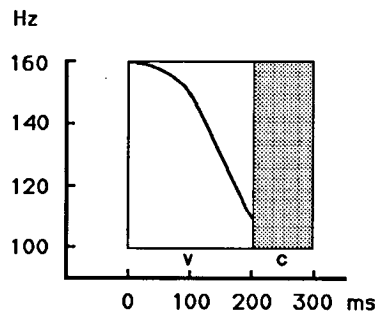


Figure 3. Illustration of the model applied to a falling tonal movement through a VC context. The model predicts perception of the tonal feature *Falling* in the vowel.

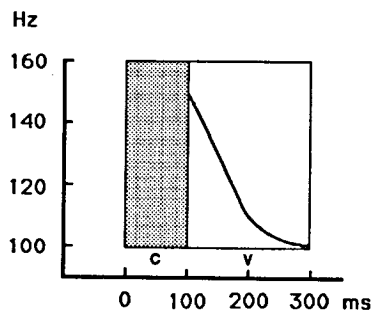


Figure 4. Illustration of the model applied to a falling tonal movement through a CV context. The model predicts perception of the tonal feature *Low* in the vowel.

In the illustrations above, a prototypically *falling* fundamental frequency contour was used. The model would deal with a *rising* contour in the same way with the same constraint conditions applying for optimal perception of the feature *Rising*.

Finally, it must be pointed out that the model described here is preliminary and does not claim to account for all possible tonal contrasts as it contains only four tonal features. More production and perception data from various languages could help expand the model as feature contrast requirements in different languages might alter the basic constraint conditions, especially where more features are needed.

3. IMPLICATIONS OF THE MODEL

3.1. Pitch Perception

The main implication of the model in terms of pitch perception theories is that it takes into consideration the constraints imposed on the perceptual system by the dynamic nature of speech. In the model, pitch movement sensitivity is strongly related to the speed of spectral change. In areas of rapid spectral change and spectral discontinuities, sensitivity is lower. In areas of spectral stability, sensitivity is higher.

The model assumes spectral analysis to be an important component in pitch perception. The use of resolved lower harmonics in pitch perception is also crucial for the model. It is the disruption and changes in the lower harmonics brought about by spectral discontinuities which give rise to the optimal recoding of tonal movement into level and movement contour features.

Thus the model supports central processing theories of pitch perception where a first order analysis of harmonic frequencies is crucial (e.g. [7]). Important to these theories is an interaction between spectral analysis and pitch extraction. This interaction is also crucial to the model.

3.2. Tone and Accent Features

Returning to the descriptive problem of contour tones versus level tones, the model clearly differentiates between contour and level features from a perception point of view. The model can be used to provide a framework for the assignment of features on a universal level. The model also implies perceptual primacy of levels over contours. These perceptual constraints can be used to explain certain aspects of universals of tone such as those reported by Maddieson [11] where it is claimed that languages do not have contour tones unless they have at least one level tone. Thus the features *High* and *Low* would be more perceptually salient and more frequent than the features *Rising* and *Falling*.

Perceptual constraints and the synchronization between tonal movement and segmental boundaries appear to be important in word accent and tone languages such as Swedish, Chinese and Thai which make use of lexical movement features. In these languages, tonal production can be seen to make optimal use of the perceptual contrast between levels and movements by means of the

critical timing of tonal movement (cf. [2], [3], [4], & [8]).

In other languages where the use of movement features is less clear, this type of synchronization may not be as important. However, data from German [9] and English [12] seem to indicate that level and movement features related to critical timing and alignment of tonal movement may play an important perceptual role for these languages as well.

3.3. Speech Perception Theories

An additional factor of importance for the model is the load on the perceptual system at vowel onset. Spectral cues at vowel onset have been shown to be crucial for segment perception in speech [14]. An implication of the model could thus be separate perceptual mechanisms for segmental cues and tonal cues. Segmental perception mechanisms would then favor rapid spectral changes and discontinuities while tonal perception mechanisms would favor spectral stability. Following this line of argument, F_0 differences at vowel onsets can function primarily as discriminatory cues for consonant features while F_0 differences during spectral stability function as cues for tonal features.

4. CONCLUSIONS

In the model presented in this paper, level features are given perceptual priority over movement contour features. For the optimal perception of contour features, constraint conditions are proposed and illustrated. Although the details of these constraint conditions may vary between languages, the principles of perceptual constraints should be applicable to many different languages on a universal level. These principles provide a framework for the assignment of tonal and intonational features from a perceptual point of view.

Finally, in view of the importance of tonal features for the overall speech perception process, the addition of a tonal component can be seen as a necessary enrichment of general models and theories of speech perception. The tonal perception model presented here is an example of such a tonal component.

5. REFERENCES

[1] ANDERSON, S.R. (1978), "Tone Features", in V.A. Fromkin (ed.) *Tone: A linguistic survey*, 133-175, New York:

Academic Press.

[2] BRUCE, G. (1977), "Swedish Word Accents in Sentence Perspective", Lund: Gleerups.

[3] GANDOUR, J.T. (1983), "Tone perception in Far Eastern Languages", *Journal of Phonetics* 11, 149-175.

[4] GÄRDING, E., P. KRATOCHVIL, J.O. SVANTESSON, & J. ZHANG (1986), "Tone 4 and Tone 3. Discrimination in Modern Standard Chinese", *Lang. and Speech* 29, 281-293.

[5] HART, J. T. & A. COHEN (1973), "Intonation by rule: a perceptual quest", *Journal of Phonetics* 1, 309-327.

[6] HOUSE, D. (1990), "Tonal Perception in Speech", Lund: Lund University Press.

[7] HOUTSMA, A.J.M. & J.G. BEERENDS (1987), "An Optimum Pitch Processing Model for Simultaneous Complex Tones", in U. Viks (ed.) *Proc. 11th ICPHS*, 4:325-330, Tallin.

[8] HOWIE, J.M. (1974) "On the Domain of Tone in Mandarin", *Phonetica* 30, 129-148.

[9] KOHLER, K.J. (1987), "Categorical Pitch Perception", in U. Viks (ed.) *Proc. 11th ICPHS*, 5:331-333, Tallin.

[10] LEHISTE, I. (1976), "Influence of fundamental frequency pattern on the perception of duration", *Journal of Phonetics* 4, 113-117.

[11] MADDIESON, I. (1978), "Universals of Tone", in J. Greenberg (ed.) *Universals of Human Language, Volume 2, Phonology*, 335-365. Stanford, CA: Stanford University Press.

[12] PIERREHUMBERT, J.B. & S.A. STEELE (1989), "Categories of Tonal Alignment in English", *Phonetica* 46, 181-196.

[13] ROSSI, M. (1978), "La perception des glissandos descendants dans les contours prosodiques", *Phonetica* 35, 11-40.

[14] STEVENS, K.N. & S.E. BLUMSTEIN (1981), "The Search for Invariant Acoustic Correlates of Phonetic Features", in P. Eimas & J. Miller (eds.) *Perspectives in the Study of Speech*, Hillsdale, NJ: Lawrence Erlbaum Associates.

[15] WILLEMS, N., R. COLLIER & J.T. HART (1988), "A synthesis scheme for British English intonation", *Journal of the Acoust. Soc. of America* 84, 1250-1261.

ON THE PERCEPTION OF DURATION OF THE CZECH VOWELS

Jevgenij Timofejev

Pedagogická fakulta
Hradec Králové, Czechoslovakia

ABSTRACT

An experiment concerning the perception of the Czech phonological duration is described in this paper. All Czech vowels synthesized by HV 02 synthesizer have been used in the experiment. The results have been compared with natural speech signals statistics.

1. INTRODUCTION

The Czech phonological duration has been studied from various points of view /J. Chlumský [3], P. Janota [4], B. Hála [1]/. The described experiment is aimed at physical relevance of Czech duration. The experiment does not include wider language parameters and has been limited with isolated words synthesized by HV 02 peripherals /Tesla Electronic Research Institute in Prague/. It is one of the possible ways for studying subject and it is based upon the precise quantity of the synthesized speech signal.

2. PROCEDURE

A list of Czech meaningful words has been prepared and synthesized by HV 02. There were two criteria for

the words preparation. All the Czech vowel were taken into consideration their duration covering the range of 30 to 240 msec with 20 msec stepping:

a ... 240 - 60
e ... 190 - 30
u ... 180 - 20
o ... 200 - 40
i ... 170 - 10

The perception test has been aimed at two specific spheres of interest. First there was an attempt to define the boundary where the phonological duration changes the meaning of the word: pás /"belt" with a long vowel "á"/ and pas /"passport" with short vowel "a"/.

A list of words presenting the entire range of temporal realisation has been listed and evaluated both in gradually descending and gradually increasing scale. Secondly there was an attempt to determine the relative temporal differences in physical realisation of vowels, which can be perceived by native Czechs as phonological temporal difference. Pairs of words with temporal contrast have been listened and estimated by respondents. The temporal contrast was both gradually descending

and gradually increasing in the range of 20 to 180 msec with stepping of 20 msec, for example bór - bor /200 - 40/. The pause between the words was 1 sec and 3 sec between pairs of words. All the synthesized Czech words have been tested in such a way. There were more than 70 respondents, native Czechs, involved into perception experiment. They were requested to write the words.

3. SOME RESULTS

3.1. Within the range of temporal parameters of the tested words the results cannot be regarded as a compact value. The boundaries between long and short vowels are rather dispersed and are perceived for various vowels as follows:

á - a : 200 - 100 msec
é - e : 170 - 70 msec
ý - u : 180 - 60 msec
ó - o : 180 - 80 msec
í - i : 150 - 50 msec

The percentage of boundary identification may be presented by following data:

á - a
200 - 180 : 2,6 %
180 - 160 : 16,0 %
160 - 140 : 56,0 %
140 - 120 : 20,0 %
120 - 100 : 5,3 %

é - e
170 - 150 : 2,6 %
150 - 130 : 14,6 %
130 - 110 : 54,6 %
110 - 90 : 26,6 %
90 - 70 : 1,3 %

ú - u
160 - 140 : 8,0 %
140 - 120 : 32,0 %
120 - 100 : 49,2 %
100 - 80 : 9,2 %
80 - 60 : 1,3 %

ó - o
180 - 160 : 2,6 %
160 - 140 : 22,0 %
140 - 120 : 40,0 %
120 - 100 : 29,2 %
100 - 80 : 4,0 %

í - i
150 - 130 : 4,0 %
130 - 110 : 20,0 %
110 - 90 : 29,2 %
90 - 70 : 44,0 %
70 - 50 : 2,6 %

The verification tests have shown upon no significant differences. The percentage of identification during two various ways of words presentation /first in gradually descending scale and then in gradually increasing one/ has been the same. These data and results have been compared with prof. B. Hála statistics [2]. The results of the comparison have indicated upon the narrow correlation between the perception test results from one side and highly reliable statistics of Czechs vowels length as presented by prof. B. Hála from the other side. We shall first discuss the temporal parameters of long vowels.

3.2. There were no responses which showed that the results of the listening test were not in any way in accordance with existing knowledge of the Czech long vowels parameters [2]. This remark is valuable for all Czech long vowels. We have not registered any word with "long" vowel, which exceeds the down limit of duration valuable for long vowels. These results are illustrated with figure 1.

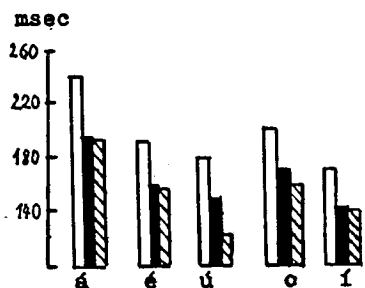


Fig. 1

□ the mean value of long vowels statistics
 ■ the results of the perception test
 ▨ the down limit of long vowels statistics

As the "short" vowels are concerned, we have registered no word with vowel, which is longer than the upper limit of short vowel statistics available in 2 - figure 2:

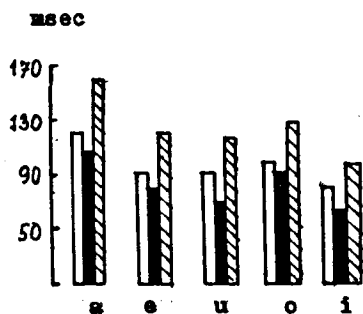


Fig. 2

The figure 2 illustrates the correlation between the mean values for long and f for short Czech vowels on the one hand and the most contrastive span of duration /from 30 to 56 % of responds/ evaluated as a bound-

dary during the perception experiment on the other hand:

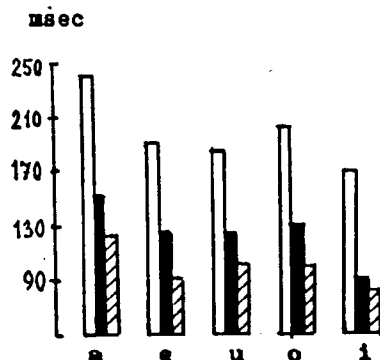


Fig. 3

□ the mean value of long vowels statistics
 ■ the results of the perception test
 ▨ the mean value of short vowels statistics

3.3. The above discussion covers the absolute values of Czech vocal duration. The relative values have been estimated on the bases of the second part of our listening test, where the pairs of two words were evaluated. We have gained rather stable results showing that the span of 100 msec is the most important for phonological quantity identification - figure 4:

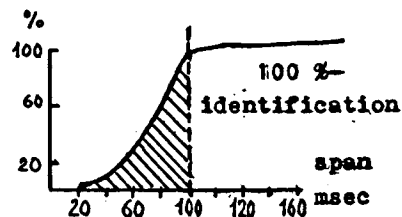


Fig. 4

4. CONCLUSION

There is no simple correlation between the natural and synthetic speech signal, neither isolated words can simulate the real language realisation. We have presented one of the possible ways of problem presentation. The results may be of interest both for natural and synthetic speech investigation.

5. REFERENCES

- [1] HÁLA, B. /1941/, "Akustická podstata samohlásek", Praha
- [2] HÁLA, B. /1962/, "Uvedení do fonetiky češtiny na obecně fonetickém základě", Praha
- [3] CHLUMSKÝ, J. /1928/, "Česká kvantita, melodie a přízvuk", Praha
- [4] JANOŤA, P. /1967/, "An experiment concerning the perception of stress by Czech listeners", Praha

DIFFERENTIATING BETWEEN PHONETIC AND PHONOLOGICAL PROCESSES: THE CASE OF NASALIZATION

M.J. Solé and J.J. Ohala

Laboratori de Fonètica, Universitat Autònoma de Barcelona, Spain
Department of Linguistics, University of Alberta, Edmonton, Canada

ABSTRACT

The aim of this paper is to differentiate between hardwired or unintended phonetic processes and phonological or language specific processes. Cross-linguistic data on coarticulatory effects of nasalization across different speech rates in American English and Spanish were obtained. The data show that in American English vowel nasalization varies (inversely) with speech rate; whereas in Spanish nasalizations has a constant duration across speech rates. Spanish nasalization is modeled as a constant additive component (dependent on vowel height), and American English nasalization as a multiplicative component. It is argued that vowel nasalization in Spanish is an unintended, vocal tract constraint unaffected by higher-level speaking rate effects, and that nasalization in American English is a phonological effect, intentionally implemented by the speaker.

1. INTRODUCTION

The aim of the study is to devise a model that can quantify and differentiate between (i) "hardwired" phonetic processes due to the mechanics of speech, and (ii) phonological or language-specific processes intentionally implemented by the speaker. The model will be formulated on the basis of original data on coarticulatory effects of nasalization in Spanish and American English.

Cross-linguistic studies using a variety of techniques [1], [3], [5], [6] show that lowering of the velum necessarily overlaps the articulatory configuration of preceding vowels and that the period of overlap varies across languages. In the present study it is hypothesized that this variation is due to the different nature of nasalization in different languages. Thus, in some languages, such as Swedish or Spanish, vowel nasalization seems to be an online or hard-wired phenomenon,

mechanically linked to the presence of a nasal consonant and by-product of the temporal organization of motor commands; whereas in some other languages, such as American English, vowel nasalization is not mechanical but intended, part of the linguistic organization of speech motor commands.

To differentiate between on-line and intended nasalization an experiment was conducted where rate of speech (i.e., time to achieve articulatory targets) was varied and its effects on velum movement (i.e., duration of vowel nasalization) were observed for Spanish and American English. This information will allow to determine which portion of the vowel (the oral or the nasalized portion) is affected when rate of speech is varied, and it will be possible to establish if the vowel is articulatorily specified as oral (with mechanical nasalization) or as nasalized. Speech rate is an intended, higher level adjustment. If the vowel is targeted as nasalized, and consequently nasalization is higher level, nasalization is expected to vary (inversely) with speech rate. If the vowel is targeted as oral, nasalization will be due to vocal tract constraints, and the nasalized portion will not vary in different rates of speech (or it will vary as a function of the velocity of the articulatory gesture).

2. METHOD

Three speakers of American English and three speakers of peninsular Spanish read a randomized word list consisting of all possible combinations of $C_1V_1V_2C_2$, where $C_1 = t, n$; $V_1 = i, a$; $V_2 = i, e, a$, $C_2 = t, n$. The carrier sentence for English speakers was "Guess ___soon". The Spanish carrier sentence was "Dos ___son" ("They are two ___"). The subjects were asked to read the 24 test sentences twice at five different speech rates: 1. overarticulated, overslow speech ("as if talking to a deaf person who was lip

reading"), 2. careful, slow speech ("as if reading out loud to a formal audience in a big lecture hall"), 3. normal conversational speech, 4. fast speech, 5. underarticulated, overfast speech ("as fast as you possibly can"). The four most equidistant speech rates were studied for every speaker.

To track the time-varying positions of the velum a Nasograph (see [7] for a description) was inserted into the subjects' nasal cavity and pharynx and the traces of velopharyngeal port opening/closing and acoustic waveform were obtained on a Siemens Oscillomink chart recording device in the standard way [3], [7]. Measurements of vowel duration and timing of soft palate lowering before nasal consonants were done in $[^hVVN]$ sequences. The measurements of vowel duration were done 1) for the aspiration period [h], 2) for V1, and 3) for V2. The method used in determining onset of velum lowering was to consider movement to begin at the time when the velocity function (slope) crosses a noise band (defined as 10% of the highest peak velocities of the velar movement gestures for each speaker) around zero. For multistage velar gestures - usually those involving a low vowel - the first lowering gesture exceeded the noise band and, consequently, velum lowering due to vowel height was included in all cases. Measurements were done by hand on Oscilomink traces.

3. RESULTS

The results for the measurements for American English are presented in Figs. 1 and 2, which show the mean duration of the oral and nasalized portion of the vowel sequence (including the aspiration period) for [iV] and [aV] sequences respectively. The onset of velic lowering is marked 0 on the abscissa and segments appearing right of 0 are the nasalized portions of the vowel sequence. Varying speech rates appear on the ordinate. Speech rate was plotted by determining the average duration for the vowel sequences. The oral portion of the $[^hVV]$ English sequences in Figs. 1 and 2 corresponds to the aspiration period as obtained from the acoustic waveform (mean oral portion for speaker JJ= 59.9ms, AV=52.3ms, MN= 49.7ms; mean aspiration period for speaker JJ= 58.3ms, AV=61.5ms, MN=49.8ms). Furthermore, in some cases velic lowering begins during the aspiration period resulting in a nasalized aspiration, [ʰ]. This

indicates that in American English the voiced portion of the vowel sequence is completely nasalized.

Figs. 3 and 4 show the results for Spanish. It can be observed that the nasalization period in Spanish shows a roughly constant duration across different speech rates. Only in the fastest speech rates some speakers (MJ [iV]; PR [iV]; JR [iV], [aV]) succeed in reducing the nasalized portion. This indicates that under unusually fast speech conditions speakers might increase the velocity of the velic lowering movement.

Comparison of Figs. 3 and 4 shows a longer nasalization period for [aV] sequences than for [iV] sequences. It seems reasonable to suggest that nasalization has a constant duration in both cases and that differences are due to further low-level adjustments due to vowel height [1], [6], [7]. The fact that for [iV] sequences the constant nasalized period (k) across speech rates for the different speakers (MJ, $k=91.6ms$; PR, $k=124.3ms$; JR, $k=97.6ms$) is longer than the minimum transitional period (40ms for [ii] sequences) is due to the fact that the computed k value is the mean of the values for different V2. This indicates that the height of V2 also has an effect on velum lowering which is presently under study. This effect is less evident for [aV] sequences.

To sum up, the results in Figs. 3 and 4 suggest that nasalization is a constant value across speech rates, and that the oral portion of the vowel varies inversely as a function of speech rate. Thus, Spanish vowels can be said to be targeted as oral and nasalization is the result of a physiological time constraint.

4. MODELING NASALIZATION IN AMERICAN ENGLISH AND SPANISH.

In American English vowel sequences followed by a nasal are nasalized throughout. Thus vowel nasalization can be modeled as a multiplicative effect (multiplied by a factor of 1):

$a=d$
where a equals the nasalized portion, and d equals the total duration of the vowel sequence (excluding aspiration).

In peninsular Spanish the nasalized portion can be modeled as a constant value (k) which depends on the height of V1 (v) (and possibly V2). The oral portion can be

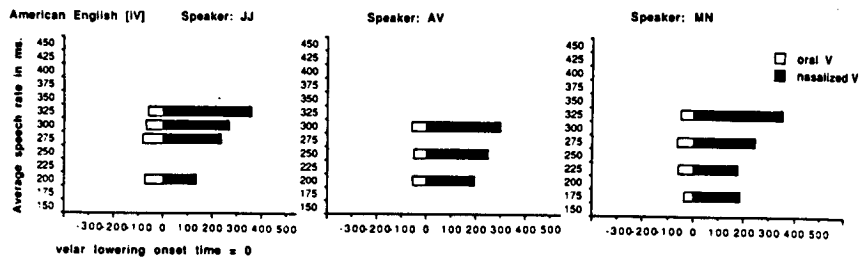


Fig. 1. Mean duration in ms. of the oral and nasalized portions of the vowel sequence [hiV] for American English on the abscissa. Onset of velic lowering is marked time 0. Average speech rate in ms. appears on the ordinate.

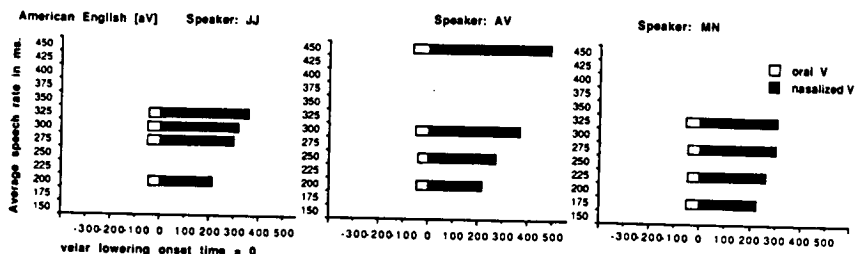


Fig. 2. Mean duration in ms. of the oral and nasalized portions of the vowel sequence [haV] for American English on the abscissa. Onset of velic lowering is marked time 0. Average speech rate in ms. appears on the ordinate.

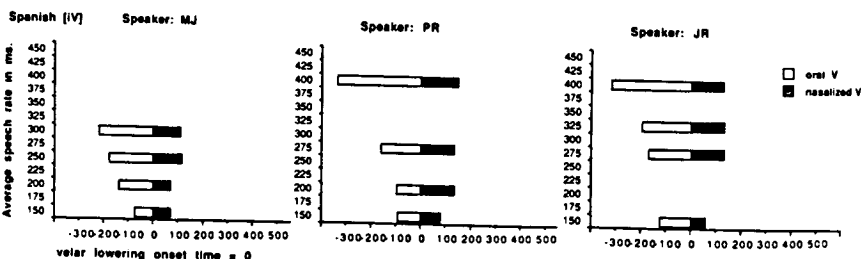


Fig. 3. Mean duration in ms. of the oral and nasalized portions of the vowel sequence [hiV] for Spanish on the abscissa. Onset of velic lowering is marked time 0. Average speech rate in ms. appears on the ordinate.

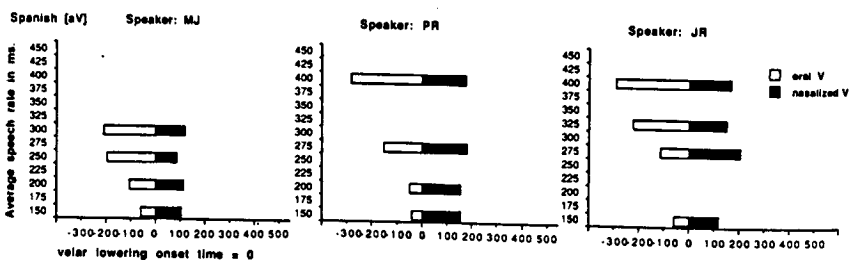


Fig. 4. Mean duration in ms. of the oral and nasalized portions of the vowel sequence [haV] for Spanish on the abscissa. Onset of velic lowering is marked time 0. Average speech rate in ms. appears on the ordinate.

modeled as the vowel's duration minus $k(v)$:

$$a = k(v)$$

$$b = d - k(v)$$

where a equals the nasalized portion of the vowel; $k(v)$ equals a constant value (incompressible beyond 40 ms) which depends on the height of V1 (and V2); b equals the oral portion of the vowel, and d equals the total duration of the vowel sequence. Thus, vowel nasalization in Spanish can be modeled as a constant which is added to vowel duration after speaking rate effects have been applied.

The working of the model for observed vs predicted values is currently under study.

5. DISCUSSION

The fact that nasalization in Spanish is an additive constant number of milliseconds indicates that it may take $k(v)$ ms to establish the articulatory configuration for the nasal consonant. Thus, nasalization in Spanish can be considered as an unintended hardware effect which is added to higher level adjustments such as speaking rate. If speakers were using vowel nasalization distinctively one would expect that the nasalized portion in one rate of speech would differ from that in another rate by an amount proportional to the difference of the duration of the vowels, rather than by a constant number of milliseconds across all rates. This is the case for American English. It seems reasonable to hypothesize [2] that multiplicative effects are phonemic and occur prior to additive ones, which reflect constraints of speech production. Since no additive component was observed for vowel nasalization in American English, it can be deduced that nasalization does not occur automatically but that it has achieved the status of a phonological rule, intentionally implemented by the speaker.

The existing models on the timing of vowel nasalization [4], [8] do not target velum position for vowels, but just for preceding and following consonants (thus, a vowel in a $[t_N]$ context is thoroughly nasalized in the transition between the two targets). According to our data these models are adequate for American English, where vowels are intentionally nasalized, but do not accurately simulate the behavior of Spanish vowels in the same context. This indicates that a timing model must be language specific.

6. CONCLUSION

The universality of vowel nasalization before nasal consonants demands an explanation that refers to some universal properties of human beings. The transition time in velic port opening is most likely the origin of vowel nasalization. However, the same phonetic effect might be unintended and low-level in one language (Spanish), and it might have been phonologized, and therefore be part of the language specific timing instructions in another language (American English). Moreover, the large number of languages (e.g. French, Hindi, Portuguese) that lose a nasal consonant distinction only to replace it with distinctive vowel nasalization indicates that phonetic nasalization can be perceived and then exploited by language users. The different nature of the same phonetic phenomenon proves the need to interpret phonetic data in terms of their phonological behavior if we are to provide an accurate account of the hardwired and softwired components of speech in different languages and implement them in automatic speech technology.

6. REFERENCES

- [1] BELL-BERTI, F., T. BAER, K.S. HARRIS & S. NIIMI (1979), "Coarticulatory effects of vowel quality on velar function". *Phonetica*, 36, 187-193.
- [2] CAISSE, M. (1988), "Vowel duration in American English", PhD Diss., University of California, Berkeley.
- [3] CLUMECK, H. (1976), "Patterns of soft palate movements in six languages", *Journal of Phonetics*, 4, 337-351.
- [4] HENKE, W. (1966), "Dynamic articulatory model of speech production using computer simulation". PhD Diss, M.I.T.
- [5] KRAKOW, R.A. (1989), "The articulatory organization of syllables: A kinematic analysis of labial and velar gestures", PhD Diss, Yale University.
- [6] MOLL, K.L. & R.G. DANILOFF (1971), Investigation of the timing of velar movements during speech", *Journal of the Acoustical Society of America*, 50, 678-684.
- [7] OHALA, J.J. (1971), "Monitoring soft palate activity in speech", *Journal of the Acoustical Society of America*, 50, 140 (Abstract). Printed in full in *Project on Linguistic Analysis*, 13, J01-J015.
- [8] VAISSIERE, J. (1989), Prediction of articulatory movement of the velum from phonetic input. In O. Fujimura, ed., *Articulatory organization: from phonology to speech signal*. In press.

8. La vélaire reste dans sa réalisation d'occlusive neutre, cf (5), [biŋá], lorsqu'elle se trouve sous un pont d'harmonie ATR, harmonie qui n'est déclenchée en mooré que par les voyelles hautes [i, u, ɪ, ũ]. Mais dans tous les autres cas où elle n'est ni initiale ni gouverneur, elle se spirantise en [ʁ], devenant donc du même coup uvulaire. Si elle était restée vélaire, la spirante aurait été représentée par la seule voyelle froide en position non-nucléaire. Mais sa localisation a été abaissée jusqu'à la zone uvulaire, manifestant ainsi l'adjonction de l'élément A⁺ en tant qu'opérateur et fonctionnant en quelque sorte comme un schwa consonantique:

$$\begin{array}{c} [ʁ] = \nu^{\circ} \\ | \\ A^{+} \end{array}$$

Il est particulièrement important de relever que le mooré fonctionne avec une propagation de A⁺ tout aussi contraignante que celle de l'élément ATR et en relation d'exclusion mutuelle avec elle. L'harmonie en A⁺ n'est déclenchée que par un segment [a] final de mot (donc par A⁺ tête) et seulement en l'absence d'harmonie ATR : c'est elle qui, en (4) transforme /bɪn + ga/ en [béngà] et /kò + ga/, "liquide", "aqueux", en [kwaàwá]. Les contextes où /g/ non-gouverneur se spirantise en s'adjoignant A⁺ (même en l'absence de cet élément dans son entourage, cf "le fait d'apprendre" en (5)) connaissent donc la même interdiction de le faire en présence d'harmonie ATR. On fait ainsi ressortir que, (a) une harmonie ATR qui s'instaure de noyau à noyau n'est pas sans incidence sur la consonne qui les sépare puisque /g/ y est ici sensible; et (b), l'antinomie de A⁺ et d'ATR se confirme, jusque dans la spirantisation autorisée/exclue de la consonne.

9. Le problème, majeur, qui reste en suspens, a trait au voisement: comment, s'ils sont des segments neutres (sans L⁻), [g] et [ʁ] peuvent-ils se dévoiser en (7) et (8)? Sa solution passe par la reconnaissance d'une distinction fondamentale à laquelle il est souvent fait référence (de manière assez trompeuse)

sous les termes de voisement spontané et de voisement phonologique, ce qui laisse entendre que le voisement spontané ne serait pas phonologique. Nous voulons démontrer que le voisement dit spontané, s'il n'est peut-être pas inscrit dans la structure en éléments des segments concernés, n'en est pas pour autant non-phonologique.

10. Appelons "sourd" le segment [k] que la présence de H⁻ rend gouverneur, et "sonore" le segment [g] que L⁻ rend également gouverneur. Les cas restants sont ceux où le segment n'est pas gouverneur: il apparaît sous ses variantes [g] et [ʁ] que nous dirons "voisées" et [g] et [ʁ] que nous dirons "non-voisées". Ces deux paires de variantes sont en distribution complémentaire: les voisées apparaissent en position d'attaques gouvernées (par un noyau plein, identifié) et les non-voisées en position d'attaques non-gouvernées (car leur noyau est vide et final). Deux interprétations sont alors envisageables.

La première consiste à dire que la distinction de voisement n'a pas à être inscrite dans la constitution interne des segments: elle est entièrement déductible du gouvernement ou non de l'attaque par son noyau.

La seconde consiste à dire que le voisement du segment d'attaque est un caractère transmis par le segment nucléaire, et qu'il n'y a absence de transmission qu'en l'absence de segment nucléaire. Il faudrait alors reconnaître que, même dans le cas non-marqué où les voyelles ne font jouer aucune distinction de voisement, elles comportent un élément de voisement spécifique, et distinct de H⁻ et de L⁻.

11. Quelle que soit l'interprétation choisie, les exemples (9) montrent que la distinction entre "voisée" et "sonore" s'impose: bien qu'appartenant à une attaque non-gouvernée, [g] reste gouverneur et donc sonore dans [béng] ou [léng], quand il ne va pas jusqu'à prendre à la nasale, non pas seulement son L⁻, mais la totalité de ses éléments pour donner ou [béng] ou [léng].

12. Au total, la répartition des différentes métamorphoses de /g/ en mooré semble

indiquer qu'il faille reconnaître comme distincts:

- des segments *sourds*, construits avec H⁻ (ici [k]).
- des segments *sonores*, construits avec L⁻ (ici [g] gouverneur d'une nasale).
- des segments *neutres*, construits sans aucun des deux éléments laryngés, et qui selon leur contexte d'apparition, se réalisent comme *voisés*, (ici [g] ou [ʁ]), ou comme *non-voisés*, (ici [g] ou [ʁ]).

13. Généralisations:

(a) Les cas de neutralisation du voisement des occlusives finales dans des langues telles que l'allemand pourraient être l'indice du fait que leur série dite sonore serait en fait, comme en mooré, une série neutre, non-marquée, s'opposant à une série de sourdes marquées par l'élément H⁻.

(b) La neutralisation du voisement des obstruents dans les langues qui, comme le mahou, n'admettent que des sonores après consonne nasale, (cf bā kwò, "derrière un cabri," mais sɔŋ gwò, "derrière une gazelle," Bamba [1]) pourrait être l'indice du fait que leur série dite sourde serait en fait, contrairement au mooré, une série neutre, non-marquée,

s'opposant à une série de sonores marquées par l'élément L⁻.

(c) Dans des langues comme le coréen, où les occlusives de la série neutre prennent à l'intervocalique une variante voisée, (/koki/ -> [kogi], "viande,") il pourrait s'agir d'un voisement comparable à celui de la vélaire neutre du mooré (mais répondant à un conditionnement différent) sans qu'ait à intervenir l'élément L⁻.

REFERENCES

- [1] BAMBA, M., (1984), *Etudes phonologiques du mahou*, Mémoire de maîtrise (non publié), Université du Québec à Montréal.
- [2] KABORE, R. (1985), *Essai d'analyse de la langue moore*, (parler de Ouagadougou), Université Paris 7, D.R.L., coll. ERA 642.
- [3] HERAULT, G. (1989), Les rections syllabiques en soninké, *Linguistique Africaine* 3, 43-90.
- [4] KAYE, J. (1989), "Coda" Licensing, ms, SOAS, University of London.
- [5] RENNISON, J. (to appear), On the elements of phonological representations: the evidence from vowel systems and vowel processes, *Folia Linguistica*.

ABOUT THE PHONETICS / PHONEMICS INTERFACE: THE CASE OF KRIOL STRESS.

B. Laks et A. Kihm

CNRS & Université Paris VIII, France

ABSTRACT.

This paper explores the interface between phonetics and phonology in the domain of stress. Stress position in Kriyol is not directly predictable from the physical parameters (f₀, i & duration) affecting the syllable. It is shown however that normalizing function can be devised such that a neuronal network with no hidden units will learn it after only 5 trials. The network is then able to predict stress position 70% of the cases.

1. PHONETICS vs. PHONEMICS

1.1. The two approaches to stress.

There are two approaches to stress, phonetical and phonemical, which are not always easy to reconcile. Indeed, the phonemical approach implies that it must be possible to uncover regular or nearly regular stress patterns, which consist in the recurrence of stress on the same location for at least a lexical class. The phonemical standpoint is thus intrinsically discontinuous, as it should be. According to the phonetical approach, on the other hand, stress is a prominence that is achieved through one syllable being maximally intense, and/or maximally high, and/or maximally long, depending on the language under consideration. Phonetic stress is thus the result of the continuous variation of three, now parallel, now divergent, scales. Hence, there does not have to be, and actually there rarely is an immediate correspondence between both analyses.

The traditional phonological practice has been to deal with, or to eschew, this difficulty in either one of two ways: either to start with the phonemical pattern - after all one knows or may know by asking native speakers "where stress is" - and try to see which physical parameters most contribute to it; or to start with an instrumental study of the physical parameters, and try to abstract stress from them. This exclusive reliance on one or the other strategy can only be fruitful, however, in those languages which exhibit a

regular correspondence between the continuous variation of at least one of the physical parameters, and the position of stress.

Asserting that this situation ought to be the normal state of affairs is an unwarranted preconception. We will show that the evidence of one language, Kriyol, at least demonstrates that it is not an obligatory state of affairs.

If this is so, the possibility of matching the models and the reality is at stake. One may renounce it and remain content with the by and large dominant separation of phonemic theory (phonology) and phonetic studies. But one may also consider that the match should exist, which then implies a serious exploration of the phonemics-phonetics interface. That there is such an interface proceeds from the necessary, we think, assumption that phonemics and phonetics ultimately address the same object, viz. the sound face of language, at different levels of abstraction.

1.2. Presenting the language

Kriyol is a Portuguese based creole language spoken in Guiné-Bissau and Casamance. A number of studies have been devoted to its syntax (see in particular Wilson 1962; Kihm 1980). Kriyol phonology, in contrast, is poorly studied yet (see Wilson 1962; Mboj 1979; Kihm 1986). The suprasegmentals, in particular, have only been cursorily considered. Kihm and Laks (1989a) is a first attempt, where we established that Kriyol is to be analysed as a stress language, like its lexifier Portuguese and like, or so it seems, the Atlantic languages (Manjaku, Balanta, Diola, possibly Wolof) that constitute both its substratum and its adstratum.

1.3. The experimental setting.

The study was conducted using a sizeable data base of nearly 400 forms, lexical items and phrases. The forms were chosen so they would constitute a representative sample of

the lexicon in terms of (a) stress patterns; (b) number of syllables; (c) syntactic category and type of construction; (d) origin (Creole, Portuguese, or African). Each form was repeated twice consecutively by a native speaker (25, male), recorded, and run through a computerized melody analyser (Philippe Martin's analyser available at the UFLinguistique in Paris7 University). We thus obtained the value of the relevant parameters intensity (db), pitch (hz), and duration (cs) for each syllable.

1.4. The problem.

The gross result of this work was that it is sometimes possible, but very often impossible to observe a regular correspondence between the relative values of the parameters and the location of the stressed syllable which is always intuitively clear. Consider the following examples (the stressed vowel is capitalized):

(1)kansera 'fatigue'	kan	sE	ra
(2nd occ.)	db	29	35
	hz	104	119
	cs	17	30
			19
(2)kansera 'fatigue'	kan	sE	ra
(1st occ.)	db	34	40
	hz	138	121
	cs	23	35
			24
3)nobresa 'youth'	no	brE	sa
	db	33	38
	hz	107	118
	cs	17	18
			21

In (1), all three parameters converge at their maximal value on the location of the stressed syllable, we designate this schema as [111]. This is by no means the most frequent schema, though. The remainder of cases is distributed among the other schemata, [101] (fit of db and cs, but not hz with the stressed syllable) as in (2), or [100] (fit of db only) as in (3). One even finds cases where the stressed syllable is neither more intense, nor higher, nor longer (schema [000]). Moreover, (1) vs. (2) shows that the same, identically stressed form may be assigned to different schemata when repeated at a few seconds interval by the same speaker. This confirms the fact that the variation is truly inherent, and cannot be explained away by the phonetic environment, or at least by any regularly recurring component of this environment. How shall we reconcile such an extreme phonetic variation with the regular phonemic stress patterns that are part of the native speaker's knowledge of his/her language? (By this, we mean the native speaker's perceptive ability to reconstruct discrete

patterns out of a continuous, relatively chaotic sound input, as well as his/her productive ability to embody these patterns into an equally continuous, relatively chaotic phonic output.) What interface may be shared by two so widely divergent objects? Before we try and answer this question, a more detailed presentation of the phonemics and phonetics of Kriyol is in order.

2. A PHONEMIC VIEW OF KRIYOL STRESS.

Although we tested both lexical items and phrases (NPs and sentences), only the former are considered in this study. Stress patterns in Kriyol are distributed according to the lexical category of the item. The following set of rules covers almost all cases:

(4)(a) Nouns and adjectives are stressed on the ultimate syllable if it is heavy (e.g. kacur 'dog'/ka'cur/), on the penultimate syllable otherwise (e.g. tabanka 'village' /ta'banka/, bonitu 'nice' /bo'niku/).

(b) Verbs are stressed on the final syllable (e.g. rispundi 'to answer' /rispundi/, mistura 'to mix'/mistura/).

3. A PHONETIC VIEW.

As already indicated, three physical parameters contribute to stress, viz. intensity (db), pitch (hz), and duration (cs). None of them in isolation is sufficient to account for the location of stress in a given item. We therefore decided to consider all three of them simultaneously, which led us to the mentioned observation that it is only in a minority of cases that the maximal values on each line fall down in a neat column supporting the stressed syllable as in (1). What we have in mind is a quasi-autosegmental framework such as parameter values supported by three autosegmental tiers are anchored to a skeletal line made of slots. Time synchronization of the different tiers will ensure that association lines do not cross. Actually, all the logically possible mismatches are attested, resulting in 8 categories or schemata, [111], [110], [101], [011], [100], [010], [001], and [000]. All possibilities are thus realized, from maximal contribution of all three parameters ([111]) to apparently no contribution at all ([000]). The figure below gives the distribution of all schemata.

(a)	[111]	23.60%	(b) [1XX]	66.29%
	[110]	6.37%	(c) [X1X]	42.70%
	[101]	20.22%	(d) [XX1]	59.93%
	[011]	7.12%		
	[100]	16.10%		
	[010]	5.62%		
	[001]	8.99%		
	[000]	11.99%		

4. PHONEMICS AND PHONETICS COMPARED.

There is therefore no overall one-to-one correlation between the phonetic facts and the phonemic stress patterns. (Recall that our schemata are tokens, whereas stress patterns are types.) Actually, such a mismatch is expected. Indeed, stress is a clear-cut phenomenon that may be expressed as discrete values, e.g. (for one-stress languages) 1 and 0. Phonetic parameters, on the other hand, are continuous scales that vary in an unpredictable (if not unaccountable) fashion each time a particular utterance occurs. It is standard practice to dismiss this variation as non-linguistic by considering it in the same way as pure individual variation in loudness, highness, and duration. We contend that, by so doing, one denies oneself the opportunity of showing that phonemics (what the speaker-hearer knows) and phonetics (what s/he effectively produces or perceives) have anything in common, as it seems obvious that they should.

We assume, then, that both the phonemic patterns and the phonetic facts are faithful images of the phonological reality of the lexical items, which means that one can start from either one or from both, following a top-bottom or bottom-top procedure, and converge onto an interface. Note in passing that such a procedure is the only one that is really safe and fruitful with unknown or poorly known languages, Kriyol being an instance of the latter category. The problem is therefore to find an integration principle for the phonetic parameters allowing this interface of a discrete and a continuous domain to stand. Let us state our general hypothesis first. We assume that stress as it patterns phonemically represents the best possible compromise of the actual values of the three phonetic parameters. By 'best possible compromise', we mean to say that those actual values are computed each time they are realized, resulting in a phonetic figure (in the Gestalt sense of the term) that can be matched against the stress pattern. This is the integration mentioned above. The matching is always approximate, sometimes quite good, sometimes very bad, but it is an integral part of stress as a phenomenon which associates two types of prominence, one cognitive and absolute, the other physical and relative.

5. A PRACTICAL SOLUTION

The first step consists in normalizing the parameter values, so that the difference between the lower and the higher values is maximized. Our procedure was to map the

indeterminate scale of real variation onto a [0...1] scale, using the following formula:

(5) Normalization formula for the phonetic parameters: Let P be a phonetic parameter, x a real value of this parameter, and y a normalized value. $x <\epsilon> \{a, \dots, b\}$, where $\{a, \dots, b\}$ is the set of the possible values of the parameter; $y <\epsilon> \{a', \dots, b'\}$ where $\{a', \dots, b'\}$ is the set of the normalized values of the parameter. The general formula is then: $y = f(x) = a' + (x - b) \cdot (b' - a') / (b - a)$

For each particular occurrence of a lexical item, there is a maximal value of P x max and a minimal value of P x min. As we chose $\{0, \dots, 1\}$ as the set of values for y, y max and y min equal respectively 1 and 0. Given this, the general formula rewrites as: $y = f(x) = 0 + (x - x_{min}) / (x_{max} - x_{min})$

Data are not modified by the normalization, which is no more than a quasi-optical means of focussing on the regularity that is embedded in the data. Let us apply this formula to the examples in (1-3). The result is given below (normalized values are on the right):

(6) kansera (2nd occurrence)

	kan	sE	ra	
db	29	35	24	0.45 1 0
hz	104	119	85	0.55 1 0
cs	17	30	19	0 1 0.15

(7) kansera (1st occurrence)

	kan	sE	ra	
db	34	40	30	0.40 1 0
hz	138	121	132	1 0 0.64
cs	23	35	24	0 1 0.08

(8) nobresa

	no	brE	sa	
db	33	38	30	0.37 1 0
hz	107	118	134	0 0.40 1
cs	17	18	21	0 0.25 1

The problem is now to integrate the normalized values in a way that gives us a number series that is parallel to the stress pattern. Two functions come to mind immediately, the Product function and the Sum function. Let us apply both to our examples (the sum is renormalized by dividing each number by the highest value in order to obtain a [0...1] scale again):

(9) kan sE ra (2nd occurrence)

	kan	sE	ra
P	0	1	0
S	0.33	1	0.05

(10) kan sE ra (1st occurrence)

	kan	sE	ra
P	0	0	0
S	0.7	1	0.36

(11) no brE sa

	no	brE	sa
P	0	0.1	1
S	0.18	0.82	1

Both functions yield the right result in (9) with a [111] item; only the sum function yields the right result in (10), a [101] item; finally, no function works in (11), a [100] item. The first result is unsurprising and could not be different. It is not so, however, with the two other results, as shown by the following tokens:

(12) lagartu 'crocodile'

	la	gAr	tu
P	0	0.5	0.3
S	0.35	1	0.52

(13) bajudas 'young girls'

	ba	JU	das
P	0	0.6	0
S	0.7	1	0.38

In (12) and (13), both functions yield the correct configuration. The only serious departure is in (12) where a potential secondary stress is assigned to the last syllable, contrary to fact.

Actually, the difficulty lies in the significance of the functions Product and Sum. Product implies that a syllable that is lowest (i.e. = 0) for at least one parameter cannot be stressed (since $n \times 0 = 0$). This is not true, as is apparent in (10) where the second syllable of kansera is stressed although it is lowest on the hz line. Sum, on the other hand, implies that stress results from the cumulation of the parameter values. This again is not always the case.

6. PROSPECTIVE SOLUTION.

Central in the present approach are the notions of interface and of best compromise. By interface, we mean some kind of device for matching continuous phonetic reality with discontinuous phonemic representations. Here, we only tested two instances, viz. Product and Sum. In fact, every possible combinatory function ought to be tested, in accordance with the idea that, given a set of parameters A, B, C, ..., and values of these parameters a, b, c, ...: Best compromise = $f(a \cdot b \cdot c \dots)$.

From our data, it is also obvious that all parameters do not contribute equally to the final matching. Actually, we think that only by assigning each parameter a specific weight, shall we come closer to a modelization of the notion of best compromise. Let x, y, z... be the weights assigned to parameters A, B, C, ..., our study has now to deal with two different sets of unknowns: function and

relative weight, so that: Best compromise = $f(ax, by, cz \dots)$. Given this, two secondary assumptions can be sustained:

(a) Relative parameter weights and combinatory function are fixed, i.e. language-specific. This roughly corresponds to the standard typological hypothesis that stress is mainly linked to intensity, or pitch, or duration depending on the language.

(b) As our data seem to show, relative parameter weight and combinatory function have to be computed for each occurrence. This leads us to an interesting question: What cognitive device(s) have to be assumed in order to achieve such a computation? Recent research has emphasized the idea that cognitive problems are a special subset of computational problems. It is therefore appealing to use neuronal networks as a modeling tool for our problem. We will be interested in seeing how such a network auto-organizes to match the phonetic cues with the stress position, and on what kind of function it will settle. As a first step, we designed a network including 24 input units (i.e. 8 db, hz, and cs triplets, as 8 syllables is the maximum length of our lexical items; with shorter words, exceeding triplets are clamped to 0). The output consists of 8 units. For learning, we used the standard back-propagation algorithm. It is worth noting that the threshold function associated with units can mirror the weight parameter mentioned before. With only feed-forward connexions and no hidden units, learning is completed after only five examples, and the network outputs the correct answer in more than 70% of unlearned cases. Obviously, such a design is too poor to efficiently cover the problem at hand. We are currently running simulations with constraint competition, thereby considering triplets as non independent figures, and having them interact to produce the correct output. Two formal solutions are conceivable. One is to set a row of fully connected hidden units each of them summing up a triplet. The other is to let the inputs compete with each other by setting lateral connexions. The latter solution is time consuming. More complex connecting patterns have to be tested before we can claim a neuronal network is able to find the kind of function we are looking for. The results obtained so far, however, show that this is indeed a promising line of research, leading to a cognitive bridging of the phonetic-phonemic gap.

M. Klein et B. Laks

C.N.R.S. et Paris VIII

ABSTRACT

In this paper we present the first step of a research program. Our purpose is to evaluate and compare production systems with connexionist machines. The research focuses on phonic syllabation in relation with the graphic parsing of isolated French words. A first implementation of phonic/graphic hyphenation is presented.

1. PROBLEMATIQUE DE LA SYLLABE

On sait qu'il n'existe pas un indice phonétique clair et parfaitement individualisé de la discontinuité syllabique. Alors que la plupart des modèles phonologiques récents font, à différent niveaux d'analyse, centralement appel à une notion de découpe syllabique, la phonétique propose quant à elle une liste de paramètres partiellement redondants et partiellement contradictoires, dont aucun pris isolément ne livre cette discontinuité (degré d'aperture, tension, courbe de sonorité, *chest pulse*, implosion/explosion, etc.). Il reste que si les locuteurs peuvent faire montre d'une intuition de découpe syllabique d'une part, et que si les modèles d'explication phonologique démontrent la pertinence de la syllabe d'autre part, on peut se fixer pour objectif d'explicitier, tant en production qu'en réception, les liens entre indices phonétiques de la syllabation et structures phonologiques postulées. Les difficultés évoquées étant pour partie liées au caractère continu du signal et à la multiplicité d'indices contradictoires, les modèles connexionnistes constituent, à priori, de bons candidats pour expliciter le lien phonétique/phonologie s'agissant de la syllabe.

La mise en oeuvre de réseaux de neurones formels pour supporter cette analyse ne préjuge pas d'une modélisation de la syllabation en termes de courbes complexes ou en termes de frontières entre constituants. En phonologie, ces deux approches ont été proposées. On peut ainsi contraster, par exemple, le modèle à segmentation de Selkirk (1982, 1984) et le modèle à courbes de Goldsmith (1990). Les implémentations neuromimétiques quant à elles peuvent produire aussi bien une segmentation du signal (Elman 1990) qu'une intégration du signal sous forme de courbe (Goldsmith 1990). Le choix d'une implémentation connexionniste ne contraint donc pas le modèle phonologique, sous ce rapport au moins. La question de la segmentation a également un statut empirique. En effet, il est non seulement nécessaire de rendre compte de l'individualisation syllabique (à l'aide de frontières ou de zones d'inflexion dans des courbes) mais aussi des phénomènes depuis longtemps relevés d'ambisyllabité et de multiplicité d'analyse pour un même mot. Soit, par exemple, les syllabations différentes du mot catastrophe :

- (1) [ka-tas-tRof]
- (2) [ka-ta-stRof]
- (3) [ka-tas-stRof] (ambisyllabité du [s])

Les systèmes à analyse unique et à coupe sont confrontés ici à des difficultés de traitement qui imposent un affaiblissement de la notion de coupe et/ou de discontinuité. Le concept de règle, central dans toutes les analyses de type parsing, doit être reformulé en termes de régularité incluant la possible existence de solutions multiples et contradictoires.

La reconnaissance de la notion syllabique ne suffit pas à définir son statut. En effet, aux analyses qui proposent une dérivabilité intégrale de l'architecture syllabique et qui postulent des algorithmes de syllabation (Anderson et Ewen 1987), on peut opposer les analyses qui font de l'architecture syllabique une partie de l'information lexicale et contestent ainsi l'existence de tels algorithmes (Encrevé 1988, Kaye, Lowenstamm et Vergnaud 1988). Les réseaux connexionnistes peuvent sans doute apporter une réponse à ces questions dans la mesure où la notion de dérivabilité y a un statut très différent. L'intégration dynamique de contraintes contradictoires et la capacité de réseaux à prendre en compte des phénomènes continus peut permettre de produire une analyse syllabique de la chaîne en termes de courbe ou de frontières sans pour autant postuler des algorithmes explicites de syllabation (Goldsmith 1990). Dans cette perspective, il reste à expliciter la nature de l'information sur laquelle les réseaux travaillent : information strictement segmentale et absence d'information syllabique vs. information segmentale et information syllabique minimale. Le codage initial sur la couche d'entrée incorpore nécessairement cette hypothèse. Ces deux conceptions du lexique sont toutes deux compatibles avec une approche connexionniste (approche *tabula rasa* de type Elman 1990 vs. approche phylogénétique de type Bienenstock 1990). Notre programme de recherche vise en particulier à tester ces deux hypothèses sur la syllabation des mots isolés en français.

2. METHODOLOGIE.

Avant même de s'interroger sur la richesse de l'information lexicale et du codage subséquent, une première difficulté doit être levée : sur quel matériel doit s'effectuer la modélisation : transcription orthographique ou phonémique, transcription phonétique, signal? A l'étape actuelle, une modélisation connexionniste prenant directement en entrée le signal digitalisé pose des problèmes pratiques et théoriques importants (constitution et taille des corpus, taille des fichiers de données, algorithmes de traitement du signal et de codage, etc.). Nous avons donc choisi, dans un premier temps de partir d'une transcription du signal.

Les transcriptions graphiques et phonétiques présentent pour notre projet les inconvénients de linéariser et de discrétiser le signal, et de filtrer les indices phonétiques dont on peut penser qu'ils constituent les bases de la syllabation. De plus, la transcription phonétique incorpore des décisions implicites de syllabation préjudiciables à l'analyse. Soit, par exemple, les transcriptions alternatives de lier :

- (4) [lje]
- (5) [lije]

La transcription graphique est un peu plus neutre sous ce rapport. Elle offre d'autre part l'avantage d'être constituée sous forme de corpus accessibles. Nous avons donc choisi comme point de départ d'appuyer nos modélisations sur un dictionnaire orthographique informatisé d'environ 90.000 formes. Deux lignes de recherche sont poursuivies en parallèle. La première, de type algorithmique, vise à expliciter des règles et à les organiser sous la forme d'un programme en C, la seconde vise à faire apprendre à un réseau de neurones formels des régularités par présentation des entrées et des sorties correspondantes.

Deux niveaux d'analyse peuvent être distingués.

a. Formalisation, sur la graphie, de règles de coupe phonique. Le sous-ensemble des règles de coupe phonique inclus dans les prescriptions de coupe graphique constitue le point de départ. Ce module est progressivement enrichi par l'adjonction de règles de coupe phonique.

b. Affaiblissement de la notion de règle et introduction des analyses multiples pour un même mot. Traitement de l'ambisyllabité. Introduction des régularités et des sous régularités.

La modélisation appuyée sur la graphie n'est, on l'a dit, qu'un choix provisoire. En effet, dès qu'on se propose de traiter des phénomènes phonologiques complexes mettant en jeu la continuité articulaire et/ou acoustique, les transcriptions discontinues de type orthographique voire phonétique/phonémique sont notablement insuffisantes. Dans une phase ultérieure, il sera donc nécessaire de prendre en compte directement le signal dans sa phase productive et réceptive. Ceci impliquera, en

retour des modifications considérables des architectures connexionnistes utilisées. L'architecture simple de réseaux à couches 'feed forward' de type PDP devra ainsi être abandonnée au profit d'architectures plus complexes incorporant en particulier des hypothèses cognitives sur l'intégration et la production du signal, à la manière de ce qui a été proposé pour la vision (cf. Bienenstock 1987). A ce niveau, la supériorité de modèles aptes à traiter le continu comme les neurones formels s'impose. Il en est ainsi, par exemple, du traitement de la syllabation dans des contextes de di-rèse/synérèse précédemment évoqués (cf. (4) et (5)).

3. PRINCIPES DE SYLLABATION

Comme nous l'avons dit, la modélisation d'un système expert de coupe syllabique et la comparaison avec les performances d'un réseau de neurones formels, travaillant tout deux sur une transcription graphique ne constitue que la première étape de notre projet de recherches. Ce sont les résultats de cette première étape que nous présentons à présent.

Au premier niveau, la syllabation est donc assurée par un système expert qui incorpore les principes suivants.

- a. Les digraphes sont insécables (exemple : ph, th, rh, etc.). Les groupes graphiques occlusive-liquide et fr/vr/fl/vl sont insécables. Les géminées graphiques sont donc considérées comme sécables mais une règle tardive d'épellation assure leur unicité et leur syllabation à l'attaque.
- b. Au niveau graphique, toute coupe syllabique laisse apparaître un noyau de syllabe.
- c. Les groupes CV ne sont jamais hétérosyllabiques. Ceci implique ce qui a été nommé "priorité à l'attaque". La syllabation d'une ou plusieurs consonnes à la coda est une conséquence de l'impossibilité de les syllaber à l'attaque.
- d. Les groupes C1C2 sont hétérosyllabiques ssi. ils ne forment pas un groupe insécable (priorité à l'attaque minimale).
- e. Les groupes V1C2 sont hétérosyllabiques sauf si C2 ne peut être en position d'attaque (cf. b).
- f. Les groupes V1V2 sont hétérosyllabiques si V1 et/ou V2 sont graphiquement accentués, à l'exception des quelques mots contenant certains groupes comme

eô, eû, oê, du groupe productif ai, et des finales de mot V1e, V1es.

g. Sauf exceptions traitées au deuxième niveau, notamment les très productives synérèses sur les groupes graphiques iV2, ainsi que les groupes eau qui ne coupent pas, les groupes V1V2 graphiquement inaccentués sont homosyllabiques ssi. V2 est un i ou un u graphiques, autrement ils sont hétérosyllabiques.

Au second niveau, les coupes prédites par le système expert sont testées manuellement et, en particulier, les ambisyllabicités et analyses multiples sont notées. On obtient ainsi la base minimale des entrées sorties fournies au réseau.

4. IMPLEMENTATIONS DES RESEAUX

La première implémentation est réalisée à l'aide d'un réseau à couches avec apprentissage par rétropropagation de type PDP (l'implémentation a été mise en oeuvre à l'ENST par C. Huynh dans le cadre de son mémoire de fin d'études). Sur la couche d'entrée on code des caractères. La fenêtre contextuelle étant une fenêtre glissante de 8 caractères avec test de coupe en position 3/4, et le nombre de caractères différents à prendre en compte étant de 43 (42 plus un symbole de fin de chaîne), la couche d'entrée comporte 344 unités. La couche de sortie sur laquelle est codée la possibilité d'une coupe en position 3/4 comprend une seule cellule dont le niveau d'activité est dans l'espace 0..1. Les niveaux de sorties supérieurs ou égaux à 0.5 sont considérés comme des réponses positives. L'optimum de résultat a été atteint avec une couche de 6 cellules cachées. La connectivité est strictement 'feed forward' sans inhibition bi-latérale.

L'apprentissage est réalisé sur un corpus de 1000 mots tirés au hasard, et les tests sont réalisés sur le dictionnaire complet de 90000 mots. Le corpus d'apprentissage est présenté 30 fois, et au total la convergence du réseau est atteinte après 45 mn de calcul sur une station SUN 3. Avec un corpus d'apprentissage au hasard, le pourcentage d'erreurs est de l'ordre de 2%. Une analyse du clustering sur la couche cachée fait apparaître que ses 6 cellules se divisent en 3 cellules excitatrices et 3 inhibitrices. Une des cellules excitatrices se comporte comme un détec-

teur typique de voyelles. On notera au passage que y graphique est, selon les cas, analysé comme voyelle et détecté par cette cellule ou bien analysé comme consonne. Le groupe des liquides est également détecté par une cellule spécialisée, mais il est de plus détecté, à un niveau faible d'activité, par la cellule spécialisée dans la détection des obstruantes. Enfin, il apparaît que les nasales et le h graphique sont détectés par une combinaison particulière des activités excitatrices et inhibitrices des cellules prenant en charge les voyelles, les liquides et les vrais consonnes. Une des cellules paraît spécialisée dans le traitement des groupes de voyelles graphiques, spécialement lorsqu'une de ces voyelles est accentuée.

L'analyse des erreurs et le caractère systématique et régulier d'un certain nombre d'entre elles nous ont conduit à modifier le protocole d'apprentissage. Le corpus d'apprentissage de 1000 mots a ainsi été divisé en 2 parties, les 500 premiers étant tirés au hasard, les 500 autres constituant une représentation statistique approchée des erreurs réalisées par le réseau dans un apprentissage strictement au hasard. On observe alors un comportement plus systématique et plus discret des cellules de la couche cachée. Le pourcentage d'erreurs est inférieur à 1%, soit de l'ordre de 600 mots, dont 1/4 environ est constitué de mots étrangers et 1/6 est constitué de mots à combinaison de caractères particulièrement rares (par exemple rhy, cz). Le reste des erreurs est constitué par quelques erreurs très systématiques portant sur des groupes lexicalement productifs de type préfixe dés + racine à h initial.

Bien que l'architecture de ce réseau n'incorpore aucune hypothèse cognitive ou phonologique sur la syllabation, on remarque qu'il couvre remarquablement le corpus des données avec un apprentissage faible. Ceci laisse supposer d'une part qu'une modification plus substantielle du protocole d'apprentissage incorporant des hypothèses sur la morphologie et la structure du lexique, d'autre part qu'une modification de l'architecture interne du réseau implémentant des hypothèses phonologiques devraient accroître encore les performances.

Ces performances pourront être comparées à celles d'un réseau traitant directement le signal en production et/ou en réception et assurant son apprentissage non seulement par une présentation simultanée des entrées et des sorties mais également par une réorganisation interne de type phylogénétique. L'architecture interne d'un tel type de réseau devra être riche et cognitivement pertinente. C'est dans cette orientation que se poursuit le travail présenté ici.

OUVRAGES CITES

- Anderson, J. et Ewen, C., 1987 : *Principles of dependency phonology*, Cambridge University Press, Cambridge.
- Bienenstock, E., 1987 : Connexionist approaches to vision, in Imbert, ed., *Models of vision perception : from natural to artificial*, Oxford University Press, Oxford.
- Bienenstock, E. et Doursat, R., 1990 : Epigenetic development of spatio-temporal patterns in the brain, in Schuster, ed., *Nonlinear dynamics and neural networks*, VCH publisher.
- Elman, J. L., 1990 : Finding structure in time, *Cognitive Science* 14.
- Encrevé, P., 1988 : *La Liaison avec et sans enchaînement. Phonologie tridimensionnelle et usages du français*, Le Seuil, Paris.
- Goldsmith, J., 1990 : *Harmonic phonology*, ms., University of Chicago.
- Kaye, J. D., Lowenstamm, J. et Vergnaud, J.-R., 1988 : *Constituent structure and government in phonology*, ms.
- Selkirk, E. O., 1982 : The syllable, in van der Hulst and Smith, eds., *The Structure of phonological representations*, Foris, Dordrecht.
- Selkirk, E. O., 1984 : On the major class features and the syllable theory, in Aronoff and Oehrle, eds., *Language sound structure*, M.I.T. Press, Cambridge.

A PROPOS DE "h" FINAL EN COREEN

J.-W. Lee

U.F.R. Linguistique, Université Paris 7, France
& Pusan.U.F.S. , Corée .

ABSTRACT

In this paper I would like to argue for the non-existence of an underlying /h/, in final position in verbal roots which are traditionally represented, for example, as "noh-"(to put) or "tah-" (to reach). Instead, I will propose a different syllable structure for these root forms, justified by an analysis based on the "transsyllabic government" proposed by Kaye, Lowenstamm and Vergnaud([2],1985).

1. INTRODUCTION

Est-ce qu'il existe vraiment un "h" en finale de morphème en coréen? Parmi les huit catégories grammaticales du coréen, on ne trouve de morphèmes considérés comme se terminant par "h", soit dans la graphie (ㅎ), soit dans des explications phonologiques, que dans les catégories "verbe" et "adjectif predicatif". Il faut remarquer que les études phonologiques sur le "h" final en coréen se sont fondées en fait sur la graphie. Dans les études [1,3,4,5.], on fait une opposition entre, par exemple, /noh-/ 'poser' et /po-/ 'voir' en postulant un /h/ sous-jacent en finale du morphème 'poser'. Pourquoi postule-t-on un /h/ abstrait qui n'apparaît jamais seul phonétiquement sinon sous forme de l'aspiration d'une géminée? Il y a une différence entre ces deux racines car on obtient [nottha] et [poda] respectivement, si on ajoute le suffixe infinitif /-ta/ au radical qui ont la même terminaison phonétique [o-]. Je vais montrer que cette différence ne vient pas de la présence d'un /h/ abstrait mais du résultat d'un processus phonologique différent dû à la structure syllabique différente des deux racines.

2. STRUCTURE SYLLABIQUE

Je propose une structure bisyllabique pour /no-/ et une monosyllabique pour /po-/, qui sont deux morphèmes phonétiquement monosyllabiques, et cette différence doit être marquée lexicalement:

(1)a. /no-/ racine de 'poser' aura une forme

	A	N	A	N
	X	X	X	X

b. /po-/ racine de 'voir' aura une forme

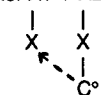
	A	N.
	X	X

Traditionnellement, dans les formes dérivées 'radical/ + /-ta/', on obtenait [th] par une métathèse de 'h+t', mais dans mon analyse, [th] résulte d'une insertion de l'élément ϕ^o - qui a comme trait [+continu], en fusion dans le deuxième segment d'une suite de '-C°C°-' qui est le résultat d'une propagation dans un domaine de gouvernement transsyllabique.

(2)a. Gouvernement transsyllabique entre deux positions d'attaque[2]:
-directionalite de droite à gauche
-tête à droite: AN + AN

b. Insertion de ϕ^o en fusion: si A1 + A2,

	X	X
		C°

insérer ϕ^o au segment qui est A2: '-C1C1' -> '-C1Ch1-'.


Ceci dit, cette insertion est due à la nécessité d'avoir le deuxième segment d'une suite '-C°C°-', aspiré ou tendu, ayant donc une représentation interne plus complexe selon les règles de distribution phonétique du coréen. A titre de référence, je rappelle ci-dessous, la représentation interne des segments consonantiques obstruents qui sont en opposition par leur complexité interne en éléments et par leur charme.

(3) les obstruents:(il faut y ajouter les sonantes pour compléter toute la representation interne des consonnes)

a. neutres: X X X X X X

		R°	R°	R°	ϕ^o
	U°	ϕ^o	ϕ^o	ϕ^o	ϕ^o
					o°
	p°s°	t°	c°k°h°		

b. aspirées X X X X

		R°	R°	v°
	U°	ϕ^o	ϕ^o	ϕ^o
	ϕ^o	ϕ^o	l°	ϕ^o
	ph	th	ch	kh

c. tendues X X X X X

		R°	R°	v°
	U°	ϕ^o	ϕ^o	ϕ^o
	H°	H°	H°	H°
	p°	s°	r°	c° k°

Nous allons maintenant passer à l'analyse phonologique en examinant des formes infinitives obtenues, en ajoutant le suffixe /-ta/ au radical verbal.

(4)a. /no-/ + /-ta/ ->[nottha]
b./po-/ + /-ta/ ->[poda]

a./no-/ + /-ta/ vs. b./po-/ + /-ta/
ANAN + AN AN + AN

X	X	X	X
n	o	p	a

R°a

?

↑

ϕ^o

[nottha] [poda]
En a, il y a d'abord propagation de /v/ au point squelettique de l'attaque vide en créant un domaine de gouvernement qui va ensuite déclencher l'insertion de ϕ^o . En b, il n'y a pas de processus phonologique de ce type (autre que le voisement de la consonne neutre expliqué par ailleurs).

3. DEMONSTRATIONS ET ANALYSES
On peut assurer que la deuxième consonne doit être plus complexe dans une suite de deux obstruents neutres '-C°1+C°2-', si on se souvient qu'en coréen, il y a un phénomène phonologique dit de 'tensification' qui transforme C°2 en tendue.

3.1. Aspirée vs. tendue
Comparons l'analyse des deux exemples suivants:

(5) a./no-/ + /-ta/ -> [nottha]
b./cap-/ + /-ta/ -> [capt'a].
En b, à la différence de a, à la place d'insertion de ϕ^o , il y a insertion de H qui est due, également à la nécessité du gouvernement transsyllabique:

a /no-/ + /-ta/ vs. b. /cap-/ + /-ta/
ANAN + AN AN + AN

X	X	X	X
n	o	c	a

R°a

U°

?

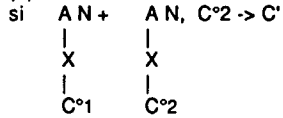
↑

ϕ^o

H°

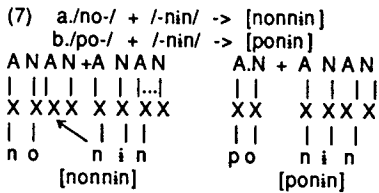
[nottha] [capt'a]
Mais cette fois-ci, /v/ du suffixe a déjà un segment concret à sa gauche, et il est obligé de prendre le degré le plus fort (les tendues, par leur charme négatif) pour pouvoir gouverner /p/ de la racine verbale 'prendre'[6,7].

(6) Insertion de l'élément H-:



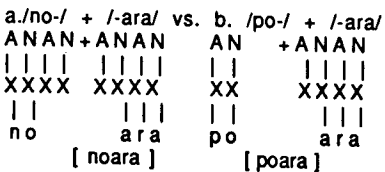
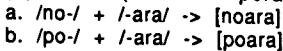
3.2. /no-/, 'bisyllabe' ?

Maintenant, indépendamment de [tth], nous allons voir s'il y a une bonne raison pour postuler une structure bisyllabique pour la racine /no-/ qui était traditionnellement représentée avec un /h/ sous-jacent en fin de morphème. En effet il y en a une, le phénomène de gémination nasale: quand on associe le suffixe de participe présent /-nin/ aux racines /no-/ et /po-/, le résultat phonétique nous montre bien que la première doit être considérée comme ayant deux syllabes sous-jacentes. Comparons les deux exemples suivants:



En a, il y a propagation du segment /n/ à la position vide ce qui provoque une gémination de /n/, mais en b, il ne se passe rien. Si on postule un /h/ sous-jacent dans le radical de /no-/, on est obligé d'introduire une règle d'élision dans ce contexte. Comparons encore pour constater que nous n'aurons plus besoin de cette règle même si on associe un suffixe qui commence par une voyelle:

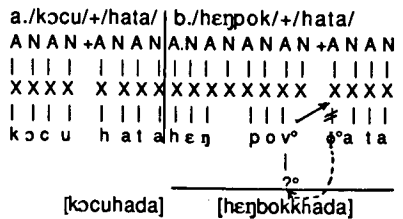
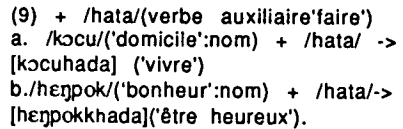
(8) +/-ara/(suffixe impératif).



Dans ce cas, nous avons une différence phonologique des deux formes abstraites mais la représentation phonétique est la même.

3.3. Cas des suffixes commençant par /h/.

3.3.1. Pour un "suffixe" commençant par /h/ (seule position d'occurrence du segment) tel que /-ha-ta/('faire'), si le dernier segment du morphème précédent est une voyelle, /h/ se maintient et se comporte donc, comme une consonne ordinaire, mais si le dernier segment est une consonne, il y a gémination de la consonne et fusion de 'h', provoquant l'apparition d'une géminee aspirée.



En a, /h/, qui a une existence phonologique, se maintient phonétiquement, mais en b, /h/ se trouve dans un domaine de gouvernement transsyllabique et comme il ne peut gouverner un segment plus complexe, il cède la position de gouverneur à /k/ ce qui produit une géminee aspirée.

3.3.2. Dans le cas du suffixe causatif ou passif, la distribution complémentaire observée dans les études classiques montre que l'on n'a besoin que d'une seule forme sous-jacente à l'attaque vide /-i/, et que les formes allomorphiques [-hi, -ri, -ki, -i] relevées dépendent du contexte et mon explication par le gouvernement transsyllabique les justifie bien.

4. CONCLUSION

En coréen, les consonnes ne se manifestent vraiment qu'en position initiale, ou plus exactement, en position d'attaque (on y trouve les trois séries d'obstruents, les sonantes et 'h'). Un phénomène phonologique capital du coréen est de constamment réorganiser la structure syllabique pour avoir des séquences réinterprétées en suites de syllabes ouvertes: A&N + A&N + A&N, etc., ce qui m'ammène à poser l'existence de syllabes sous-jacentes vides. Mon interprétation de 'h' va dans ce sens et un /h/ non-initial n'existe pas: l'aspiration transsyllabique est le résultat d'un type de processus phonologiques que la théorie de KLV[2] appelle: "gouvernement transsyllabique".

5. REFERENCES

[1] AHN, S.C.(1986), On the nature of h in Korean, *Studies in the Linguistic Sciences*, 16 -2, 1-13.
 [2] KAYE, J.D., LOWENSTAMM, J. & VERGNAUD, J.-R.(1985), The internal structure of phonological elements: a theory of charm and government, *Phonology Yearbook* 2, 305-328 .
 [3] KIM, C.W.(1970), A theory of Aspiration, *Phonetica* 21, 107-116.
 [4] KIM, S.H. (1990), *Phonologie des consonnes en coréen*, Thèse de doctorat, Ecole des Hautes Etudes en Sciences Sociales, Paris.
 [5] KIM-RENAUD, Y.K.(1975), *Korean Consonantal Phonology*, Ph.D. , University of Hawaii.
 [6] LEE, J.W.(1987), *Quelques problèmes phonologiques du coréen*, mémoire de DEA, Université Paris 7.
 [7] LEE, J.W.(1990), La structure syllabique et les segments, *Wetenonchong (Annales de l'université des langues étrangères de Pusan)* 8, PusanUFS, 171-209.
 [8] MARTIN, S.E.(1951), Korean phonemics, *Language* 27, 519-533.

UNDERSPECIFICATION AND PHONOLOGICAL ASSIGNMENT OF PHONETIC STRINGS: THE CASE OF CLASSICAL MANDAIC [qen:a:] 'NEST'

Joseph L. Malone

Barnard College and Columbia University

ABSTRACT

The Classical Mandaic (CM) coinage of the verb [qəna:] 'to build a nest' on the basis of the phonologically isolated noun [qen:a:] 'nest' poses a puzzle for linear phonology by implying that the underlying representation of the noun was taken to be marked /qen?aa/ rather than unmarked /qennaa/. However when the situation is re-analyzed in terms of nonlinear underspecificational phonology the puzzle vanishes, the nonlinear counterpart of /qen?aa/ turning out to be unmarked after all.

In [7] Sanford Schane proposed that a phonetic form which is indeterminate with respect to its phonological structure be automatically provided with whatever phonological structure might be determined by universal theory to be least marked for the phonetic string in question. In [4] I adduced a prima facie counterexample from CM, which I will briefly review here.

The CM noun [qen:a:] 'nest' had become lexically isolated and hence phonologically opaque. Though its original phonology had been */qennaa/, synchronically it might just as legitimately derive from either /qen?aa/

or /qe?naa/ by assimilation of /ʔ/ to /n/. In accordance with Schane's hypothesis, [qen:a:] should certainly re-affiliate with its original phonological representation as /qennaa/, /nn/ being patently a less marked origin of [n:] universally than either /nʔ/ or /ʔn/. But in fact the Mandeans' subsequent coinage of a verb 'to build a nest' on the basis of [qen:a:] clearly revealed that /nʔ/ was the underlying solution; see [4] for justification.

QED--or so I thought in 1970. However, the advent of autosegmental, syllabic, and underspecificational phonology (cf. specifically for this study [1,2,3,5,6]) has led to a complete reevaluation, as I shall now show.

Taking off from the observation that /ʔ/ was merely an SPE-vintage abstract segment (though historically the reflex of a true phonetic laryngeal (*[ʔ]) or pharyngeal (*[ʕ]), and should rather be replaced synchronically by a featurally unspecified melodic unit (/ʔ/), let us start with the derivation in (1).

First, melodies associate to whatever skeletal positions are syllabically marked. Archangeli's approach [1] allows marking of syllable heads, indicated in (1a) by a vertical line over an X; and also of positions in the

domain of a syllable head, indicated by a slant line over an X. Hence the melody e associates to the simple nucleus X while a associates to the complex nucleus XX. These associations are given in the step (1a) to (1b), making for the short e of the stem and the long a of the suffix.

Next, remaining melodic segments are associated with syllabically unspecified positions in the step (1b) to (1c).

Then remaining syllabic specifications are provided in moving from (1c) to (1d). This is guided in part by universal regularities, and in part by language-specific patterns. Thus for CM, assignment of Onset (O) to the X associated with ʔ is not hampered by the featural vacuity of the latter, since Mandaic-impossible syllables would result from any other assignment. The X in question cannot be associated leftward, since Coda (C) adjuncts are admitted only under quite restricted circumstances. Neither can the X associate rightward, since three-mora Nuclei (N) are strictly disallowed.

Finally, an anchoring convention dictates that an unspecified melody reassociate from its skeletal position to whatever adjacent melody the syllabic assignments will tolerate: to the lefthand melody in this case, only n but not also (righthand) a comprising a possible Onset.

The derivation in (2), corresponding to the historical /qennaa/ analysis, falls out even more simply, since there are no unspecified melodies. Beyond that, the only notable difference from (1) is in step (b) to (c), where the melody n spreads to two tandem X's.

However, when we attempt to apply this treatment to a

form with a second-radical ʔ, in (3), an apparent difficulty emerges, since the phonotactics of the language will allow the ʔ to assume either Coda value, in (3d), or Nucleic value, in (3d'), with the consequence of predicting alongside correct [qen:a:], in (3e), also incorrect *[qi:na:] in (3e')-- [i:] instead of [e:] following by a rule of raising.

But this is not a difficulty per se. Though not considered in [4], this is a potentially correct result, one brought out virtually automatically under the joint autosegmental-underspecificational assumptions adopted here. Though lexical "freezing" forestalls pervasive free variation, the overall reflexes of nouns of this stem shape with original 2/ *ʔ or *ʕ are pretty much split between resolutions like [qen:a:], and those like the unattested alternant *[qi:na:].

We are now ready to consider how the paradigmatically isolated noun [qen:a:] 'nest' might "choose" among the likes of (1,2,3) upon the occasion of the Mandeans' fielding a new paradigm to the tune of a denominal verb 'to build a nest'. Which of these, (1a) or (2a) or (3a), might provide the best suited underlying representation, all else being equal?

It seems to me that (1a) does, for three reasons: (1) Both (1a) and (3a) should be favored over (2a) because each of the former contain three-radical roots, which all hands down represent the unmarked state of affairs in CM and all other Semitic languages. Thus the root in (1a) is /qəʔn/ and that in (3a) is /qəʔn/. So-called geminate roots, on the other hand, are normally analyzed as

biradical autosegmentally (see e.g. [5,6]). Thus the root in (2a) would be the two-radical /qn.

Two factors give the edge to (1a) over (3a):

{II} First of all, (3a), as we have seen, allows vacillation in phonetic stem-shape, between a long-consonant resolution like [qen:a:] and a long-vowel resolution like †[qi:na:]. (1a), on the other hand, provides unambiguous stem-stability, in terms of just long-consonantal [qen:a:].

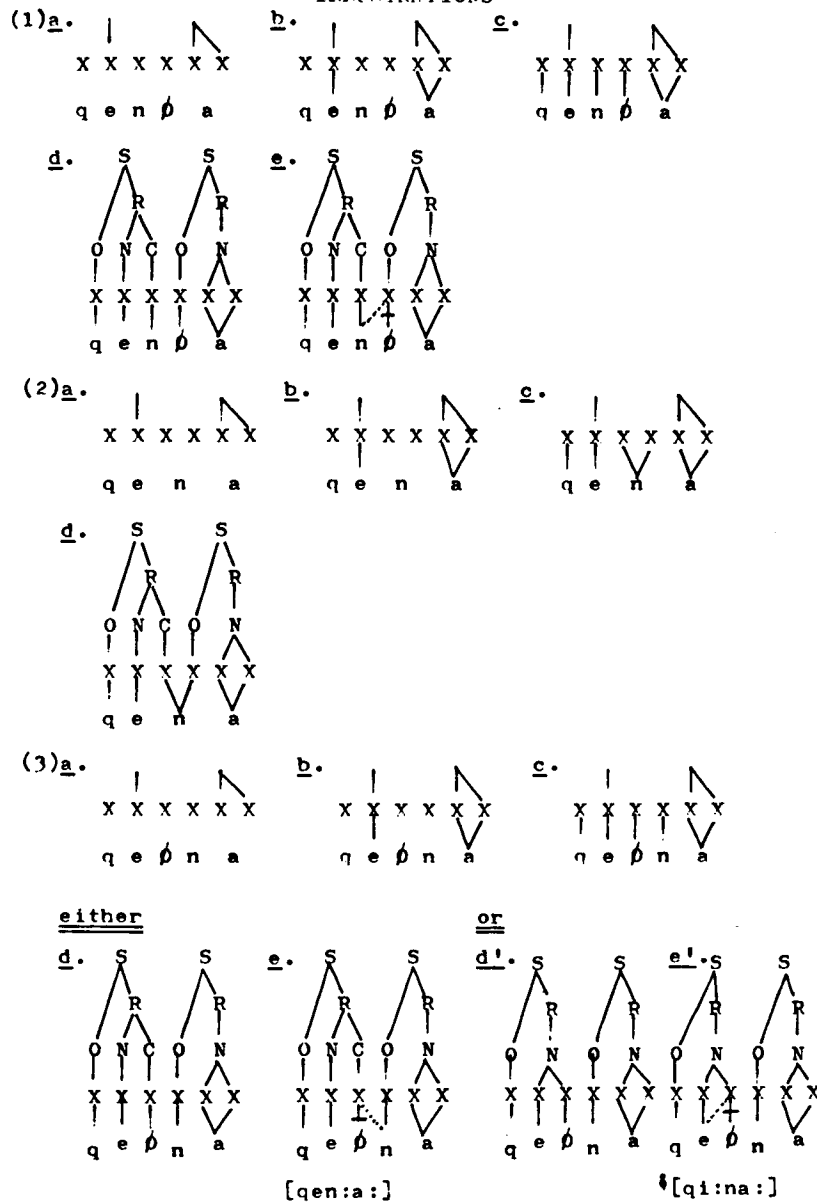
{III} Finally, the verb actually coined on the basis of (1a) turns out to be appreciably closer to the unmarked (strong verb) canon [CəCaC]. Thus [qəna:], the actual verb 'to build a nest', is phonetically closer to a strong verb like [ləYaʔ] 'to take' than would either †[qan], corresponding to (2), or †[qa:n], corresponding to (3).

Thus there need be nothing at all maverick about the restructuring of [qen:a:] as phonological /qenβaa/. On the contrary, if the analysis just proposed is approximately correct, it instantiates one of the most mundane of all analogical change types: assimilation to the least marked available model--much like Schane [7] proposed way back in 1968 after all.

REFERENCES

- [1] ARCHANGELI, D. (1985) "Yokuts harmony: evidence for coplanar representation in nonlinear phonology," *Linguistic Inquiry* 16, 335-372.
 [2] LEVIN, J. (1985) "A metrical theory of syllabicity," MIT Ph.D. dissertation.
 [3] LOWENSTAMM, J. and J. KAYE. (1986) "Compensatory lengthening in Tiberian Hebrew," L. Wetzel and E. Sezer, eds., *Studies in compensatory lengthening*, Dordrecht: Foris.
 [4] MALONE, J. L. (1970) "In defense of non-uniqueness of phonological representations," *Language* 46, 328-335.
 [5] MCCARTHY, J. (1981) "A prosodic theory of nonconcatenative morphology," *Linguistic Inquiry* 12, 373-418.
 [6] MCCARTHY, J. (1986) "OCP effects: gemination and antigemination," *Linguistic Inquiry* 17, 207-264.
 [7] SCHANE, S. (1968) "On the non-uniqueness of phonological representations," *Language* 44, 709-716.

ILLUSTRATIONS



LOW LEVEL PHONETIC IMPLEMENTATION RULES: EVIDENCE FROM SINDHI

Paroo Nihalani

National University of Singapore

Abstract. The traditional notions of segmental phonetic representation and rule systems formulated in terms of discrete operations have paid little attention to the processes of "phonetic implementation" as opposed to "physiological implementation". This paper argues that some details of speech, such as timing and coordination of articulatory gestures, have language-specific conditioning, and therefore should fall within the scope of phonology. Evidence will be provided from implosives in Sindhi and other languages in support of the premise, and the status of low level phonetic implementational phenomena in phonological theory will be discussed.

1. INTRODUCTION

Chomsky & Halle (1968) and Goldsmith (1980) characterize sound contrasts on the phonological level in terms of binary feature values. They consider each feature to be both a component at the phonological level and a single physical scale. Recently, Ladefoged (1981) and Lindau and Ladefoged (1983) have shown that relating a feature to a single physical scale often constitutes an oversimplified view of feature correlates.

Sounds of one language may differ from those of another because of the phonetic value of the segments along the same continuum. To take an example, the linguistic specification that distinguishes between [p] and [b] in English is that they are [-voice] and [+voice] respectively. The articulatory instruction that accompanies the feature

[+voice] is "vibrate the vocal folds". In order to implement this instruction, a number of articulatory instructions have to be performed, such as keeping the vocal folds sufficiently lax, reducing the distance between the vocal folds, keeping the airflow through the glottis powerful enough to cause vibration, and maintaining the difference between the subglottal and supraglottal air pressure by lowering the larynx, allowing air to escape through a small velic opening, and/or expanding the walls of the pharynx. "Vibrate the vocal folds", however, is the primary instruction that is associated with the linguistic feature [+voice], and the rest of the articulatory gestures are ways of implementing this instruction. Speakers of different language backgrounds choose different combinations of parameters for the implementation of voicing in stops. The phonetic implementation of these differences is as much important as those in the sound patterns. In order to illustrate this point, I will discuss some phonetic differences between implosives in Sindhi and few other languages.

Implosives have been traditionally characterized as glottalic ingressive sounds produced by lowering the vibrating glottis (Catford, 1939; Pike, 1943). Lindau (1984, p. 152) notes that Hausa implosives are produced with aperiodic, inefficiently closing vocal cord vibrations and that there is considerable speaker to speaker variation between implosives in languages, and that languages may differ in the way that they maintain distinction between implosives and the corresponding plosives. Ladefoged (1964, p.6) noted that his Igbo implosives only produced negative

pressures 8% of the time. Ladefoged (1971, pp. 25-26) therefore observes: "Although these sounds may be called implosives, in ordinary conversational utterances air seldom flows into the mouth when the stop closure is released."

In this connection, Painter (1978, p. 254) observes: "Despite Ladefoged's caveat (1964, p.6) that his Igbo implosives only produced negative pressures 8% of the time... my physiological data for Ga, Sindhi and Yoruba show negative pressures most of the time." More recently, Nihalani (1986) has shown that there exist natural languages like Sindhi (spoken in India and Pakistan) and Kalabari (spoken in Nigeria) in which implosives do involve an ingressive air flow in addition to the downward displacement of the vibrating glottis. The quantitative measurements of the air flow dynamics run counter to Ladefoged's assumption that there are no real implosives.

Ladefoged (p. c.) has commented that Nihalani's findings are based on his own speech (one single speaker), and that the aerodynamic data are collected from citation forms. Ladefoged has valid criticism in that we should always use large enough sample to base our generalizations. It is obviously crucial to any study of this sort to have as many speakers as practicable, in order to increase the possibility of making meaningful language-specific generalisations.

The purpose of this study was to expand the data on pressure-flow dynamics from much larger number of informants in order to explore aerodynamic characteristics of implosives in Sindhi and also to determine whether these articulatory strategies are consistent within a language or vary only according to speaker-specific idiosyncrasies.

2. TEST MATERIALS

Data on the intraoral pressure and oral air flow were collected from 3 speakers (1 male and 2 females, based in Los Angeles). A minimal pair representing the bilabial implosive sound positioned syllable-initially was selected. The language informants were requested

to utter words in a carrier phrase: "hi: 6.576"

3. INSTRUMENTATION

The language informant speaks into a specially constructed mouthpiece pressed against the face, which takes the oral air flow through calibrated resistance so that a pressure transducer provides a signal that is directly proportional to the rate of air flow. If one can find a language informant who is willing to tolerate a nasal catheter, then it is possible to record the pressure build up behind stop closures anywhere in the vocal tract. Alternatively, a simple way of obtaining supraglottal air pressure and air flow data on just bilabial sounds was used by inserting a small tube between the lips.

All these parameters were digitized along with the audio signal from a microphone at the rate of 11000 samples/sec. Figure 1 is an example of the aerodynamic data recorded in the Phonetics Lab, UCLA. The top channel records the audio-signal, the middle channel represents oral air flow and the bottom channel represents intraoral air pressure.

4. RESULTS

Figure 1 gives the aerodynamic record of the word [ʃaru] 'child'. The closure phase in the articulation of the implosive sound is characterized by a straight line Q-C (channel 2) indicating absence of air flow in either direction through the mouth. The large periodic fluctuations in the delimited segment R-S on the the pressure tracing (channel 3) reflect the vibrations of the vocal cords. A mid-line was drawn through these ripples by hand. The maximum pressure was measured on the mid-line. The measurements of the P_{supra} were made at the point of release of closure. Table 1 presents the Peak P_{supra} values of the syllable-initial implosives/explosives.

Table 1. Peak Measurements of Supraglottal air pressure in cm H₂O.

	b	6	Difference
HW	7.5	-2	5

In the production of the implosive [ɓ], the vocal folds are brought together before the larynx is lowered. Vocal folds remain fairly tightly together throughout the articulation so that air will not pass through the glottis in such large volume as to destroy the negative pressure necessary for an implosive. Lowering of the larynx obviously enlarges the supraglottal cavity behind the oral closure which results in generating negative pressure inside the mouth. Since the larynx lowers only after the vocal folds are constricted, the lips brought together and velopharyngeal port closed, the rarefaction process in the expanding supraglottal cavities is not affected so much so that the air is sucked in when the outer closure is released. These results are typical of other female speaker as well.

Another interesting feature was noted consistently in the speech of both speakers. Implosives are produced with a relatively short closure duration. Table 2 presents the duration of voicing in both 'implosives' and 'explosives'.

Table 2. Duration of voicing in ms.

	b	ɓ	Difference
HW	14	10	4
SS	12.5	9	3.5

Note that the voicing of implosives ranges between 70% to 72% of the corresponding explosives.

The third speaker, however, produced implosives with a voiceless beginning of the closure. The closure displays highly aperiodic vibration, whereas the voiced plosive [b] in the speech of the third speaker has periodic voicing vibrations during the closure phase. So the voicelessness or aperiodicity in the case of third speaker may serve to keep the implosives apart from the voiced plosive. However, the spectrograms made from the independent recording of the same speaker clearly indicate presence of vocal fold activity throughout the period of closure in the

articulation of implosives. I don't know how to resolve this anomaly.

5. DISCUSSION

The aerodynamic records show in the case of 2 out of 3 speakers that the movement of the larynx occurs while the vocal cords are vibrating. This downward movement of the vibrating glottis enlarges the supraglottal cavity behind the closure. These vibrations are maintained by a small amount of lung air which is not of sufficient volume to destroy the partial vacuum caused by the downward laryngeal movement and thus prevent the occurrence of suction-pressure. The negative pressure range between -2 cmH₂O to -5 cmH₂O was generated in the mouth. On separation of the articulators, the airflow was found to be ingressive. Thus the quantitative measurements, on the whole, confirm the results reported earlier by Nihalani (1986).

6. THEORETICAL ISSUES

The preceding discussion makes it clear that Sindhi implosives show negative pressure most of the time in contrast to the implosives observed by Ladefoged in which negative pressure was produced only 8% of the time. The first question that comes up is: Should the linguistic characterization of implosives be based on negative pressure/suction, with the greater degree of downward displacement of the larynx being a physiological consequence of the need to maintain the pressure difference for suction, OR should the linguistic characterization specify (as Ladefoged implies) the greater displacement of the larynx?

Suppose we took the position that the linguistic instruction that is associated with the production of implosives is "lower the larynx". Voiced explosives and the implosives would then be linguistically distinguished from each other in that the instruction to lower the larynx is implementational in the former (the larynx is lowered in order to keep the vocal folds vibrating), while it is phonological in the latter. This distinction in the phonological function of

the articulatory gesture of 'larynx lowering' is parallel to that of 'velum lowering'. In the production of nasal sounds, the instruction to lower the velum is phonological in that it is associated with the feature [+nasal], while in the production of voiced plosives in Sindhi the lowering of the velum is only a means of implementing the vibration of the vocal folds (Nihalani 1975).

The distinction between the implosives in Hausa, on the one hand, and Sindhi, on the other, in "not having" and "having" ingressive airflow would then be a difference in the implementation of the instruction to lower the larynx. In Hausa, the oral closure is released only when the supraglottal air pressure is neutralized with the ambient pressure, while in Sindhi the oral closure is released when the supraglottal air pressure is less than that of the atmospheric pressure. As a result, there is an ingressive airflow in Sindhi but not in Hausa.

An alternative would be to hold that the relevant phonological feature of implosives is [+suction], which is associated with the instruction "create an ingressive air flow". The lowering of the larynx would then be a procedure for the implementation of this instruction. That this instruction is not actually realized in languages like Hausa would then be analogous to the fact that the phonological instruction to vibrate the vocal folds fails to apply prepausally and postpausally during the closure period of voiced stops in languages like English.

An interesting theoretical issue that arises from the study of implosives in Sindhi is the status of implementation phenomena in phonological theory. There has been a growing body of literature in phonetics and phonology during recent years arguing that some details of speech, such as timing and coordination of articulatory gestures, have language-specific conditioning, and therefore they fall within the scope of phonology (Ladefoged, 1980, 1985; Liberman, 1983; Port, Al-Ani and Maeda, 1980; Port and Mitleb, 1983; Mohanan, 1986; Cohn, 1990; Huffman, 1990).

These processes of "phonetic implementation" as opposed to "physiological implementation" pose a challenge to the traditional notions of segmental phonetic representation and rule systems formulated in terms of discrete operations, and are therefore of profound interest.

Until recently, a widely accepted view, following Chomsky & Halle was that phonetic implementation was universal and this was discussed explicitly in terms of coarticulation. Phonetic implementation or the physical realization of the abstract patterns represented by the phonology was assumed to be mechanical. As a consequence, a phonological output was assumed to have a unique physical realization. It was also assumed that phonetic differences occurred cross-linguistically. Within this framework, the distinction between phonetics and phonology appeared clear-cut. Phonology involved language-specific rules, whereas phonetics was the universal mechanical realization of the phonology. Since the mapping was thought to be universal, little attention was paid to the phonetic implementation of phonological representation from a linguistic point of view. However, the more phoneticians looked for cross-language phonetic generalisations, the more exceptions they found to possible universal phonetic generalisations. Many phonetic processes that were assumed to be mechanical and to follow automatically from physiological factors, on clearer examination, turned out to demonstrate significant differences between languages. Differences of each language therefore will have to be described in terms of language-specific low level rules of "phonetic implementation", and these must form part of the phonological description of natural languages. Thus an understanding of the mapping processes from discrete, categorial and timeless phonological units to continuous articulatory and acoustic quantitative physical manifestations is a real central issue in the general understanding of phonology, and is the important goal of linguistic phonetics.

EVIDENCE FOR FINAL DEVOICING IN GERMAN? AN EXPERIMENTAL INVESTIGATION

H. G. Piroth, L. Schiefer, P. M. Janker and B. Johné

Institut für Phonetik und Sprachliche Kommunikation
der Universität München, Germany

ABSTRACT

This investigation deals with the question of whether morpheme- and word-final devoicing in German is a case of complete or partial neutralization. Durational parameters measured from systematically varied utterances of two South German speakers lead to the suggestion that concerning vowel, occlusion and release durations, final devoicing is incomplete in some morphosyntactic positions.

1. INTRODUCTION

Final devoicing ("Auslautverhärtung") is one of the standard cases for the neutralization of a phonological contrast [7]. During the last ten years investigations were undertaken to show that neutralization of voicing in German final obstruents is incomplete [3]. In subsequent experiments O'Dell & Port [4] and Port & O'Dell [6] reported that in words with underlying voiced stops the duration of the preceding vowel is significantly longer, that 'voicing into closure' is also longer on average, whereas occlusion and aspiration are shorter in this case. Since their results were gained in a reading task Fourakis & Iverson [2] claimed that the incompleteness of neutralization measured might be due to hypercorrection in reading. Therefore they performed an oral word conjugation test instead that gave no hint in favour of an incomplete final devoicing. Charles-Luce [1] draw attention to the question of whether neutralization might depend on the position and context of the affectable word-final obstruent in the sentence frame. Although his results were not systematic, in some cases of significant differences between underlying voiced and voiceless alveolars, position and context effects could be detected. Port & Crawford [5]

discuss the effect of speaking styles and task conditions to approach the question of whether incomplete neutralization is artificial (e.g. orthographically induced) or not. In their production experiment they presented three words affectable by neutralization and their counterparts under different conditions (the words disguised in sentences, the words directly contrasted in sentences, and the words in isolation randomly presented). Their results suggest a voiced/voiceless contrast in the neutralization position when the crucial word pairs were directly contrasted in single sentences. The contrast they found for the isolated words in our opinion seems to be due to the fact that the word list was so small that Ss could gain evidence of the experimental purpose.

Considering as important the point brought into the discussion by Charles-Luce [1] we looked for a test design that is (i) suitable to vary German stops supposed to be affected by final devoicing (FD) systematically over the relevant contexts and (ii) complex enough to hinder the Ss from recognizing the experiment's objective.

2. TEST MATERIAL AND DESIGN

To meet both requirements in the examination of the range of neutralization in German stops words were chosen that allow the influence of final devoicing to be tested in five different positions: The final position representing the standard case for final devoicing (FD), subdivided into (1) the utterance-final and (2) the word but not utterance-final position, the morpheme-final but not word-final position in compounds, subdivided into (3) morpheme-final position with voiced and (4) with voiceless continuation. The inter-

vocalic position (5) was added as control context which should not be affected by neutralization.

Therefore, words were selected which can be arranged in pairs fulfilling the FD condition in their rhymes and can easily be used to build compounds with voiced and voiceless continuation as well as word forms with the FD-affectable consonant in the intervocalic control position. Each place of articulation (labial, alveolar, velar) is represented by three word pairs with at least two different nuclei, one containing vowel+/1/ before the stop. All word forms are shown in Tab. 1.

Table 1: Word Material

Words are arranged according to place of articulation and position of the stop:

- (1,2) utterance- or word-final,
- (3) morpheme-final in voiced context,
- (4) morpheme-final in voiceless context,
- (5) intervocalic position

	(1,2)	(3)	(4)	(5)
labial				
Bub	Büblein	Bübchen	Buben	
Hup	Huplaut	Hupverbot	hupen	
Hieb	Hiebwaife	hiebfest	hieben	
Piep	Piepmatzen	piepsen	piepen	
Kalb	Kälblein	Kälbchen	kalben	
Alp	Alpweide	Alphorn	Alpen	
alveolar				
Rad	Radlager	Radfahrer	Räder	
Rat	ratlos	Ratschlag	Rates	
Ried	Riedweg	Riedkanal	Riedes	
miet	Mietwagen	Mietvertrag	Miete	
Wald	Waldlichtung	Waldvogel	Waldes	
alt	Altmetall	Altflöte	alte	
velar				
Betrug	Truglicht	Trugschluß	betrügen	
Spuk	Spukmärchen	Spukschloß	spuken	
Berg	Bergluft	Bergsteiger	Berge	
Werk	Werkmeister	Werksfahrer	Werke	
Balg	Bälglein	Balgtreter	balgen	
Kalk	Kalklager	Kalkfuhrer	kalken	

In preparation of the test stimuli these word forms were embedded in a sentence frame "Ich sage ... nochmal". For the utterance-final condition the word "nochmal" was omitted ("Ich sage...") resulting in 90 test sentences (18 words x 5 conditions).

To ensure that the subjects have no evidence of the experimental purpose also words with fricatives, nasals or liquids instead of the stop were used to construct derivatives of a similar shape and presented in the same frames. These sentences were read from cards containing the orthographic form of one sentence each by two South German native speakers (1f/1m) three times in randomized order.

Subjects were seated comfortably in a chair within a soundproofed room in front of a Neumann 11304-8 cardioid microphone. The sessions were recorded on audiotape (Telefunken M 15). The test words were analyzed for durational parameters by means of a Kay DSP Sonagraph 5500 (wide band 8kHz). The parameters are the duration of the vowel, the occlusion, the release and the word stem. By definition, vowel duration is measured from F2-onset after the preceding consonant or consonant cluster to F2-offset (including the liquid if present) before the occlusion. Occlusion starts from that point and ends at the beginning of the release consisting of the burst and the following aspiration (if present). If a fricative followed the stop, then the release ends at the point with a clearly visible change in the spectral structure of the frication. Otherwise it ends when no energy was visually detectable in the sonagraph (at an input sensitivity of 45dB). Word duration is counted from the beginning of the consonant or consonant cluster which precedes the vowel to the end of the release, thus covering the word stem only.

Additionally it was registered whether the stop was realized as voiced or voiceless, whether the consonant following the release was voiced or voiceless and whether it occurred within 40 ms or more.

3. RESULTS

Since the registration of voiced and voiceless bursts showed that only 51.9% of the phonologically voiced stops in the inter-

vocalic control position were phonetically voiced and since only one case of a phonetically voiced burst was found in the remaining material, only durational parameters were statistically analyzed. A 5x2-factorial ANOVA (5 positions and 2 phonation types) was calculated for the durations of the word, vowel, occlusion and release pooled over subjects, words and places of articulation.

For the analysis morpheme- and word- (but not utterance-) final cases were omitted if the pause between the stop and the following consonant was 40 ms or more. Main effects and interactions are shown in Tab. 2, as well as a posteriori pair comparison results (Scheffe) for significant main effects and simple effects for significant interactions ($\alpha = 0.01$). Position and phonation type are always of significant influence, the interaction between both only for the variables occlusion and release. For word duration the rank order of positions as shown by the Scheffe procedure reflects the fact that within a compound the target word stems are shorter than in the

Table 2: Analysis of variance results
PHON(1,2): Category of underlying stop (voiceless/voiced)
POS(1,5): Stop position (utterance-final, word-final, morpheme-final in voiced context, morpheme-final in voiceless context, intervocalic)

Main effects and interactions

	d.f.	F	p
Vowel			
POS	4,464	16.638	p<0.001
PHON	1,464	11.757	p=0.001
POSxPHON	4,464	1.478	p=0.208
Occlusion			
POS	4,464	78.292	p<0.001
PHON	1,464	55.766	p<0.001
POSxPHON	4,464	10.025	p<0.001
Release			
POS	4,464	74.286	p<0.001
PHON	1,464	36.017	p<0.001
POSxPHON	4,464	3.556	p=0.007
Word			
POS	4,464	77.268	p<0.001
PHON	1,464	27.783	p<0.001
POSxPHON	4,464	0.926	p=0.449

Simple effects within significant interactions

Occlusion

PHON within			
POS(1)	1,422	11.348	p=0.001
POS(2)	1,422	3.683	p=0.056
POS(3)	1,422	2.853	p=0.092
POS(4)	1,422	3.851	p=0.050
POS(5)	1,422	136.771	p<0.001

Release

PHON within			
POS(1)	1,422	13.473	p<0.001
POS(2)	1,422	6.943	p=0.009
POS(3)	1,422	1.493	p=0.222
POS(4)	1,422	0.085	p=0.770
POS(5)	1,422	38.990	p<0.001

Scheffe a posteriori pair comparisons (p<0.01) for significant main effects

Vowel 4 < 3 < 2 < 5 < 1
x ——— x
 x ——— x

Occlusion 4 < 5 < 3 < 2 < 1
x — x x — x
 x — x

Release 4 < 5 < 3 < 2 < 1
x — x x — x
 x — x

Word 4 < 3 < 5 < 2 < 1
x x ——— x x

intervocalic control position while they are longer in word- and utterance- final position. The fact that word duration in utterance- final position (1) is significantly longer than in any other position might at least partly be due to final lengthening which should not occur in the other positions. On the other hand, in position (4) being the only position in voiceless context it is significantly shorter than in all positions with voiced context which are statistically not different from one another.

Scheffe pair comparisons for vowel duration separate the morpheme- final position on the one hand from the utterance- final and the control position on the other hand. For occlusion and release pair comparisons show the same structure. Duration in word- and utterance- final positions are significantly larger than the

others. The morpheme- final positions differ from one another with the control position in between, which can be explained with respect to the significant interactions that were encountered concerning occlusion and release. In both cases the interaction is based on the phonation effect which occurs as expected for the control position and the fact that as expected as well there is no effect of phonation within the morpheme- final positions. Interestingly, there is a clear effect of phonation in the utterance- final position and additionally in the word- final position for release only (occlusion and release are longer in the voiceless case). The varying influence of phonation on the occlusion and release durations in different positions can be seen in Fig.1. Especially, in the control position (5) the durational differences between voiced and voiceless stops are evident.

4. DISCUSSION

Taking the results overall, it emerges that neutralization in final stops in German is not simply final devoicing. Even for the control word forms with intervocalic non- neutralized stops phonetic voicing plays no important role, since only half of the realizations have voiced releases, while the durational differences are distributed according to phonation types under several conditions. For morpheme- final positions in compounds no effect of phonation type could be found in terms of vowel, occlusion and release durations. On the other hand, there is a clear effect for release in word- final and for release and occlusion in the utterance- final position from which the standard examples for final devoicing in German stops are taken, so that in these cases definitely no neutralization occurs.

As these results are taken from durational data in future work we will measure the distribution of spectral parameters over phonation types. Since the data are taken from only two South German speakers we plan to include Mid and North German speakers as well. Furthermore we intend to expand the material to contain fricative pairs as well.

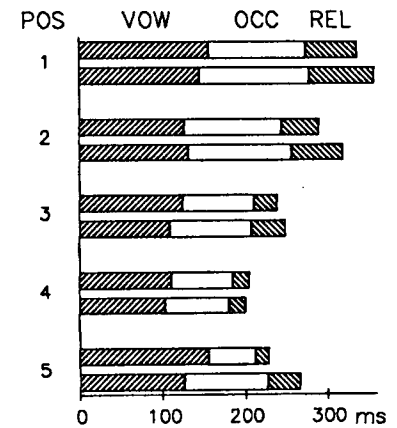


Fig. 1: Durations of vowel, occlusion and release for positions 1 to 5 (lower bar: voiceless; upper bar: voiced)

5. REFERENCES

[1] CHARLES-LUCE, J.(1985), "Word-final devoicing in German: effects of phonetic and sentential contexts", *J.Phon.*, 13, 309-324.
[2] FOURAKIS, M. & IVERSON, G. (1984), "On the 'incomplete neutralization' of German final obstruents", *Phonetica*, 41, 140-149.
[3] MITLEB, F.(1981), "Temporal correlates of 'voicing' and its neutralization in German", *Research in Phonetics: Report No. 2*, 173-192, Bloomington.
[4] O'DELL, M. & PORT, R.(1983), "Discrimination of word final voicing in German", *J.Acoust.Soc.Am.*, 73, Suppl.1, S31.
[5] PORT, R. & CRAWFORD, P.(1989), "Incomplete neutralization and pragmatics in German", *J.Phon.*, 17, 257-282.
[6] PORT, R. & O'DELL, M.(1985), "Neutralization of syllable-final voicing in German", *J.Phon.*, 13, 455- 471.
[7] TRUBE TZKOY, N.(1939), "Grundzüge der Phonologie", *TCLP*, 7, Prague.

SOUND DISTINCTION: UNIVERSAL INVENTORIES
OF PHONIC SUBSTANCE OR LANGUAGE-SPECIFIC SYSTEMS?

Vulf Y. Plotkin

Siberian Branch, Academy of Sciences
Novosibirsk, Russia

ABSTRACT

Since each language system is a unique segmentation of universal substance shaped into its elements, ultimate phonological units are not items in universal inventories of so-called 'distinctive features' to be incorporated into language systems. Each phonological system segments the universal human arsenal of sound-distinctive capacity in its own way, forming a specific set of neuromuscular impulses, of which each activates a fully automatized aggregate of articulations.

Phonological theory experiences today tremendous difficulties on account of its growing isolation from the needs of modern phonetic technology, which, finding little response to its requirements, has to rely on its own groping solutions of phonological problems. This gap between theory and practice is the inevitable result of the failure to work out an adequate answer to the fundamental question of the ultimate phonological unit.

It is widely believed that phonology as a linguistic science started by assigning that status to the pho-

neme - hence its other name 'phonemics'. However, as early as in 1936 Josef Vachek showed that the phoneme was not the smallest indivisible phonological unit, because it could contain smaller non-successive simultaneous units, e.g. sonority, palatality etc. [13]. The idea was developed by Bohumil Trnka at the 3rd International Congress of Phonetic Sciences in Ghent [12]. Then Roman Jakobson devoted four decades of pioneering work to the search for the phonological quantum - the ultimate language unit named 'distinctive feature' [7; 8; 9]. The best-known result of the quest is the universal inventory of a dozen items, of which phonological systems are built for all languages. The inventory was later revised theoretically and enlarged threefold by Noam Chomsky [4]. While fully recognizing the great scientific and practical value of R. Jakobson's achievements, we have to admit nevertheless that the entities he discovered and catalogued are not what he thought they were, i.e. the ultimate phonological units. It stands to reason that no items from a universal set can be directly employed as units in a language system

[5]. N. Chomsky was therefore quite consistent in stressing the language-independent nature of his inventory of 'features' [3]. What R. Jakobson and N. Chomsky really inventoried is indeed universal, it is the common human arsenal of sound-distinctive capacity. Naturally, all phonological systems are based on it as their substance foundation. But no part of underlying substance can be directly integrated into any system, and language systems are not exceptional in this respect. The elements of the universal anthropophonic distinctive potential listed in the above-mentioned inventories are certainly not ready-made units to be selected by and included into a concrete language system. A language unit is not a mere piece of substance, but substance shaped as an element to fit into the unique structure of the given language system. Consequently, elements of different systems cannot be identical with elements in other systems, however close they might seem in substance. This has long been accepted for phonemes, but not for 'distinctive features', which, according to R. Jakobson, coincide with the same 'feature' in other languages [8]. Regrettably no theoretical explanation was offered for this deviation from the general principle that precludes the compilation of universal inventories for phonemes, morphemes, words from all languages. The phonological system of any language is a specific way of segmenting the universal potential of phonic distinction and molding the

segments obtained into language units - ultimate phonological quanta. The segments are not produced by selecting some 'features' as relevant and discarding the rest as redundant; they are rather aggregates of several articulatory movements together with their auditory correlates. In acquiring the sound pattern of a language a child achieves automatic combination of the uniquely aggregated movements, and the whole aggregate is then activated by a single neuromuscular impulse. The impulse is in fact the substance vehicle for the realization of the corresponding ultimate phonological unit. In many languages (e.g. German, French) vowel labiality and tongue position are separate units, while in many others (e.g. Russian) they are parts of the same aggregated unit; in the latter case there is no point in regarding one of them as relevant and the other as redundant - they are jointly relevant within the same unit in the given phonological system. As for the part which each of these phonic actions plays within the aggregate, its automatic regulation is performed at a lower sublinguistic level. In French and English the consonantal subsystems contain ultimate units of post-centrality combining in their aggregates the phonic features of velarity, palatality and alveopalatality [7; 8; 11]. But the features are differently grouped and realized in the two languages, and despite their similarity together with the unavoidable common designation each unit is unique in being an element of a specific language system.

Full recognition of the status of language units for the phonological quanta calls for the creation of a suitable term, the customary designation as 'distinctive features' being vulnerable in its two components. To begin with, the word 'feature' is incompatible with the status of a language unit in its own right, as a feature is a mere attribute of a unit of higher rank. Indeed, the term appeared when the phoneme was regarded as the basic phonological unit possessing certain characteristic features. Now, when that notion has been replaced by establishing the ultimate phonological unit as belonging to an independent tier in the system, it must be given a designation that would correspond to the new status and not be an adjunct to the phoneme.

Secondly, the new designation should avoid a reference to distinction as the primary function of the unit in question. Units of every language level fulfil that function, and all language units are equally distinctive. At the same time they are all constitutive within higher units. Consequently, language systems have no need for separate distinctive units, for all distinction is achieved by the use of different constitutive elements. The ultimate phonological unit is no exception: phonemes are distinguished by containing different units of this level. Together with the customary designation we must therefore decline the term 'merism' [2].

The best term for the unit in question was suggested by Jan Baudouin de Courtenay at the beginning of the

century - the blend 'kinakeme' [1], containing the Greek roots for 'movement' and 'hearing' together with the suffix -eme.

Like all the other language units of every level in the macrosystem, the kinakemes are elements in a subsystem of their own, which is naturally not a mere inventory but a well-structured body. Its structure displays two principles. One is thorough binarism - all kinakemes are paired into oppositions of positive vs. negative. Positive kinakemes are materialized as neuromuscular impulses to perform the respective movement or recognize the respective auditory signal; their negative counterparts are realized in the absence of the impulse.

The other structural principle provides for a hierarchy of tiers in the kinakemic subsystem: it always contains two categories (modal and local) with possible subcategories in them and with a further division into kinakemic oppositions. The resulting variety of structural patterns is vast, so that each language usually has a very individual organization of its kinakemic subsystem [10; 41].

The purely negative step of discarding the obsolete notion of universal inventories for ultimate phonological units is obviously insufficient. It must be followed by constructive steps in two directions: first, the kinakemic subsystems are to be described for as many languages as possible; second, a typology of kinakemic subsystems is to be worked out to find their common properties as well as possible diversity in them.

REFERENCES

- [1] BAUDOUIN DE COURTENAY, J. (1910), "Les lois phonétiques", *Rocznik slawistyczny*, III.
- [2] BENVENISTE, E. (1966), "Problèmes de linguistique générale", Paris: Gallimard.
- [3] CHOMSKY, N. (1972), "Language and Mind", New York: Harcourt.
- [4] CHOMSKY, N. & HALLE, M. (1968), "The Sound Pattern of English", New York: Harper & Row.
- [5] FISCHER-JØRGENSEN, E. (1975), "Trends in Phonological Theory: A Historical Introduction", Copenhagen: Akademisk Forlag.
- [6] GRUCZA, F. (1970), "Sprachliche Diakrise im Bereich der Ausdrucksebene des Deutschen. Beiträge zur allgemeinen Sprachtheorie", Poznań: PWN.
- [7] JAKOBSON, R. (1962), "Selected Writings", I, The Hague: Mouton.
- [8] JAKOBSON, R. (1976), "Six leçons sur le son et le sens", Paris: Editions de Minuit.
- [9] JAKOBSON, R. & WAUGH, L. (1979), "The Sound Shape of Language", Harvester Press.
- [10] PLOTKIN, V. (1976), "Systems of Ultimate Phonological Units", *Phonetica*, 33, 2.
- [11] PLOTKIN, V. (1978), "The Kinakeme as the Ultimate Unit of Language", *Kwartalnik Neofilologiczny*, XXV, 3.
- [12] TRNKA, B. (1939), "On the Combinatory Variants and Neutralization of Phonemes", *Proc. 3rd Int. Cong. of Phonetic Sciences*.
- [13] VACHEK, J. (1936), "Phonemes and Phonological Units", *TCLP VI*. Reprinted in: "A Prague School Reader in Linguistics" (1964), Bloomington: Indiana University Press.

MEDIEVAL AND EARLY MODERN ENGLISH SYSTEMS OF VOWEL ORDER: FROM ALPHABETIC TO ORGANIC SCHEMES

Horst Weinstock

Institut für Anglistik, Aachen, Germany

The paper traces the evolution of the vocalic subsystem from its Classical, Medieval, and Early Modern English alphabetic but inorganic order *a e i o u* to its organic but nonalphabetic scheme *i e a o u*. The essentials of the cardinal-vowel system date back to 1762.

Greek and Latin copied the alphabetic pattern of Hebrew. All Hebrew letters formed a macroalphabet or pansystem of names and meanings. Its purely consonantal acrophones and acrographs served as a microalphabetic pansystem of sound- and number-values. The practice of syllabography worked without scripting vowels. Eventually, sound-evolution vocalized acrophonic *Aleph, He, Waw, Yod, Ayin*.

After the vocalization, the Phoenician alphabet reached Greece. The Greeks incorporated, complemented, and regularized the vocalic subsystem. They specified the new vowels as *epsilon, omicron, ypsilon, omega*. The compounded names signalized *psilon* 'plain, simple', i.e. 'monophthongal', *micron* 'small, short', and *mega* 'large, long'. The quantitative and qualitative distinctions expanded the Greek subcatalogue to seven vowel-letters. Dionysius Thrax (2nd cBC) rendered their linear sequence as $\alpha\epsilon\eta\iota\omicron\upsilon\omega$. Roman usage up to Varro (1st cBC) established the Latin scheme *a e i o u*, standardizing the 'megaphonic' type of vocalic length and canonizing the graphic norm and optics of vowel-letters in spelling. In Ireland, the Roman Christian mnemonic *a e i o u* soon ousted the Gaelic Celtic order *a o u e i*.¹

An early breakthrough in anatomic, organic, or phonetic letter-sounds of the microalphabet occurred between the 3rd and the 6th century AD. In the cabbalistic *Sepher Yetsira* or 'Book of Creation', an anonymous Talmud scholar classified letters according to the flow of their breathstream from throat to mouth.² He identified the places of articulation as those stretches of the oral tract along the comparatively static or immovable speech-organs which faced the protruding back, front, blade, or tip of the dynamic or movable tongue. Describing a purely consonantal alphabet, the *Sepher Yetsira* quite naturally skipped the (nonexistent) scripted Hebrew vowel-scheme. Yet in spite of its organic transposition of dentals after labials verbalized for consonants, a merely hypothetical classification of vowels according to the throat-to-mouth arrangement would suggest *u o a e i*.

Although *lingua* as both 'tongue' and 'language' must remain of prime significance for anything linguistic, subsequent grammarians and commentators of the *Sepher Yetsira* could not fail to adjust the monocausal but polyfactorial model of vowel articulation. Dunash Ben Labrat (10th c) and Solomon Ibn Gabirol (11th c) improved the anatomical description of speech-organs and corrected the order of letter-sounds to gutturals, linguals, tectals (or tectals, linguals), dentals, and labials.³

The ways and habits of Roman thinking as well as Patristic epistemology ignored the monocausal but polyfactorial considerations. Not only from

Varro (1st cBC) via Tertullian (2nd/3rd cAD) to Donatus (4th cAD) and Priscian (early 6th cAD) did Roman and Latin grammarians hold on to the mnemonics of the vowel scheme *a e i o u*. Taking on trust any letter's harmony within *nomen-figura-potestas*, the everyday practice of the Latin Middle Ages managed to perpetuate alphabetic aspects in both Romania and Germania. Apart from the identical order of the vowels, the growing neutralization of vowel-length in Romania led to the alphabetic subscheme and mnemonic pattern $\bar{a} \bar{e} \bar{i} \bar{o} \bar{u}$ as against the lengths preserved $\bar{a} \bar{e} \bar{i} \bar{o} \bar{u}$ in Germania. For the phonemic and paradigmatic triad *nomen-figura-potestas*, Boniface (8th cAD) observed but underemphasized the phonetic and syntagmatic transience of contextual vocalization and coarticulation outside the microalphabet.

Aelfric's *Grammar* before 1000 presented the pansystem of the Latin alphabet, expressly adding the unaltered subsystem *a e i o u*.⁴ Byrhtferth's *Manual* in 1011 appended a column with a vocalic *A E I O V*.⁵ About 1150, the *First Grammatical Treatise* just integrated the Germanic unlauted tone-colours of Old Icelandic into an otherwise stable Latin scheme. Its alphabetic insertions followed graphic conventions and largely etymological antecedents.⁶

$a, \bar{a}; e, \bar{e}; i, \bar{i}; o, \bar{o}; u, \bar{u}; y, \bar{y}.$

Aelfric's *Grammar* and Byrhtferth's *Manual* based their vowel schemes upon the *figurae* or written shapes of the letters. Clinging to the alphabetic order, it must have dawned upon the First Grammarian that inadvertent teachers of a harmonious Latin *nomen-figura-potestas* doctrine had been neglecting *nomen* and *potestas*.

An early insular attempt at considering articulatory and acoustic aspects of vocalic order stems from mid-13th-century Oxford. An Oxford Bodleian, a London British Library, and a Paris Bibliothèque Nationale manuscript each hold some pseudo-Grosseteste treatise.⁷ Elaborating upon the Aristotelian differentiation of a vocalic *sonus in motu* from a consonantal *sonus in*

potentia in the Bodleian Digby version, the pseudo-Grosseteste defined vowels as *simpliciter* and consonants as *secundum quid*. A vowel's 'substantial' motion (*motus*) flows without any obstruction, whilst a consonant's 'accidental' motion takes shape from an obstruction at one or more of the speech-organs. With all its inconsistencies, a further treatise by the Digby phonetician (in accordance with the pseudo-Grosseteste) construed the table of vowels upon the particular motions along the speech-organs and points of articulation *guttur, lingua, palatum, os, labia*. The phonetic scheme *a u i o e u* rendered what the pseudo-Grosseteste held to mirror the spectrum of apertures within the oral cavity. Diagrammatically, the types of articulatory motions and acoustic generations resembled geometrical figures and concentric configurations (lines, curves, circles, triangles, and columns).

Roger Bacon (1214-1292) in *Linguarum Cognitione* closed his mind to Robert Grosseteste's (1175-1253) metaphysics of light, acoustics, or cosmology, and to their obvious echoes in the pseudo-Grosseteste. Bacon propagated the Latin scheme *a e i o u* and their Continental pronunciation. His essentials of Hebrew transliterated syllabographic *b* as *ba be bi bo bu*. Even his supralinear equivalents for *Aleph* and *Ayin* just as his phonographic guide to Hebrew punctuation adhered to Latin alphabetic *a e i o u*.⁸

The 14th century yielded no vocalic schemes in sources such as John Mandeville or John Trevisa.

In the 15th-century "De Vigilia Pentecostes", John Mirk recalled the universal importance of the vowel letters and the Varronian and Donatian subsystem *A E I O V*.⁹ In 1499, the anonymous *Promptorium Parvulorum* provided no entry under *vocalis* or *vowel*. The entry under *vocalis* in the anonymous *Ortus Vocabulorum* of 1500 lacks complete schematic exemplification and enumeration.

In the 16th century, the initial phase of the Great Vowel Shift might have stimulated the grammarians' and phoneticians' awareness to reconsider

the hitherto unsuspected conception of harmony in the problematic nature, correlation, coordination, and interdependence of *nomen-figura-potestas*. Yet on the whole, insular Renaissance humanists and Tudor scholars widely studied written sources from a graphic and alphabetic angle; they stabilized the Classical Latin five-vowel subset. Some 23 Tudor authorities went on arranging the vowels in alphabetic order.¹⁰ Smith differentiated between still alphabetic semivowel-plus-vowel clusters and nonalphabetic digraphic monophthongs or peak-and-glide diphthongs. This practice, however, failed to convince prompt imitators.

The 17th century brought no fundamental change. Some 35 publications went on propagating the alphabetic schemes *a e i o u* or *A E I O U*. Sporadically since about 1550, a minor change started taking firm ground: 17th-century phoneticians used to add the Greek allograph *y* for *i*. As marks of a major change, pretty regular inclusions of syllabophonic *ba be bi bo bu* (and *ab eb ib ob ub*), dual schemes of *ā ē ī ō ū* versus *ă ě ĭ ŏ ũ*, and a supplement "*a e i o u* silent" betray a growing sensitivity to nonalphabetic aspects. Realizing shades of timbre or duration as well as a disharmony between vowel-names and sound-values, the corpus attracts attention to contextual (allographic, phonetic, syntagmatic, transient) views of the phonic structure. In principle, the sources did not break with the graphic tradition of the pansystemic alphabet.

Prepared to some extent by Robinson's (1617) "Scale of Vowels" *u o a e i* from back to front, by Price's (1665) "Throat Vowels" *u o e i a*, and by Wilkins's (1668) "Sound Chart" and "Organic Alphabet", William Holder's *Elements of Speech* (1669) advanced the phonetic sciences considerably.¹¹ Conceding a concurrent share of lips and throat in the generation of vowels, Holder recognized the free passage of "Breath Vocalized" through the cavity of the mouth. The shape and mechanism of tongue and oral cavity form the main cause of the number and the main reason for a natural or organic order of the various vowels.

"... and then the Series of the Vowels according to their degrees of aperture, and recess towards the *Larynx*, will be thus, *i e, æ, a, α, o, oo*; to which may be added *u* and *y*."¹²

Although Holder's theory did succeed in beating a path to phonetics, his (like Price's and Wilkins's) practice fell back upon the old vice of alphabetic order. The graphic schemes *a e i o u* (*y*) or *A E I O U* (*Y*) continued to survive in some 24 late-17th-century teachers.

18th-century documents carried on alphabetic schemes and aspects in some 42 arts of poetry, dictionaries, dissertations, elementaries, essays, grammars, guides, institutes, introductions, repositories, rudiments, spelling-books, and treatises. Again, slightly more than one in four authorities thought fit to specify the morpho-phonemic sound-values by means of syllabographic *ba be bi bo bu* (*by*) and *ab eb ib ob ub*.

Confronted with notational needs in a period of no adequate transcription, 18th-century phonologists resorted to diacritical, numerical, or typological devices. Some augmented the alphabetic order of vowels by supraposing accents or figures above polyphonic characters; others implemented an etymological alphabet of historical "representatives" or allographs for phonemic transcription.

All in all, the tentative solutions in marking, listing, and ordering the spectrum of vocalic timbres got phoneticians nowhere.

In 1762, Henry Home (1763 Lord Kaimes/Kames) published his three-volume *Elements of Criticism* which, within seven years, went into four editions. Referring to Harris's *Hermes* (1751) and to the then contemporary anatomists, John Rice in 1765 rejected Lord Kaimes's suggestion that the five vowels showed the same extension of the windpipe but different openings of the mouth, and that the vowel scheme formed a regular series of sounds descending from high to low in the organic order *i e a o u*.

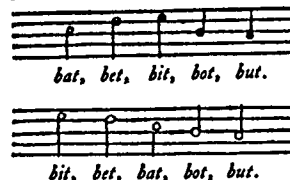
"Neither a higher nor lower Note can proceed from the Lips of the Mouth, than first proceeds from the Lips of the *Glottis*."¹³

Rice compared the musical notation of alphabetic *a e i o u* with the rise and

fall of the syllabophones *bat bet bit bot but*. He produced evidence that

"... they are not all equally grave or acute."¹⁴

Sequencing the musical notation in a steady series from high to low



the orthoepist concluded from the reverse *bit bet bat bot but* that

"... there must at least be Syllables of five different Lengths: And this is what I mean by the natural Length of Syllables."¹⁵

Whatever John Rice's "natural length" and "glottal tone" may have meant (duration, fundamental frequency, harmonics, timbre, wave-length rather than pitch), none of the celebrated authorities after him seems to have recognized the gigantic stride of his contribution. Rice's doctrine failed to gain acceptance with some 33 British and American authorities before 1800.

Nevertheless, John Rice anticipated Daniel Jones's cardinal-vowel system as a standard invariable scale. For the sake of a universal and uniform phonetic notation, Rice abandoned the alphabetic order and graphic orientation of the vowel-scheme. The long lost harmony of *nomen-figura-potestas* had ended up in an uncontrolled history of partly etymological or allographic and partly contextual or syllabophonic inconsistencies. Rice promoted a nonalphabetic method which keyed one chief cause to several companion factors. He integrated the static firmament of the more or less immovable speech-organs into the dynamic zeniths of the movable tongue-positions, fixing the sound-values of vocalic articulation and modulation to a scale from "front high" via "mid low" to "back high". Under the circumstances of the reverse directions of thinking and writing, the new order perfectly agreed with the old of the *Sepher Yetsira*. John Rice's phonic model of 1765 as an oral alphabet converted the alphabetic but inorganic and unnatural order of the graphic sub-

system *a e i o u* into the nonalphabetic but organic and natural scheme *i e a o u* of the cardinal-vowel system.

- 1 PARSONS, J. (1767/1968), *Remains of Japhet*, English Linguistics 1500-1800, No. 64, Menston: Scolar, 404, 412. — For a historical introduction to "The Letters of the Alphabet", see MICHAEL, I. (1987), *The Teaching of English*, Cambridge: C. U. P., 25-32.
- 2 BACHER, W. (1892-95/1974), *Die Anfänge der hebräischen Grammatik und Die hebräische Sprachwissenschaft vom 10. bis zum 16. Jahrhundert*, Amsterdam Studies in the Theory and History of Linguistic Science, III: Studies in the History of Linguistics, 4, Amsterdam: Benjamins, 20-22.
- 3 BACHER, op. cit., 21, n. 2.
- 4 SOMNER, W. (1659/1970), *Dictionarium Saxonicum-Latino-Anglicum*, English Linguistics 1500-1800, No. 247: *Aelfrici Grammatica: "De Litera"*, Menston: Scolar, 2.
- 5 CRAWFORD, S. J. (1929/1966), *Byrhtferth's Manual (A.D. 1011)*, Vol. 1, Early English Text Society: Original Series, 177, London: O. U. P., 195-196A.
- 6 HAUGEN, E. (1972), *First Grammatical Treatise: The Earliest Germanic Phonology*, London: Longman, 17.
- 7 REICHL, K. (1976), *Tractatus de Grammatica: Eine fälschlich Robert Grosseteste zugeschriebene spekulative Grammatik*, Edition und Kommentar, Münchener Universitätschriften: Veröffentlichungen des Grabmann-Institutes, (New Series) 28, München, Paderborn: Schöningh, 24-27, 135, n. 16.
- 8 NOLAN, E./HIRSCH, S. A. (1902), *The Greek Grammar of Roger Bacon and a Fragment of His Hebrew Grammar*, Cambridge: C. U. P., 202. — BRIDGES, J. H. (1893/1964), *The Opus Majus of Roger Bacon*, Pars Tertia: *Linguarum Cognitione*, Frankfurt/Main: Minerva, 91.
- 9 ERBE, Th. (1905/1973), *Mirk's Festival*, Part I, Early English Text Society: Extra Series, 96, Millwood, New York: Kraus, 156-157.
- 10 For the corpus of Early Modern English texts, see ALSTON, R. C. (1967-1972), *English Linguistics 1500-1800*, 365 vols., Leeds/Menston: Scolar.
- 11 For Robinson's "Scale of Vowels", see JONES, Ch. (1989), *A History of English Phonology*, London: Longman, 213-218. — HOLDER, W. (1669/1967), *Elements of Speech*, English Linguistics 1500-1800, No. 49, Menston: Scolar, 24-91, 132-143.
- 12 HOLDER, op. cit., 90.
- 13 RICE, J. (1765/1969), *An Introduction to the Art of Reading*, English Linguistics 1500-1800, No. 161, Menston: Scolar, 23.
- 14 RICE, op. cit., 40.
- 15 RICE, loc. cit., 41.

EVALUATION QUANTITATIVE DE L'ALTERNANCE PHONETIQUE DU /ə/ IMPORTANCE DE L'ENTOURAGE CONSONANTIQUE

J. VAN EIBERGEN

Institut de la Communication Parlée, URA CNRS n° 368
Grenoble, France

ABSTRACT

This study attempts to show the statistical importance of the realisations [±] of the french voyell /ə/. It also brings out the phonological rules of its obliteration on a phonetic level. Finally, the influence of the consonantal context on this voyell is brought out.

1. INTRODUCTION

La voyelle /ə/ appelée couramment "muette" ou "latente" et plus spécifiquement "schwa" par les phonologues, constitue un des particularismes du français. Elle se caractérise par une variation de réalisations [±] dans la chaîne parlée. A cause de cette alternance vocalique unique en français, nous qualifierons ce /ə/ de "bifide" pour donner l'image de deux unités séparées dont chacune d'elles appartient à la même entité. L'élaboration des règles d'effacement ou de maintien de cette voyelle sur un plan phonologique et phonétique, est d'autant plus complexe que les facteurs déterminant son fonctionnement sont nombreux et qui plus est, se situent à des niveaux différents (phonétique, phonostylistique, sociolinguistique..)

2. EVALUATION QUANTITATIVE DE L'ALTERNANCE DU /ə/

Nous nous proposons de présenter tout d'abord des résultats statistiques [4] concernant la double réalisation du /ə/ à partir d'un corpus de français vernaculaire, à caractère spontané et implicatif, de 17 000 phones environ [3]. C'est dans ce type de français que les effacements du /ə/ sont les plus fréquents en effet, le nombre d'occurrences de sa réalisation [-] le classe au deuxième rang après le [a], par rapport à l'ensemble des unités vocaliques et consonantiques recensées. D'autre part, parmi les autres unités phoniques réalisées [-], le /ə/ représente, à lui seul, plus de 80% de ces

effacements. Le nombre de /ə/ réalisés [-] par rapport à ceux réalisés [+] est trois fois supérieur. Par ailleurs, 65% des réalisations du /ə/ sont en position interconsonantique dont un quart seulement se réalise [+]. Son rôle de liant ou d'isolant consonantique est bien un des plus importants et ceci explique que la majeure partie des travaux de recherche porte sur la bifidité du /ə/ dans cette position. Le pourcentage des groupes consonantiques résultant de l'effacement du /ə/ correspond à 33% de l'ensemble des groupes recensés. Le pourcentage de chaque type de groupes augmente proportionnellement en fonction du nombre de consonnes contenues dans ces groupes: le /ə/ est responsable de la réalisation de 30,4% des groupes de 2 consonnes, de 50,2% des groupes de 3 consonnes, de 90,5% des groupes de 4 consonnes et des quelques rares groupes de 5 consonnes. Parmi ces groupes, 80% se situent à la frontière de mot.

Enfin, dans la mesure où le /ə/ peut toujours passer du zéro phonique à sa réalisation effective, il peut apparaître aussi pour jouer le rôle de processus d'attente ou d'hésitation et cela représente statistiquement 37% de ses réalisations [+]. Il existe d'autres pauses vocaliques mais le /ə/ représente de loin le pourcentage le plus élevé, soit 66,8% des occurrences de ce type de processus d'attente.

3. APPLICATION DES RÈGLES PHONOLOGIQUES D'EFFACEMENT DU /ə/

Nous avons ensuite testé la validité des règles phonologiques d'effacement du /ə/ [2], à partir de notre corpus. Ces règles concernent majoritairement le /ə/ en contexte consonantique et nous nous limiterons à ce type de contexte problématique. Il existe de nombreux facteurs qui expliquent la bifidité du /ə/ en position interconsonantique, tels que

sa place par rapport à l'accent, par rapport à la coupe syllabique. Intervient aussi le type de français réalisé en fonction d'une situation de communication donnée. F. DELL a retenu surtout la place du /ə/ par rapport à la pause, le nombre de consonnes précédant ou suivant le /ə/ et la nature de ces consonnes. C'est le nombre de consonnes précédentes qui détermine principalement les réalisations de la voyelle bifide. Interviennent ensuite, dans certains cas seulement, celles qui suivent le /ə/ en fonction du nombre de celles qui le précèdent. Enfin, la nature des consonnes entre en jeu et leur place les unes par rapport aux autres.

Nous rappellerons brièvement les règles de F. DELL en donnant pour chacune, des exemples cités par lui-même et les résultats obtenus tirés de l'application de ses règles sur notre corpus.

3.1. Précédé d'une consonne.

En début de groupe rythmique et en syllabe initiale de mot "*v(e)nez ici*", "*d(e)vant moi*". La règle d'effacement est facultative et le /ə/ tombe d'autant plus facilement qu'il est éloigné de l'accent principal de groupe rythmique. Précisons que F. DELL considère que les /ə/ appartenant à des monosyllabes fonctionnent de la même manière que ceux en syllabe initiale de mot. Le /e/ ne s'efface pas, en revanche, dans un contexte consonantique occlusif "*te casse pas la tête*", "*de quoi tu te plains*". Nous avons rencontré plusieurs contre-exemples dans des monosyllabes, "*j* trouve*", mais aussi dans des polysyllabes, "*d*puis*". Dans ces exemples, les deux consonnes en contact, après effacement du /e/ sont sourdes, de nature ou par assourdissement. Le caractère sourd du contexte semble faciliter l'effacement du /ə/ et nous observerons ce phénomène dans bien d'autres positions. Retenons malgré tout qu'il s'efface une fois sur cinq environ dans ce type de position sauf dans le mot outil "*je*" très fréquent dans le français vernaculaire. En effet, plus de 77% de /ə/ dans cette position se sont réalisés [-], que ce soit dans un contexte assourdi après effacement, "*j* prends*", ou dans un contexte sonore, "*j* vois*".

A l'intérieur du groupe rythmique et toujours en syllabe initiale de mot, le /ə/ s'efface facultativement, "*la s(e)crétaire*". Cette règle d'effacement est facultative. Dans notre corpus, une forte majorité de /ə/ se sont effacés dans cette position. Par ailleurs, F. DELL parle d'exceptions à cette règle, des mots tels que "*guenon*", "*peser*", "*vedette*". Parmi les dites

exceptions le /ə/ se réalise [+] pour éviter des groupes consonantiques rares en français, "*guenon*", "*guenille*" et inexistant dans cette position. IL se maintient aussi dans certains entourage consonantiques tels que - b + e + d - = "*bedeau*", - l, s + e + v r - "*levraut*", "*sevrer*" dans lesquels interviennent les consonnes qui suivent. Dans de nombreux mots dits "littéraires" et souvent bisyllabiques le /ə/ ne s'efface pas non plus, "*semonce*", "*ledir*". Enfin, rares sont malheureusement les cas d'opposition phonologique du type "*belette*", "*blette*".

A l'intérieur de mot, "*ach*teur*", "*massiv*ment*", le /ə/ s'efface obligatoirement. F. DELL relève par ailleurs des exceptions à cette règle, "*champenois*", "*attendant*", "*dépecer*", et considère ces mots d'un usage peu fréquent. Mis à part dans le premier exemple, il y a la présence de préfixe; le /ə/ fonctionnerait alors comme en syllabe initiale de mot. Dans notre corpus, nous avons rencontré le même phénomène "*enregistré*", "*démasure*". F. DELL précise aussi que le /ə/ se maintient devant un groupe [lj] même s'il n'est précédé d'une seule consonne "*hôtelier*". Ceci montre que les consonnes qui suivent interviennent aussi dans le fonctionnement du /ə/.

En fin de polysyllabe, "*une vieill* courtisane*", le /ə/ tombe obligatoirement. F. DELL exclut évidemment les cas où les mots qui suivent commencent par un "h" aspiré. Il écarte aussi les contextes où le /ə/ est suivi du mot "*rien*", dans lesquels sa réalisation est variable, "*il mang(e) rien*". Dans notre corpus, la règle d'effacement s'est appliquée sauf lorsque le /ə/ était un processus d'attente car il peut apparaître parfois, dans le groupe rythmique, "*une page spéciale*".

En finale de polysyllabe à la fin du groupe rythmique, "*elle est trop petit**", la règle d'effacement est obligatoire. F. DELL propose de considérer les mots "*lorsque*", "*puisque etc.*" non comme des exceptions mais comme des mots composés dont le deuxième terme est "*que*". Il fonctionnent alors comme des monosyllabes qui ne s'effacent jamais en fin de groupe rythmique devant une pause. Hormis ces mots outils, nous avons dénombré de nombreux contextes dans lesquels le /ə/ a été maintenu. Il joue alors le rôle de processus d'attente. A quel niveau doit-on formaliser des règles rendant compte de ce type de réalisation [±] ?

3.2. Précédé de deux consonnes

La chute du /ə/ dépend essentiellement de la nature des consonnes qui le précèdent. En syllabe initiale de mot et à l'initiale ou à l'intérieur de groupe rythmique, lorsque les deux consonnes appartiennent au même mot le /ə/ n'est jamais réalisé [-], non pas à cause de sa position, mais parce que les groupes consonantiques dans cette position, sont des suites - occlusive + constrictive -, "prenez tout", sauf dans le mot "squelette", dans lequel le /ə/ sera aussi maintenu.

En syllabe initiale de mot, lorsque les deux consonnes sont séparées par une frontière de mot, "j'arriv* demain", le /ə/ ne s'efface jamais sauf dans certains mots qui seraient prononcés avec un débit très rapide, "quell* s*main*". Nous avons trop d'exemples dans notre corpus dans lesquels le /ə/ s'est réalisé [-] pour considérer qu'il s'agit d'exceptions, "pour d*main", "j'ai pas mal c* matin". Son effacement dépend, dans ce cas, de la nature du groupe consonantique le précédant. En effet, si ce groupe commence par un [R], il peut se réaliser [-]. S'il commence par un [l], cette liquide doit être suivie par une constrictive; si au contraire elle est suivie par une occlusive le /ə/ se maintient, "pas mal de copains".

À l'intérieur d'un polysyllabe, "surgelé", "entretien". Pour F. DELL, le /ə/ ne s'efface jamais dans cette position. Comme en syllabe initiale, tout dépend de la nature des consonnes. Dans nos exemples, le /ə/ s'est effacé après un groupe consonantique commençant par un [R] dans des mots comme "gouvern*mental", "charg*ment", "vers*ment" etc. On a des réalisations [+] du /ə/ dans un même contexte consonantique mais dans des mots peu fréquents, "prosternement", "regorgement", "ressourcement". On peut en conclure que le /ə/ s'effacerait dans ce contexte à condition que le groupe de consonnes qui le précède commence par un [R] et à condition que le mot soit très fréquent.

En fin de groupe rythmique, le problème est identique à celui du /ə/ précédé d'une consonne. La voyelle bifide peut se réaliser [+] et jouer alors le rôle de processus d'attente.

2.3. En syllabes contiguës

F. DELL traite à part les suites de /ə/ contenues dans des syllabes qui se succèdent. Il énonce un principe fondamental qui régit sa réalisation [-] : les règles d'effacement peuvent supprimer autant de schwas qu'on veut tant que leur

effacement n'engendre pas de groupes de trois consonnes dont les deux dernières étaient isolées par un /ə/. Autrement dit, dans la suite -vcəcə- le premier /ə/ peut s'effacer; le deuxième est alors précédé de deux consonnes et doit être maintenu; seul un schwa peut s'effacer dans cette suite. D'après lui, l'énoncé "il veut que ce travail soit bien fait" n'a que deux réalisations possibles: "veut qu* ce travail" ou "veut que c* travail". La formation de groupes de 3 consonnes dus à l'effacement de cette voyelle est malgré tout possible si la distribution des éléments consonantiques par rapport au(x) /ə/ permet l'application des règles d'effacement. Dans l'exemple de F. DELL "prenez l* train", le /ə/ du monosyllabe n'est pas maintenu car il n'est précédé que d'une seule consonne. Nous avons rencontré 16 contextes de /ə/ en syllabes contiguës dans lesquels deux effacements successifs se sont produits en contradiction avec le principe fondamental énoncé par F. DELL, "il faut que* j* travail*.". Le caractère sourd par nature ou par assourdissement des consonnes, après la chute du /ə/ semble faciliter son effacement.

4. CONCLUSION

La fréquence des réalisations [±] du /ə/ est extrêmement élevée dans la chaîne parlée et cette voyelle pose encore de nombreux problèmes quant à son fonctionnement. Pour expliquer cette variation de réalisation il faut tenir compte des différents facteurs déterminant son fonctionnement et tenter de les hiérarchiser les uns par rapport aux autres de telle sorte que l'on puisse trouver des règles conditionnelles contextuelles d'effacement. Ces règles devraient expliciter d'une part, les exceptions aux règles obligatoires, d'autre part, le caractère facultatif des règles de F. DELL.

5. RÉFÉRENCES

- [1] LUCCI, V. (1978), "Reconnaissance des variantes socioculturelles et situationnelles du français parlé. Evaluation des paramètres". Bulletin de l'Institut Phonétique de Grenoble, 7, 33-66.
- [2] DELL, F. (1985), "Les règles et les sons", Hermann.
- [3] VAN EIBERGEN, J. (1986), "Corpus d'un français vernaculaire à caractère spontané et d". Bulletin de l'Institut de Phonétique de Grenoble, 15, 35-73.
- [4] VAN EIBERGEN, J. (1986), "Réalizations et rôles du g bifide", Thèse de 3ème cycle, Institut de Phonétique de Grenoble, 337 p.

TYPOLOGY OF GERMANIC MORPHOSYLLABISM

YU. K. Kuzmenko

Institute of linguistics, Leningrad, USSR.

ABSTRACT

Syllabic languages where the syllable is always a minimally meaningful unit and represents one morpheme possess phonological features which are common for these languages irrespective of their genetic relationships (relevant syllable division, monosyllabism, tones specific syllable structure with syllable initials and finals differing both in number and quality). The subject of this paper is to trace phonological changes in Germanic languages which increase their affinity with syllabic languages.

1. CHANGES IN CONTACT TYPE AND SYLLABLE DIVISION

One of the most important changes in history of Germanic languages is the change of the correlation between vowel length and syllable division within the world. In old Germanic languages CVCV-sequences had open syllables irrespective of vowel quantity and such free length is preserved in some modern High Alemannic, South Bavarian and Scandinavian dialects. In modern Germanic languages the syllable is always closed after a short vowel (close contact) and open after a long one (loose contact). Thus modern Germanic languages show the development from CV-CV language to CVC-V one. In the overwhelming number of close

contact words of CVCV(C) type CVC sequences represent a root morpheme and the syllable and morpheme boundary coincide. Standard High German is a typical example of a language with the contact correlation where the type of contact reflects chiefly the preceding opposition V:C-VC:. However in many Germanic languages the number of close contact words increases at the expense of loose contact words, thereby increasing the number of words with the coinciding syllable and morpheme boundaries. The first change that increases the number of close contact words and leads to the monophonemisation of original VC-sequences is the contact shift in the combinations V: + j, w that occurred in Middle English (cf. OE *growan*, ModE *grow*), in Frisian, Dutch and Low German dialects. This trend is quite obvious if we compare Middle West Frisian which possessed 6 so called long diphthongs (i.e. biphonemic combinations of V+C) with the modern Frisian dialect of Schiermonnikoog where 5 of them were shortened and j and w got incorporated into the syllable nucleus [1]. Though this change is not often the case in High German dialects it can be observed even there (cf. Low Alemannic /sauə /, /sdeia / Standard High German *sagen*, *steigen*)[2]. The same type

of change is now taking place in Danish (cf. /bre'vəd / > /breu'əd / *brevet*, /fla'ʝən / > /flai'ən / *flagen*). Not only [j] and [w] are apt to change the contact type and to become a part of a monophonemic diphthong but also the resonants /r/, /l/ and /n/ can vocalize merging with the preceding vowel. Such is the development of the postvocalic /r/ in English, Danish, Low and High German dialects, the development of /l/ in Low and High German dialects, Dutch and English (cf. modern trend to vocalize /l/ both in *filling* and *feeling*) and the incorporation of /n/ into the nasalized vowel in various modern Germanic vernaculars.

The other type of contact shift leading to the increased number of close contact words affects root morphemes with voiceless stops and high vowels. We know that the vowel duration is dependent on the vowel height and on the quality of the consonant (the shortest are narrow vowels followed by voiceless stops). The degree of V + C contact seems to depend on the same factors [3]. The change of the contact type (loose > close) of vowels (especially narrow) + voiceless or tense plosives can be observed in English, Frisian, Dutch, Low German and Danish dialects. In Frisian this change affects chiefly the combinations which are most suitable to be shortened (narrow vowels + voiceless stops). In many words here the contact shift is already completed (e.g. *dyk*, *bite*, *buk*) in some words it is still in progress (cf. free variations of contact type in *siik* /si:k/ - /sik/ or *broek* - /bru:k/ - /bruk/). If Selkirk [4] and Kukolshchikova [5] are right and the syllable

with postvocalic tense stops in English are always closed irrespective of the quantity and the quality of the preceding vowel (words like *pity* and *peaty* having the same type of contact and the same type of syllable division), we can suggest that the close contact ousted the loose one in all words with original long vowels followed by tense stops. In Dutch the contact type changes in the combinations of original /i:/, /y:/ and /u:/ with any consonant except /r/. Vowel length and syllable division in the words like *gieten* and *boeken* are the same as in the words *pitten* and *putten*. In both cases we have the same type of closing command after the short vowel [6].

2. INCREASE IN DIFFERENCE BETWEEN INITIALS AND FINALS

One of the most apparent phonological features of the syllabic languages is the qualitative and the quantitative difference between initials and finals. This difference coupled with morphologically determined syllabification indicates a particular manifestation of the morphological boundaries in a text. The number of initials chiefly consisting of released consonants, glottal stop, /h/ and consonant clusters exceeds considerably the number of finals which can vary from 13 in Mon Khmer languages to 3 in eo (i, u and n). Consonant clusters are intolerable as finals. The processes resulting in forming the same type of correlation between initials and finals are going on in Germanic languages. The simplification of final clusters CC occurs here according to two patterns: vocalization and nuc-

leation of the first consonant or deletion of the second consonant. In both cases the pattern CVCC is ousted by the pattern CVC. Vocalisation affects at the first place the resonants and it is characteristic of English, Dutch, Afrikaans, Frisian, Danish, Low and High German dialects (cf. the changes VrC>VC, Vlc>VC, Vnc>VC). The second pattern is the deletion of stops. In Afrikaans two types of stop final clusters were simplified chiefly by the deletion of final stop /t/ after obstruents and all stops after resonants [7]. The deletion of final stops is a characteristic feature of the Jutlandic Danish [8], some Low German and English vernaculars.

3. TONES

Every morphosyllable in the syllabic languages is characterized by a special tone. Most typologically similar to the tones of the syllabic languages are tones in Danish (Jutlandic) and Low German dialects where they occur exclusively in monosyllabic words. The tonal distinctions reflect here the original distinctions of monosyllabic and bisyllabic words (cf. Jutlandic Danish /hu:s/ - /hu:s/ Standard Danish hus, huse). In Franconian dialects the tonal distinctions are also largely characteristic of monosyllabics and reflect original opposition of mono- and bisyllabic words but due to the so called spontaneous and combinatory accentuation the tone of the apocope can occur both in original monosyllabic and preserved bisyllabic words. Even though the problem of the origin of the tonal distinctions in Germanic languages can not be considered as finally

solved there is much evidence that the traditional idea that the tones in Danish, Low German and Franconian dialects appeared in the period of the apocope is valid. Spontaneous and combinatory accentuation in words with original long broad vowels and voiced consonants in original monosyllabics and preserved polysyllabics in Franconian can be explained as depending on their longer duration connected with the quality of the corresponding vowels and consonants. The tonal distinctions become relevant in the period of the apocope, one of the phonetic features of the apocopated words being length. At this moment phonetically longer duration of the broad vowels and of the vowels before the voiced consonants become apocopically accentuated even in words which were not affected of the apocope. Thus the Low Franconian dialects where the longer duration is one of the features of the apocopated words and of the words with spontaneous and combinatory accentuation reflect the older stage of the development whereas the central Franconian "Scharfung" in apocopated and spontaneous/combinatory accentuated words is the result of metatony. In English there is a trend to an abrupt ("entering") tone to be formed in the words with unreleased tense stops. In West Jutlandic dialects we can see two types of the same kind of abrupt tones.

All above mentioned changes in spite of their seeming differences are the expression of one trend, the trend of morphosyllabism which is characteristic of the development of Germanic languages.

4. REFERENCES

- [1] Spenter A. (1968) *Der Vokalismus der akzentuierten Silbe in der Schiermonnikooger Mundart*. Kopenhagen.
 [2] Жирмунский В.М. (1956) *Немецкая диалектология*. М.:Л.
 [3] Gondaillier J.P. (1973) *Coup ferme et coup lâche, application de ces conceptions au passage voyelle-consonne occlusive en koinè de Luxembourg-ville*. - *TIPhS* N 5.
 [4] Selkirk E. (1982) "The syllable", *The structure of phonological representation*. ed. van der Hulst, H. Smith. T.2. Dordrecht.

- [5] Кукольщикова Л.Е. (1984) Об одном спорном случае слога деления в английском языке // *Экспериментально-фонетический анализ речи*. Вып. 1. 29-38.
 [6] Nooteboom S.G. Slis I.H. (1972) "The phonetic feature of vowel length in Dutch", *Language and speech*, v.15, n 4..
 [7] Ponelis F.A. (1989) "Ontwikkeling van klusters op sluitkranke in Afrikaans, Suid-Afrikaanse Tydskrif vir Taalkunde", v.7, n 1.
 [8] Bengtson D.B. (1985) *Yngre regionaldansk*, *Danske folkemaal*, 27.

ISOMORPHOUS AND ALLOMORPHOUS CHARACTERISTICS OF THE
CAUCASIAN AND SOME INDO-EUROPEAN LANGUAGES IN THE
FIELD OF PHONETICS AND PHONOLOGY

Lily A. Ponomarenko

Pedagogical Institute, Zhitomir, USSR.

ABSTRACT

Similar and distinctive phonetic and phonological features of a number of languages of diverse types (analytical English, synthetic Russian and Ukrainian, and agglutinative Caucasian languages with some touch of polysynthetic characteristics in north-western branch and fusion in north-eastern branch) have been ascertained.

The main methods used in the investigation were: method of analytical comparison and questionnaire method.

1. INTRODUCTION

The choice for analysis of the languages of diverse types was conditioned by the fact that their fundamental characteristic features matter not only to the morphological and word-building levels, but also to other ones, including phonetics and phonology. For example, the leading feature of the agglutinative languages is haplosemy that is the attachment of one element of the form to one element of the content, which provides a higher degree of stability of the language system than the availability in the inflexional languages of synthetosemy (simultaneous polysemy) creating asymmetry. The latter involves fluctuating articulatory norms.

2. RESULTS AND DISCUSSION

If we compare such Indo-European languages as English, Russian, Ukrainian with the Caucasian languages, we shall observe more advanced articulation of the former set of languages. The sound systems of the Caucasian languages contain velar, pharyngeal and partly laryngeal phonemes. In both English and Ukrainian there is only one pharyngeal phoneme (rendered by the letter "h" in English and "r" in Ukrainian). Russian has no pharyngeal sounds at all. In all three languages there are no laryngeal consonants.

One may note some tendency for rapprochement of Caucasian phonological systems to those of the Germanic and Slav languages under review. We mean the advance of the articulations of pharyngeal and laryngeal series in the Caucasian languages. The strong glottalized affricates turned into the corresponding aspirate sounds this way in Tindin.

Separate phonemes of the Caucasian languages are articulated differently; for example, lateral consonants in some Caucasian languages (Georgian, Zan, Rutul, Udi) are similar to the corresponding Russian and Ukrainian phonemes (dental), in some (Lezgian, Lack, Dargi, Agul, Tabasaran, Tsakhur) to

English ones (alveolar), in some — different from the corresponding phonemes of the Indo-European languages under review (front palatal — in Budukh and Hinalug; noisy — in Kabardian).

Besides the privative binary opposition according to a distinctive feature "resonance/lack of resonance", inherent in consonantal systems of all the languages under review, in the Caucasian languages there is one more opposition closely interwoven with the former; breath consonants can be aspirate and checked. That equipollent opposition embrace only obstructive sounds. It does not apply to spirants.

English is vocalic, while the Caucasian languages, as well as Ukrainian and Russian, belong to a consonantic type. The consonantal system is especially developed in Ubykh (80 consonants), Abaza (66 consonants), Hinalug (59 consonants), etc. The availability of the small number of vowels (2 - 3 vowels in some Caucasian languages) predetermines the absence of restrictions in their use and vice versa: the availability of the large number of vowels creates prerequisites for such limitation. For instance, open syllables cannot be concluded with short vowels in English, where there are many (27) vowels.

Accumulation of a great number of diverse consonants is a rare phenomenon for all the languages. This universal is connected with the tendency of effort economy: it is difficult to pronounce the great number of consonants without vowels. But in a small quantity of cases such clusters are found even in vocalic English. As to the

possibility of flowing several diverse consonants together, it is on the average more characteristic of Caucasian languages than of English, Russian and Ukrainian because there are fewer vowels in the former. The location of adjacent consonants and their highest possible number is individual for each language. For example, consonant clusters in the final position are typical for Svan and Tabasaran whereas ones in the initial position and inside the word are typical for Georgian. Russian can tolerate a cluster of four consonants in preposition, while English permits only three.

Some Caucasian languages (different sets) have oppositions; analogous to English — short/long vowels (Chechen, Ingush, Hunzib, Lack), open/close vowels (Chechen); analogous to Russian and Ukrainian — hard/soft consonants (Adyg, Abkhazian, Abaza, Ubykh). It is necessary to mention that the force of the opposition "open/close vowels" is great neither in English, nor in Chechen.

Some Caucasian languages (different sets) have phonological oppositions absent in English, Russian and Ukrainian: a) palatalization/lack of palatalization of a vowel (Svan, Udi); b) orality/nasality of a vowel (Batsby, Botlikh, Godoberin, Karatin, Hunzib); c) labialization/lack of labialization of a consonant (Abaza, Abkhazian, Adyghe, Kabardian, Ubykh).

From suprasegmental units we shall dwell on accent. The accent in all the languages under review, except the Rutul language and the Munib subdialect of Andy characterized by tonic (musical) accent, is dynamic.

It is in Russian where the accent is expressed most strikingly, a little bit less — in English, still less — in Ukrainian, and quite slightly — in many Caucasian languages.

Within the Caucasian languages themselves, even closely related, the accent has different intensity. For example, it is weaker in Andy than in Avar (both refer to Avar-Ando-Tsez subgroup of Daghestan group of languages); it is weaker in Georgian and Zan than in Svan (all three refer to Kartvel group of Caucasian languages). Weak stress of Modern Georgian literary language resembles the sea after a storm /5, 14/.

The degree of unstressed vowels reduction is connected by direct dependence with accent intensity. That is at the bottom of intensive reduction of unstressed vowels in Russian. That is the reason of stability for phonetic changes in Modern Georgian where there is weak stress and, on the other hand, frequent reduction of vowels (right up to their falling out) in the Old Georgian language, where strong stress dominated. The relatively strong accent of Modern Svan, Abkhazian, Abaza, Lezgian also results in the reduction of unstressed vowels.

English prefers close syllables, Russian and Ukrainian give preference to open ones. Caucasian languages are not identical in this respect. Even within one south-Caucasian branch the indices, according to our calculation, are quite different: close syllables prevail in Svan, whereas open ones dominate in Georgian and Zan, share of open syllables in Zan exceeding their share in

Georgian. Note should be taken that a close syllable was typical for Georgian historically /2, 29/. In the course of the Georgian language development the quantity of open syllables was being increased. For instance, the number of close syllables is greater in "Hero in tiger's fell" by Shota Roostavelli than in the works of literature by modern Georgian writers, although even there the quantity of open syllables prevailed over close ones.

Vowels serve as syllable-building sounds in Lack, Russian and Ukrainian, while in English not only vowels but also sonorous consonants "n" and "l" can fulfil that function: [kə:|tn], [ou|ld].

Let us dwell on living phonetic processes.

In contrast to English, Ukrainian, and Kabardian, where the neutralization of the opposition "voicelessness/resonance" in the final consonants does not take place, in Russian, Budukh, Lezgian the final voiced consonants are devoiced.

The neutralization of that opposition occurs in an ultima in Ingush, but the direction of the phonetic process is opposite: voiceless fricative affixal consonants "c", "ш", "x" are sonorized.

The notions "accommodation" and "assimilation" are often mixed being used as absolute synonyms. We consider that accommodation presupposes the adaptation within the bounds of one phoneme, while assimilation presupposes the substitution of one phoneme for another.

In the languages under review one can meet both regressive and progressive assimilation. But their proportion in various languages is different. In some

Caucasian languages (similar to Russian and Ukrainian) regressive assimilation prevails over progressive (Bezhitin, Zan, Adyghe, Tsez), and in some (similar to English) — quite the reverse (Chechen, Tabasaran).

Vowel harmony — non-contiguous assimilation of affix vowels to root ones — functions in some Caucasian languages like in Turkic languages (and often under their influence). A final vowel is liable to likening in Avar. Preverbs receive vocalism depending on a vowel in the root of the verb in Tabasaran. That phenomenon is not observed in the Indo-European languages under review.

Dissimilation is peculiar to a number of Caucasian languages (Abkhazian, Andy, Lack, Svan, Zan). It is found comparatively seldom in English, Russian, and Ukrainian (English "laurel" came from "laurer", Russian "верблюд" came from "велоблюд", Ukrainian "лицар" came from "рицар").

In all the Indo-European languages under review the speech of a male and a female is less differentiated than, for instance, in Hushtadin subdialect of the Bagvalin language. Intervocalic "д" turns into "p" in women's speech, while in men's speech "д" is not changed.

3. CONCLUSION

The comparative research of the given languages is interesting not only from the point of view of typological theory but also from the standpoint of practical application for the intensification of education process.

Reference to the isomorphous phenomena in a

mother tongue will save time on explanation, while attention to the allomorphic phenomena will help to avoid interference.

4. REFERENCES

- /1/ АРАКИН В.Д. (1979), "Сравнительная типология английского и русского языков", Ленинград: Просвещение.
- /2/ ДЕНЕРИЕВ Ю.Д., ЧИКОБАВА А.С. и др. (1967), "Языки народов СССР", Москва: Наука, т. 4.
- /3/ ПОНОМАРЕНКО Л.А. (1990), "Типология фонологических систем английского, русского, украинского и иберийско-кавказских языков", Английский язык. Научно-методические материалы. Иттомир: ИВУРЭ. — Вып. 2, с. 4 — 6.
- /4/ ШВАЧКО К.К., ТЕРЕНТЬЕВ П.В., ЯНУКЯН Т.Г., ШВАЧКО С.А. (1977), "Введение в сравнительную типологию английского, русского и украинского языков", Киев: Вища школа.
- /5/ SCHUCHARDT H. (1895), "Über das Georgische", Wien.

THE TYPOLOGY OF SPEECH SEGMENT UNITS

R.K. Potapova

Moscow State Linguistic University, USSR

ABSTRACT

The study of speech communication presupposes a preliminary segmentation involving the identification of cues segments and their characteristics. The task of the present investigation consisted in the determination and comparative description of different speech segments correlated in duration in spoken utterances for German, English, Swedish and Danish.

1. INTRODUCTION

The organization of spoken connected speech implies the selective extraction of linguistic objects which is impossible without segmentation. There are three principal difficulties arising in the process of solving a task of speech recognition both in the context of natural speech communication and in the application of automatic recognition devices: optimum segmentation of an object which is very difficult for the units to be recognized have no clear-cut boundaries; accounting for the variability factor of object characteristics (for instance, variability of articulation and acoustic characteristics of a speaker); identification of a set of key characteristics. The orientation to -

wards key segments enables us to differentiate between micro-, medi- and macrosegmentation. In microsegmentation key segment we include audial and acoustic intrasound/intersound transient processes, occlusions, frictions, explosions, subsonic and sound segments, syllabic segments. The following can be cited as segmentation universals at the audial level: segmentation at the neuron level with the orientation towards changes on the domain of F_0 ; reaction on the maximum values of spectrum energy and changes in timing of speech signal energy. Certain acoustical characteristics of the microsegmentation can also be defined as universals: presence/absence of F_0 ; instantaneous change of F_0 at the transition from a consonant to a vowel and from a vowel to a consonant; presence/absence of spectrum noise; noise localisation at the frequency scale; noise intensity; noise duration; presence of low and/or high frequency spectrum energy. The objects of macrosegmentation are phrases, sentences and fragments of spoken text. Between the objects of micro- and macrosegmentation is the phonetic word (the accentual group with proclitic and enclitic syl-

lables), which makes it possible to separate missegmentation, where we are guided by the following characteristics: specific realization of prosodic and spectral characteristics of juncture sounds; time correlation between segments within phonetic word; integral intensity of phonetic word; qualitative and quantitative characteristics of stressed/unstressed vowels within phonetic word. In implementing the macrosegmentation of speech prosodic characteristics are the principal ones as well as an account for syntactic and semantic information.

2. PROCEDURE

The investigation was based on the German, English, Swedish and Danish material. We have recorded 20 speakers male and female for each language. The experimental corpus contained a set of interlingual identical words which were included into a set of sentences having identical rhythmic and syntactic structures. All acoustical characteristics were extracted at the first stage by intonograph and at the second stage by MICRO SPEECH LAB-ver. 3. One aim of our investigation is to systematically describe and to define realizations of duration of different speech segments in connected text. To discover correlation for duration data of different speech segments the correlation coefficient $\bar{\rho}$ was determined:

$$\bar{\rho} = \frac{\sum_{i=1}^n x_i y_i - \bar{x} \bar{y}}{\sigma_x \sigma_y}$$

The H_0 -hypothesis of the independence of relation of data was tested by means of t-criterion with $p=5\%$. It was necessary to determine: a) the nature of the time

correlation between the adjoining subsonic segments within syllable of the following types: occlusion (separately for voiced/unvoiced segments), explosion, friction, aspiration, transition; b) the nature of the time correlation between adjoining sound segments within phonetic word; c) the nature of the time correlation between the syllables within phonetic word; d) the nature of the correlation in duration between unstressed and stressed vowels within sentence.

3. DISCUSSION

In the course of the study of speech behaviour it was discovered that the temporal and structural organization of speech movements calls for a complicated speech programme in the brain. It should be noted in this connection that at present the following problems are being studied in the domain of the temporal organization of spoken utterance: -the determination of synthesis algorithms of rhythmic pattern of the utterance (the problem which is often referred to as preprogramming of the temporal organization of speech signal being part of a more general problem of synthesis (generation) of prosodic patterns of an utterance; -the determination of speech signals used in the information exchange between linguistic and physiological levels of speech analysis and synthesis; -the determination of the rules for the conversion of transformation of the rhythmic patterns of an utterance into real time intervals between articulatory positions and movements [1]. As a result on the data pro-

crossing it was demonstrated that the level of subsegment allowed us to establish a specific temporal correlation between such segments as, for instance, a voiceless occlusion and following frication. In Swedish a tendency was observed for establishing a statistical relevant correlation between the duration of a voiceless occlusion of a tense occlusive consonant and the duration of the following frication segment. In Danish a regular correlation was traced between the duration of the frication segment of a tense occlusive consonant and the duration of the following and the preceding vowels [2]. It was demonstrated that the consonant and vowel segments are characterized by different degrees of a sufficiently reliable temporal correlation. Thus, for instance, it was proved that in German and English exists a negative temporal correlation between sound segments in the VC - sequence. The English language was characterized with the negative temporal correlation between the vowel and the following consonant in the syllable and phonetic word irrespective of its position in sentence. In German the correlation of the same type was revealed for the sequence "vowel-consonant", but it was mainly observed in the final position of sentence. It should be noted that this data were gotten for two-syllable words with a short stressed vowel. For the long stressed vowels in German there exists temporal correlation between a long vowel and a preceding consonant. No such correlation was observed between a long vowel and a following

consonant. The temporal correlation established for the "vowel-consonant" sequence underwent no substantial change dependent on position, which show a relative stability of the temporal relations between the sound segments which form closed syllables in the languages. The data obtained indicate that the pattern of temporal compensation at the microsegmentation level is not substantially modified even in those cases, when the sound - and subsegment sequence becomes part of a rhythmic structure of a higher linguistic order. This rule is typical for connected speech in German and English. In Swedish the situation is somewhat different, and the temporal correlation may be registered for sequences of sound segments "consonant-vowel", as well as for sound segments "vowel-consonant". This type of temporal correlation was observed not only in different words but also within the same word. In Danish the temporal correlation should be noted between sound segments only in the sequences "vowel-consonant". It follows from all that was said above

that the temporal relation at the level of microsegmentation within a syllable in connected speech is looser in some languages and stronger in other languages. This means, that the involvement of sound segments into a sequence of the speech continuum may substantially modify the type of the temporal relation between adjoining microsegments, or may have no significant impact leading to a temporal redistribution. In the former

case, the temporal relation between the constituents of a syllable is sufficiently mobile, which brings about new types of temporal correlation, while in the latter case, the temporal relation between the constituents is sufficiently strong, which preserves the relative integrity of the syllabic structure in the connected speech. This conclusion appears rather well-founded, - which is an evidence of natural flexibility of the syllable as a fundamental material quantum, on which the whole speech "building" places. For the microsegmentation positive temporal correlation between syllables of different type within a phonetic word is characteristic of some languages. The temporal correlation was observed between CVC and CV-segments within phonetic word. No correlation was observed between other types of syllable combinations: GV -(C)C.V within the same phonetic words in utterances. The comparison of duration data for vowel - sequences in sentences in case of macrosegmentation revealed a positive temporal correlation between them. The data obtained suggest that the duration of units of micro, - medi- and macrosegmentation in connected speech may be realized with different degrees of regularity and hierarchical character. It may be concluded that the architectonics of a speech utterance does not merely amount to a simple sum of duration data of a set of micro, - medi, - macrosegments, but emerges instead as a more complex structure, comprising

some relatively autonomous units, the temporal organization of which is predetermined by their own micro, - medi- and macrosegmental properties, as well as by the corresponding prosodic properties of the whole speech structure in general [3].

4. CONCLUSION

The variability of the prosodic organization of a utterance is brought about by the interaction of the following factors: the physiological one as predetermined by the constitution of human speech organs; the physiologically-linguistic one as predetermined by the laws of coarticulation in accordance with the features of the pronunciation basis of a given language; the linguistic one as predetermined by the phonemic, morphemic and syntactic rules of a language; the psycholinguistic one as predetermined by the communication act as a whole. The above mentioned factors are overlapping in the process of temporal programming and all types of segmentation of a spoken utterance.

5. REFERENCES

- [1] ČISTOVIĆ, L. (1972), "Problemy issledovanija vremenoj organizatsii reči" Leningrad.
- [2] POTAPOVA, R. (1986), "Slogovaja fonetika germanских jazykov", Moscow, Vysshaja škola.
- [3] POTAPOVA, R. (1990), "Strukturno-tipologičeskoje segmentovedenije zvučasnoj reči", Tipologija jazykov, Moscow, MGLU, 70-87.

PHONOLOGICAL COMPONENT
IN THE QUANTITATIVE LANGUAGE TYPOLOGY

L.G.Zubkova

Peoples' Friendship University, Moscow, USSR

ABSTRACT

As the language is a system connecting meaning and sound, the connection itself must be systemic. Hence, the complete typological characteristics of the language system should include information concerning the mode of correspondence between sound and meaning.

1. INTRODUCTION

The phonological typology of the basic language unit - the word - reflects its constitutive and paradigmatic relations, its functional-semantic and grammatical characteristics and as such correlates with the morphological typology [1,2]. This makes it possible in addition to morphological and syntactic indices of the quantitative typology suggested by J.Greenberg to introduce: 1)indices disclosing constitutive hierarchical relations of the units of different levels; 2)indices reflecting the degree of variability/uncertainty of the units of different levels (in their interrelation); 3)indices characterizing the sound shape of morphemes and words and not only their syllabic and suprasegmental organization (as suggested by V.Scalička), but their phonemic structure depending on functional and semantic characteristics as well. As different word groups may not have similar typological tendencies, all indices should be determined both for the word as a whole and for separate parts of speech.

2. CONSTITUTIVE RELATIONS' INDICES

In constitutive relations between language units the frequency of factual coincidence of the units of one level with the units of another one is of special typological significance. On the one hand, the frequency of one-word sentences, one-morpheme words, one-phoneme (one-syllable) morphemes essentially characterizes the plane of expression of each of the meaningful units. On the other, the frequency of words separately making up a sentence, morphemes making up a word, phonemes (syllables) making up a morpheme indicates the degree of autonomy of the lower-level units towards the higher-level ones. The higher is the frequency the less is the degree of autonomy. Taking into consideration the last index a phoneme must be regarded as the universal minimal functional sound unit. Even in languages with the high index of grammaticality in the majority of usages a phoneme functions as a purely phonological entity. In classical agglutinative languages the degree of autonomy of a phoneme increases and reaches its maximum value in isolating languages as being most "lexemic". The notion of a syllable seems irrelevant because of non-autonomy of a syllable in reference to a morpheme even in syllabic isolating languages.

3. VARIABILITY/UNCERTAINTY INDICES

3.1. On the Level of Words

It's been noted that typological differences in the degree of

extension of lexical polysemy correlate with differences in the degree of synthesis and the length of a word: simple and short words tend to have more lexico-semantic variants (meanings) than derived and long words. This tendency is apparent in the opposition between analytic and synthetic languages and in synthetic languages in the opposition between the underlying members of derivational chains and high-stage derivatives.

3.2. On the Level of Morphemes

The degree of semantic uncertainty (polysemy) of words and morphemes is inversely proportional to the degree of allomorphic variability. It is extremely limited in isolating languages and more or less developed in synthetic languages. In accordance with the principle of morphological structuring of a word and monosemy or polysemy of inflexional affixes in agglutinative languages allomorphic variability occurs mostly in affixes and is normally automatic, predictable, whereas in inflecting languages it occurs in roots and is mainly unpredictable. Meanwhile inflecting languages reveal clear parallelism between semantic and allomorphic variability between words of different stages in derivational chain. According to the data of Russian, polysemy and polymorphism of a root/stem in inflexional paradigm (fusion tendency) is more typical for the basic (non-derived) words. The higher the derivational stage, more limited are polysemy and allomorphic variability. High-stage derivatives usually possess one meaning and are characterized by monomorphism of a root/stem (agglutinative tendency).

3.3. On the Level of Phonemes

Due to a greater length in syllables and in phonemes and thus a greater occurrence probability of phonemes in positions of neutralization, high-stage derivatives differ from non-derived words by a greater phonematic uncertainty. It is not accidental

that the Russian scientific text with the degree of synthesis of a word equal to 3.21 morphemes and the average word length equal to 3.4 syllables contains 45% of weak phonemes, whereas in the colloquial speech with the degree of synthesis equal to 2.47 morphemes and the length equal to 2.8 syllables the frequency of weak phonemes decreases to 36%.

Apart from phonologically conditioned uncertainty of phonemic identification of sound segments in the analysis of variability on the level of phonemes the frequency of phonemes manifesting themselves as marked (derived) members of morphological alternations should also be taken into account.

This group of indices sides with indices characterizing the degree of preferable use of consonants in a definite position within morpheme/word and, respectively, the degree of differentiation between positions: within the morpheme, at morpheme juncture and at word juncture. The coefficients of rank correlation of consonants in comparable positions may serve as the abovementioned indices. Since the position of morpheme juncture (opposite to word juncture) reveals good positive correlation with within-the-morpheme position the degree of positions differentiation in consonantal structure of a simple word and in the root enables one to consider the type of the affixation used and its functional load. Positional differences are weakened in the following order: root-isolating languages, prefixing languages, suffixing languages, languages with developed bilateral affixation.

4. MEANINGFUL UNITS SOUND SHAPE INDICES

The sound shape of morphemes and words is indispensable to meaning. Fundamental typologically significant semantic difference is the difference between lexical and grammatical meanings.

Contrasting of lexical and grammatical is reflected in constitutive and distinctive functions of sound units and in the degree of phonematic uncertainty. For example, in the Russian speech the coincidence of a morph with a syllable is more frequent in prefixes and roots as more lexical units, the coincidence of a morph with a phoneme -in suffixes and inflexions as more grammatical ones. The phoneme is quite autonomous in respect to the autosemantic morpheme. With lessening of lexicality and increase of grammaticality of morphemes, phonemes become less autonomous and may not possess this quality, if the given type of a morpheme always or mostly expressed by one phoneme as, for instance, is the case with inflexions in Marathi and Arabic.

Phonemes constitute a morpheme not as an autonomous element but as an integral part of the word. In line with the degree of autosemanticity of a word the phoneme in the Russian language obtains maximum autonomy in reference to a substantive root, less autonomy in reference to a verbal root and still less autonomy in respect to a pronoun root. The distinctive properties of phonemes within a morpheme go in line with the main rules of segmental word structure. Since the middle part of the word as distinct from marginal positions is usually irrelevant to distributive restrictions, the distinctive possibilities of phonemes in middle-of-the-word morphemes prove to be more effective. For example, in the Russian speech the frequency of weak phonemes in roots and suffixes amounts to 33-34% and in prefixes and flexions - to 62 and 59%. The part-of-speech function of a word is also significant for distinctive function of phonemes. In notional parts of speech performing nominative function the frequency of weak phonemes is higher than in pronouns performing substitutive and demonstrative

functions.

Morphonological differences between parts of speech are also essential for sense discrimination. For example, in the Indonesian language the voiceless consonants/nasals interchange frequency in the root-initial position amounts to 81-85% in predicatives, and to 51-53% in nouns. The opposition between lexical and grammatical greatly influences the indices which characterize the sound shape of meaningful units.

4.1. Phonemes' Inventory

In autosemantic morphemes making up an open list all phonemes and their combinatory possibilities are realized on a larger scale. In syntactic morphemes making up a closed list the inventory of phonemes is restricted: the more so, the less the number of the given morphemes and more grammatical their meanings. For example, in Russian and English in derivational suffixes the number of generally used phonemes amounts to more than 80%, in flexions it lessens to 33% in Russian and to 18% in English. The degree of restriction as to the phonemic inventory of morphemes is different in different parts of speech. In particular, the number of phonemes constituting Russian derivational suffixes lessens in the consequence: nouns (78.5%) - adjectives - adverbs - verbs (31%).

4.2. Phonemes' Quality

More or less strong tendency to attach phonemes to certain meanings first and foremost manifests itself in preferable usage of vowels to express grammatical meanings, and consonants to express lexical meanings. The degree of lexicalization of consonants and the degree of grammaticalization of vowels as well as the degree of lexicality/grammaticality of morphemes themselves may be indirectly observed in the consonantal coefficient, which reflects the proportion of consonants and

vowels in accordance with their frequency of occurrence within different types of morphemes. In the Russian speech inflexions possess the minimum (0.43) and roots - the maximum (2.13) value of the given coefficient. The more lexical is the root the higher is the coefficient. Consequently, it is higher in noun roots than in verb roots.

In derivational morphemes, combining lexical and grammatical meanings, the proportion of consonants and vowels is more equal (consonantal coefficient in suffixes equals to 1.44, in prefixes to 1.17).

4.3. The Meaningful Units' Length in Phonemes and Syllables

It is already known that concrete meanings are expressed by longer than the abstract one elements. The syllabic or non-syllabic form of meaningful lower-level units is determined by their free or bound position within higher-level units and finally by degree of autosemanticity. In tendency grammatical morphemes and words are shorter than lexical ones. For example, in isolating Yoruba and in inflecting Russian the syntactic root is shorter than the autosemantic root (respectively 1.00 and 1.35 syllables in Yoruba, 0.9 and 1.28 - in Russian), the pronoun root is shorter than the notional one (1.12 and 1.42 in Yoruba, 0.7 and 1.4 in Russian), the verbal root is shorter than the noun root (1.13 and 1.57 in Yoruba, 1.1 and 1.5 in Russian).

4.4. Morph Junctures and Syllable Boundaries Interrelation

Strong coincidence of syllable and morpheme boundaries in syllabic (isolating) languages is determined by lexicality of morphemes and their possibility to manifest a word. In non-syllabic languages different types of morpheme junctures in many ways correspond to syllable boundaries as the mode of combination and variability of morphemes depend

on their meaning, position within a word, and on whether they are added to the stem or to the word as a whole. For example, in Russian in accordance with the agglutinative character of prefixes and fusional character of flexions the coincidence of morpheme and syllable division is more probable at the prefix-root juncture and very rare at root/suffix-flexion juncture.

4.5. Morphemes' Suprasegmental Characteristics

The frequency of morphemes (autosemantic morphemes in particular) marked by suprasegmental means seems also to reflect the degree of lexicality/grammaticality of the language. It is not by chance that such prominence can be observed more often in tone isolating languages which are most "lexemic".

5. CONCLUSION

The language system integrity and unity can be clearly seen in good or average sufficient correlation of monophonemic morphs' frequency and the frequency of morpheme junctures within a syllable with indices of lexicality/grammaticality (+0.918 and +0.775), agglutination/fusion (+0.898 and +0.837) and synthesis (+0.716 and +0.536) in 11 languages of different types. The lower is the lexicality index and thus the higher the index of grammaticality, the higher are the indices of synthesis and fusion, hence more oftener occur in the text monophonemic morphs and respectively more frequent are morph junctures within the syllable.

6. REFERENCES

- [1] Zubkova, L.G. (1987), "Aspects of the Sound Form of the Word", Proceedings XIth ICPhS, vol.2, Tallinn.
- [2] Zubkova, L.G. (1990), "Fonologičeskaja tipologija slova", Moscow.

NEUTRALISATION DES VOYELLES NASALES CHEZ DES ENFANTS D'ILE DE FRANCE

I. Malderez

U.F.R. Linguistique, Université Paris 7, France.

ABSTRACT

A tendency towards the neutralization of oppositions between /ã/ and /ɔ/, on the one hand, and /ã/ and /ɛ/ on the other hand, is observed in the speech of Ile-de-France and Parisian youngsters. This phenomenon is also reflected in children's spelling.

1. APERCU DES TRAVAUX ANTERIEURS.

La neutralisation de l'opposition /ɛ/ vs /ɛ/ au seul profit de /ɛ/, déjà mentionné en 1821 par le père DESGRANGES [2] et un siècle plus tard par H. BAUCHE [1], semble aujourd'hui en voie d'aboutissement en français standard. H. WALTER [10] et P. LEON [7] signalent la corrélation positive qui s'établit entre l'âge du locuteur et la distinction des deux voyelles nasales. L'étude de O. METTAS, menée auprès de locutrices parisiennes âgées de 18 à 35 ans, montre que le plus souvent /ɛ/ "se confond avec /ɛ/" [9].

A ce phénomène déjà ancien s'ajoutent deux nouvelles tendances de neutralisation. En effet depuis 1972 certains auteurs signalent des déplacements ou des chevauchements entre les voyelles /ɛ/ et /ã/ d'une part et entre /ã/ et /ɔ/ d'autre part. Pour O. METTAS, le "phonème /ã/ tend à se rapprocher de [ɔ] du parler neutre [...] Cette dernière réalisation est l'un des indices les plus

fréquents du sociolecte, dans cette génération". De même /ɛ/ est "identifié comme une réalisation de /A/, oral ou nasal" dans certains cas. H. WALTER signale aussi le cas d'un locuteur parisien prononçant /ã/ "avec un arrondissement qui peut compromettre la distinction de l'opposition /ã/ vs /ɔ/." I. FONAGY étudie le traitement des nasales chez sept jeunes parisiens [3]. "La tendance générale du déplacement de *in* vers *an* n'a pas empêché un locuteur et deux locutrices à s'opposer à ce courant, et de substituer dans certains cas /ɔ/ à /ã/." Pour un des locuteurs, le ɔ a été perçu "comme /ã/ par la moitié ou la plupart des auditeurs."

P. LEON attache à ce chevauchement une valeur sociolinguistique "de certain parler chic" [8]. I. FONAGY et G. BOULAKIA supposent aussi cette valeur stylistique. "On a l'impression que les variantes /ɛ/ qui s'approchent de /ã/ font 'plus jeune', 'plus branché', 'plus désinvolte'." [4].

Ce type de glissement est aussi réalisé dans la parole des enfants d'Ile de France.

"Tu aurais pu mettre que les chevaux, ils ont tous un nom (A.V., 10 ans).

- C'est pas vrai! Ils ont pas tous un an! (S.B., 9 ans) "

Nous avons mené une étude systématique dans deux écoles primaires rurales du sud de l'Oise où sont scolarisés des enfants âgés de 5 à 12 ans.

2. TEST DE PERCEPTION

2.1. Test

Nous avons enregistré quatre enfants de 8 ou 9 ans lisant un corpus présentant 6 triplets minimaux en position finale. Un test de perception à choix limité a été effectué auprès de l'ensemble de la classe où étaient scolarisés les quatre locuteurs. Ainsi nous avons calculé l'indice d'audibilité IA¹ de chaque locuteur sur chaque énoncé, chaque phonème et sur l'ensemble du test.

2.2. Résultats

La locutrice A.T.C. se distingue par 100% de IA égal à 1. Notons que sur l'enregistrement d'une saynète (style moins formel) nous avons constaté que A.T.C. produit des nasales ambiguës. Chez les locuteurs C.D. et V.H. les IA globaux sont respectivement de 0,93 et 0,94. Chez ces deux enfants, la neutralisation des voyelles est faible dans ce style formel. De même la conservation des oppositions n'est pas non plus effective à 100%. La moitié des énoncés chez C.D. et 81% chez V.H. ont des IA inférieurs à 1. Les indices tombent à 0,76 (C.D.) ou 0,80 (V.H.) pour certains énoncés. De plus, bien que les écarts restent faibles, ces locuteurs effectuent des traitements différents pour les trois phonèmes : les indices varient de 0,91 pour /ɔ/ à 0,95 pour /ɛ/ chez C.D. et de 0,93 pour /ã/ à 0,96 pour /ɛ/ chez V.H.. La locutrice P.M. obtient à peu près les mêmes résultats que les deux garçons pour les phonèmes /ɛ/ et /ɔ/. Par contre, le /ã/ dans la parole de P.M. est perçu [ɔ] 107 fois sur 131 dans le test, soit un IA de 0,18 pour ce phonème. Trois énoncés obtiennent un IA égal à 0,05.

¹IA: indice de perception en accord avec l'énoncé proposé au locuteur dans le corpus.

Ainsi chez P.M. le déplacement obéit à une règle pratiquement catégorique. Néanmoins, elle perçoit correctement les trois voyelles nasales lorsqu'elle participe aux tests de perception. Un énoncé portant sur le phonème /ã/ atteint un IA de 0,65.

2.3. Statuts des oppositions

Ce test permet d'observer que chez ces enfants les oppositions entre /ã/ et /ɔ/ sont les plus fragiles, celles entre /ɛ/ et /ɔ/ les plus stables. Par ailleurs les déplacements repérés sur d'autres enregistrements ne concernent que la paire /ã/, /ɔ/. Ces résultats² sont en partie en accord avec les travaux antérieurs. En effet, si des confusions de voyelles nasales sont réalisées dans la parole des enfants et des jeunes parisiens, il n'existe pas de loi générale établissant une neutralisation plus avancée qu'une autre, ni un sens privilégié de déplacement. "Le /ɛ/ s'approche de /ã/ dans certain cas, dans tel mot plus souvent que dans tels autres, dans la parole de certains locuteurs plus souvent que dans celle d'autres locuteurs" [3]. D'autre part il nous semble difficile de considérer cette perte d'opposition comme une variante stylistique. Les enfants qui se sont prêtés à cette étude appartiennent à des classes socio-professionnelles de type 'ouvrier' ou 'employé' et habitent à la campagne dans des villages de moins de 1500 habitants.

3. CE QUE REVELE L'ORTHOGRAPHE

3.1. Données

Nous avons été confronté à des fautes d'orthographe assez atypiques. M.R. écrit *jombon* pour "jambon"; P.M. écrit *on* pour "en" et inversement; P.C. *den* pour "dont"; L.A. *versans* pour

²Tableau 1

"versons"; G.C. avont pour "avant"; O.B. resombe pour "ressemble" et je mon nuis pour "je m'ennuie"; A.V. elle sont les odeurs pour "elle sent..." etc...

A la suite d'IFONAGY, on peut affirmer que "les fautes d'orthographe d'enfants de 6-7 ans reflètent souvent leur conception phonologique." [3; 5]. Nous pensons pouvoir étendre cette constatation à tout individu de plus de sept ans, enfant ou adulte, en difficulté vis à vis de l'écrit.

Les divers manuels de français issus des derniers programmes officiels ne proposent aucun travail quant à la discrimination orthographique des voyelles nasales pour les enfants de 8-12 ans. Cet état de fait souligne le caractère récent de la perte partielle des oppositions entre voyelles nasales. Les manuels s'attachent par contre à faire acquérir les différentes graphies in, ain, un. Pour les élèves du CE1³, on trouve quelques rares exercices concernant l'opposition /ã/ vs /ɔ/. [6].

3.2. Test

Nous avons proposé la série de trois exercices⁴ à 118 élèves (6-12 ans) d'une

3. Cours Élémentaire 1^{ère} année. (7 ans).

4. 1 Complète les mots avec an ou on.

Les hir...delles s...t mainten...t de retour. Dès les premiers ray...s de soleil, elles arrivent en volet...t, en ras...t le sol et en cri...t. Qu...d il fait froid, elles v...t en Afrique.

2 Choisis le mot qui convient.

1. Est-ce que les oiseaux ont une {langue, longue}?
2. Il y a des {rongées, rangées} d'hirondelles sur les fils.
3. Le faisan a brusquement disparu à {l'angle, l'ongle} du bois.

3 Complète les mots avec en ou on.

école en mai 1989, à 71 élèves du même établissement en mai 90, et enfin à 41 enfants (7-12 ans) d'une seconde école en septembre 90.

Nous avons constaté des confusions orthographiques de nasales dans tous les cours y compris chez les plus vieux. Notons que P.M. n'a fait aucune erreur au test; dans son cas l'orthographe est bien fixée et ne révèle pas les déplacements mis en évidence dans sa parole lors des tests de perception.

Sur 4208 réponses exprimées apparaissent 368 "erreurs"⁵, soit un taux global de réussite de 91%. Dans le 3^{ème} exercice, nous avons accepté renger, antrée, et monten. Selon les mots en référence, ce score varie de 76 ("angle") à 98% ("serpent"). Les lexèmes les plus usités n'atteignent pas systématiquement les meilleurs scores : "sont" 92, "quand" 87, "vont" 93. Par contre si on considère les onze lexèmes entrant dans une paire minimale ("vont" vs "vent"), la moyenne des scores est de 89% contre 94 pour les autres.

L'appartenance à une paire minimale favorise donc les confusions orthographiques de nasales.

4. REFERENCES

- [1] BAUCHE, H. (1946), *Le langage populaire*, [1920], Paris: Payot.
 [2] DESGRANGES, J.C.L.P. (1821), *Petit dictionnaire du peuple à l'usage des quatre cinquièmes de la France*, Paris: Chaumerot.

- | | |
|----------------|----------------|
| 1. un serp...t | 5. j...gler |
| 2. l'...trée | 6. ...f...cer |
| (3. r...ger) | 7. le m...t... |
| 4. gr...der | 8. ...vir... |

5. Tableau 2

[3] FONAGY, I. (1989), "Le français change de visage?", *Revue Romane*.

[4] FONAGY, I. & BOULAKIA, G. (1989), "Tendances de neutralisation des oppositions entre voyelles nasales dans la parole des jeunes parisiens", *Actes du ICPHS de Budapest*.

[5] FONAGY, I. & FONAGY, P. (1971), [Comment faire usage des fautes d'orthographe en hongrois?], *Magyar Nyelvőr*, 95, 70-89.

[6] GUEREAULT, D. & LEON, R. (1978), *Le français au CE1*, Paris : Hachette Ecoles.

[7] LEON, P.R. (1973), "Modèle standard et système vocalique du français populaire de jeunes parisiens", *Contributions canadiennes à la linguistique appliquée*, Rondeau, G. Ed, Montréal, 55-79.

[8] LEON, P.R. (1979), "Standardisation vs diversification dans la prononciation du français contemporain", *Current issues in the phonetic sciences*, Hollien, H. et P. Eds, Amsterdam: Benjamins, 541-549.

[9] METTAS, O. (1973), "Les réalisations vocaliques d'un sociolecte parisien", *Travaux de l'Institut de Phonétique de Strasbourg*, 5, 1-11

[10] WALTER, H. (1977), *La phonologie du français*, Paris: Presse Universitaire de France.

Tableau 2 : Pourcentages d'erreurs selon les mots en référence:

mots en référence	nbre. occu.	nbre. erreurs	% de erreurs
hirondelle	223	8	4
sont	225	17	8
maintenant	224	13	6
rayons	223	12	5
voletant	216	39	18
rasant	219	32	15
criant	222	28	13
quand	226	19	13
vont	224	16	7
langues	220	20	9
rangées	223	7	3
angle	222	54	24
serpent	229	4	2
entrée	227	13	6
gronder	228	10	4
jongler	228	16	7
enfoncer	210	21	10
menton	210	22	6
environ	209	15	7

Tableau 1 : Nombre d'occurrences et répartitions des perceptions:

sont perçus	ē			ɔ̃			ã		
	ē	ɔ̃	ã	ɔ̃	ã	ē	ã	ɔ̃	
C.D.	97	2	3	93	7	2	96	5	1
V.H.	111	0	4	108	6	1	128	3	7
P.M.	124	5	3	127	5	0	23	1	107

EVALUATION OF VOICE AND PRONUNCIATION CHARACTERISTICS OF MEN AND WOMEN

Mirjam T.J. Tielen and Florien J. Koopmans-van Beinum

Institute of Phonetic Sciences, University of Amsterdam,
Herengracht 338, 1016 CG Amsterdam, The Netherlands.

ABSTRACT

In this paper results are presented of an extensive listening experiment on the evaluation of male and female voice and pronunciation characteristics. Thirty male and thirty female speakers from three profession categories were recorded while reading aloud several texts. Three main research questions were involved: 1. Are voice and/or pronunciation of men and women evaluated differently? 2. Do listeners' judgments and subjects' opinions about these characteristics reveal similar results? 3. Can speakers from different professions be distinguished by voice and pronunciation cues only?

1 INTRODUCTION

Clear differences in acoustic characteristics exist between male and female voices. An interesting question in this connection is which perceptual characteristics would be related more to male voices and which to female voices. A study by Kramer [2] revealed that some perceptual characteristics were more associated with female speech (e.g. gentle, melodious), while other characteristics were associated more with male speech (authoritative, loud). In the present experiment, it was tested to what extent male and female voice and pronunciation would be evaluated differently. Judgments based on actual presentation of voices and subjects' opinions were compared. All evaluation scores were collected by means of semantic scales [3]. Another question was whether voices of different professions would be evaluated differently. If so, this would imply that listeners are able to distinguish voices with respect to

profession (see also [4]).

2 METHODS

2.1 Speakers and listeners

Thirty male and thirty female representatives from three profession categories (nurses, managers and information agents) were selected. These particular speaker groups have been chosen, because these groups differ clearly from one another with respect to the number of men and women working in these professions, and because speech is an important aspect of the work in all three categories. Twenty male and twenty female students of the University of Amsterdam (all native speakers of Dutch) participated as listeners in the experiment.

2.2 Stimuli and Recordings

The speakers were asked to read aloud text passages taken from actual speech. Three texts dealt with topics associated with the three professions. An additional text was included which dealt with a neutral topic with respect to the speaker groups. The sixty speakers were recorded at various places. They were allowed to prepare the texts in the way they desired; also, they were free in choosing their own tempo and intonation. At the end of the recordings, the speakers gave their opinion about their own voice and pronunciation by means of the already mentioned rating instrument (results in 3.3).

2.3 Perceptual evaluation

Voice and pronunciation were evaluated by means of the semantic 'twin scales' (seven pairs of related notions) as developed for Dutch by Fagel et al. [1].

In addition, four new scales were included that were supposed to differentiate between the profession categories, if the professions would differentiate at all. The English translations of the Dutch scales as used in the present experiment are shown in Table 1. The four scales at the bottom of the table are the additional ones.

2.4 Experimental procedure

The listeners were instructed to evaluate the voices by means of the eighteen scales (results in 3.1). The seven texts as well as the sixty speakers were randomized. The listening sessions were performed at the Language Centre of the University of Amsterdam, where listening facilities were available that allowed for selective presentation to listeners individually and simultaneously. During the listening sessions, the forty listeners were also asked to identify the profession of the speakers presented by choosing from a list of six possibilities (results in 3.4).

Apart from evaluating the voices, the listeners were also asked to give their opinion about typical voice and pronunciation characteristics of men and women in six profession categories, including the three already mentioned (results in 3.2). This task was also performed by means of the same rating instrument.

3 RESULTS

3.1 Evaluation of speakers' voice and pronunciation

Mean scale scores were derived for all three speaker groups for men and women separately (see Table 1). It appeared that the differences between male and female voice and pronunciation are not very large, although except for scales no. 1,9,11,12 and 18, the scales appeared to be significantly different (t-test; sign. level was set to 0.003, because of the repetition of t-tests). Not surprisingly, the largest differences are found for the scales 'high-low' and 'shrill-deep'. Smaller differences are found for the scale 'dull-clear', with female voices considered to sound clearer than male voices. Furthermore, female nurses and managers were evaluated as more monotonous than male

nurses and managers. Some scales differentiated the profession categories. Managers were evaluated as speaking in a little more polished and cultured way. Factor analyses were performed on the correlations between the scales in order to look for the underlying patterns of relationships between the data. Analyses performed on the male and female data separately, revealed that four factors explained about 50% of the total variance in rather well interpretable dimensions that can be characterized as 'Appreciation quality of the voice', 'Personality evaluation', 'Pronunciation quality' and 'Pitch' respectively, with some minor differences between the two sexes. The scales 'broad-cultured' and 'pleasant-unpleasant' attained the highest communality estimates, which implies that these scales are responsible for a considerable part of the variance.

3.2 Opinions about typical voice and pronunciation characteristics

The mean scale scores obtained by the non-auditively based judgments reveal that there a smaller number of assumed differences between the voices of the two sexes exist in comparison to the voices of the three profession categories. Only the scales 'high-low' and 'shrill-deep' differed significantly for male and female voices.

Most extreme scores were found for the scales 'slovenly-polished' and the related scale 'broad-cultured'. Managers (male as well as female) were supposed to possess the highest degree of culture and polishment in their pronunciation. Also, rather extreme mean scores are found for attributes like powerful, authoritative and business-like with respect to managers. One more appealing finding was the high mean score for male information agents with respect to melodiousness.

3.3 Evaluation of voice and pronunciation by the speakers themselves

The mean scores as given by each of the speakers with regard to their own voice and pronunciation also tended to the centre of the interval. Nevertheless, female information agents judged their pronunciation as more polished and cultured than the other speaker groups.

Table 1.
Mean scale scores for female (F) and male (M) nurses (20), managers (20), and information agents (20) as evaluated by all (40) listeners over all text presentations. In the last two columns the overall means for female and male speakers are presented. The 18 scales used were seven-point interval scales.

SCALES		NURSES		MANAGERS		INF.AGENTS		ALL	
		F	M	F	M	F	M	F	M
1. slovenly	polished	4.58	4.39	4.87	4.75	4.35	4.63	4.60	4.59
2. broad	cultured	4.28	4.42	4.67	4.59	4.08	4.85	4.34	4.55
3. high	low	3.76	4.98	3.90	5.05	3.50	4.85	3.72	4.89
4. shrill	deep	3.88	4.80	3.99	4.90	3.47	4.49	3.78	4.73
5. dragging	brisk	4.34	3.89	4.36	3.97	4.19	4.85	4.30	4.17
6. slow	quick	4.14	3.56	4.10	3.70	4.06	4.33	4.10	3.86
7. husky	not husky	4.13	4.37	4.24	4.28	4.31	4.47	4.23	4.37
8. dull	clear	4.59	4.05	4.78	4.06	4.75	4.30	4.71	4.14
9. weak	powerful	4.25	4.21	4.56	4.38	4.32	4.48	4.38	4.36
10. soft	loud	4.14	3.97	4.31	3.94	4.51	4.18	4.32	4.03
11. ugly	beautiful	4.23	4.02	4.35	4.19	3.72	4.15	4.10	4.12
12. unpleasant	pleasant	4.49	4.29	4.63	4.44	3.96	4.43	4.36	4.39
13. monotonous	melodious	4.40	3.95	4.66	4.09	4.34	4.25	4.47	4.10
14. expressionless	expressive	4.37	3.91	4.68	4.06	4.37	4.22	4.47	4.07
15. severe	sweet	4.45	4.14	3.96	3.93	4.07	3.96	4.16	4.01
16. business-like	emotional	4.07	3.74	3.59	3.53	3.99	3.57	3.88	3.61
17. vulgar	distinguished	3.87	4.06	4.31	4.40	3.80	4.16	3.99	4.21
18. timid	authoritative	3.86	3.97	4.40	4.17	4.14	4.23	4.13	4.12

Female managers and information agents scored their voices as more expressive than the other speaker groups. Male speakers regarded their own voices less husky than the female speakers did.

3.4 Identification of profession by voice cues alone

Most listeners reported that identification of the profession of a speaker was a difficult task. The alternative options of 'shop assistant' and 'teacher' (categories that were actually not present in the speaker groups) appeared to be options that were rather frequently chosen. It appeared that female nurses were iden-

tified correctly more often than male nurses. Moreover, female nurses were scarcely confused with female managers in contrast to male nurses with male managers. Male information agents were classified most often as teachers, whereas female information agents received most scores on the categories of nurse and shop assistant. The significance of the different scores was tested in an analysis of variance. The Anova analysis (mixed model with repeated measures on the factors

Table 2.
Results of the analysis of variance (mixed model; repeated measures) on the correct profession identification scores. Sl=Sex of listener; Ssp=Sex of speaker; Psp=Profession of speaker.

FACTOR	DF	SS	F RATIO	SIGN. OF F
Sl	1, 38	10.37	1.48	.23
Ssp	1, 38	27.26	5.94	.02
Psp	2, 76	143.52	11.49	.00
Ssp * Psp	2, 76	136.41	13.87	.00
Sl * Ssp * Psp	2, 76	4.25	.43	.65

'speaker sex' and 'profession of speaker') gave rise to the results as presented in Table 2. From that table it can be seen that sex of speaker was one of the differentiating factors.

Also, speakers with different professions were identified differently and the interaction between sex of speaker and profession of speaker also reached the level of significance, which implies that male and female nurses, managers and information agents received different scores. From the differences in percentages between the three speaker groups, it appeared that the manager scores are higher when manager voices had been presented. The same holds for the nurse scores. Although not tested in the Anova, the influence of text condition on the profession appointments appeared to be very clear as well. The scores on each of the categories were increased when the text content fitted with that particular profession.

4 DISCUSSION

Although considerable differences existed between the individual speakers of each group as evaluated by the listeners, in general the differences between male and female speakers as well as between managers, nurses and information agents appeared to be rather small. More differences were appointed with respect to male and female speakers if voice and pronunciation characteristics were evaluated without actual presentation of voices. These supposed characteristics were also more extreme. The finding that the differences are considered to be

larger than were actually found, may point to the existence of prejudices. The differences between 'high-low' and 'slow-quick' were enlarged for male and female managers. Differences in 'high-low', 'ugly-beautiful' and 'monotonous-melodious' were enlarged for male and female information agents. The differences between the three professions were even more enlarged; managers are supposed to speak e.g. very polished and with authority in comparison with the other groups. Evaluation by the speakers themselves revealed not such clear group differences. Identification of profession on the basis of someone's voice was far from perfect, but neither a random choice.

5 REFERENCES

- [1] Fagel, W.P.F., Herpt, L.W.A. van, and Boves, L. (1983), "Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation", *Speech Communication* 2, 315-326.
- [2] Kramer, Ch. (1977), "Perceptions of female and male speech", *Language and Speech* 20, 151-161.
- [3] Osgood, C.E. and Suci, G.J. (1955), "Factor analysis of meaning", *J. of Exp. Psychology* 50, 325-338.
- [4] Tielen, M.T.J. (1990), "Perception of the voices of men and women in relation to their profession", *Proc. ESCA Workshop on speaker characterization*, 192-197.

SOCIAL DISTRIBUTION OF LONG-TERM AVERAGE SPECTRAL CHARACTERISTICS IN VANCOUVER ENGLISH

J. H. Esling**, B. Harmegnies* and V. Delplanq*

**University of Victoria, Canada

*Université de Mons, Belgique

ABSTRACT

This project examines differentiation in voice quality setting across varieties of English in urban Vancouver, Canada. Two different techniques of long-term average spectral (LTAS) analysis are used to investigate extended samples of continuous text across groups collected in a sociolinguistic survey. The first procedure uses smoothed spectra and canonical discriminant analysis. The second procedure uses nonsmoothed spectra and the SDDD dissimilarity measure. Results thus far suggest that overall SES effects are greater than aging effects, particularly for the MMC group.

1. THE VANCOUVER SURVEY

Data are drawn from the Survey of Vancouver English, conducted in 1979-80 [4] [5]. Analysis focuses on 192 randomly-selected male and female English speakers native to the Vancouver region, in the three age groups of the survey: O (over 60), M (35-60) and Y (16-34). Four socioeconomic status (SES) categories, middle and upper working class (MWC/UWC) and lower and middle middle class (LMC/MMC), are compared. The sample text is drawn from a reading passage with local content.

2. INITIAL LTAS PROCEDURE

2.1. Method

In the initial LTAS procedure, frames below voicing threshold (silences and voicelessness) are discarded, and FFT power spectra of voiced speech with smoothing applied are integrated over successive nonoverlapping 20 ms windows of the first 60 s of the survey reading text [2]. LTAS distributions are

compared using principal component and canonical discriminant analysis to compute the Mahalanobis distance. Initial analysis focuses on the older and middle-aged groups.

The relationship of SES-group spectra with the LTAS of articulatorily modelled settings (by the first author) are expressed in generalized squared distance. Categories of the phonetic description of voice quality settings used to compare and evaluate LTAS results from the sample are those defined by Abercrombie [1] and Laver [10]. Sets of models for LTAS analysis based on these categories have been assessed by Nolan [11] and Harmegnies, Esling & Delplanq [9].

2.2 Results

Results of initial LTAS analyses are summarized in TABLES 1 and 2. The generalized squared distance measure of intragroup variability in LTAS indicates that SES groups are more homogeneous for female subjects than for male subjects in the survey in general. For middle-aged males, only the LMC and MMC groups are differentiated by the LTAS analysis of voiced speech used in this study, corroborating the significant separation between LMC-MMC men found using vowel formant data.

Classification of each model setting by group is inconclusive. The association of the velarized setting with UWC males is tentative.

Middle-aged MWC and UWC women contrast in LTAS with middle-aged MMC women, in conformity with the vowel-formant distributions separating these groups. Other LTAS relationships, however, do not confirm over the long term the more vowel-specific findings of formant analyses [3]. LTAS peak

TABLE 1. LTAS relationships between middle-aged and older male groups. Canonical Discriminant Analysis; Probability > Mahalanobis Distance. (Figures in bold represent groups which are significantly separated.)

		Older				Middle-aged		
		MWC	UWC	LMC	MMC	UWC	LMC	MMC
Middle-aged	MWC	0.76	0.59	0.46	0.38	0.07	0.33	0.45
	UWC	0.12	0.39	0.01	0.31		0.35	0.004
	LMC	0.15	0.84	0.02	0.12			0.01
	MMC	0.42	0.06	0.94	0.23			
Older	MWC		0.31	0.40	0.65			
	UWC			0.12	0.37			
	LMC				0.39			

TABLE 2. LTAS relationships between middle-aged and older female groups. Canonical Discriminant Analysis; Probability > Mahalanobis Distance. (Figures in bold represent groups which are significantly separated.)

		Older				Middle-aged		
		MWC	UWC	LMC	MMC	UWC	LMC	MMC
Middle-aged	MWC	0.09	0.33	0.05	0.002	0.36	0.52	0.13
	UWC	0.01	0.27	0.04	0.001		0.46	0.02
	LMC	0.20	0.86	0.42	0.03			0.43
	MMC	0.84	0.28	0.31	0.34			
Older	MWC		0.19	0.25	0.47			
	UWC			0.74	0.04			
	LMC				0.20			

locations suggest a similarity between older MWC women and middle-aged MMC women. Middle-aged MWC women, on the other hand, contrast with the LTAS pattern of both older MWC and MMC speakers, as well as with the middle-aged MMC distribution. Older MWC women, where the formant shift suggests tongue retraction as for uvularization or pharyngalization, are classified together with the faucalized, uvularized and velarized models in descending order of probability.

3. SDDD LTAS PROCEDURE

3.1. Method

The second LTAS analysis procedure adopts the SDDD statistical techniques introduced by Harmegnies & Landercy [6] [7] [8] to improve the reliability of interspeaker LTAS comparisons.

The 192 speakers in the survey were split into categories according to three main variables: sex (male, female), age (younger, middle, older) and socioeconomic status (middle working class, upper working class, lower middle class, middle middle class). A full factorial partitioning model was used, and therefore resulted in 24 subsamples (2 sexes x 3 ages x 4 SES categories) of 8 speakers each.

Interspeaker comparisons of the speakers' LTAS were performed by means of the SDDD dissimilarity index. This was considered the dependent variable of the study.

Age and SES were, in turn, each considered as the independent variable of the study. Both these analyses were performed separately in the male and in the female group.

TABLE 3. Average SDDD values for intra- and inter-age comparisons (96 male subjects)

		AGE		
		O	M	Y
AGE	O	5.4	5.7	5.8
	M		5.7	5.8
	Y			6.0

TABLE 4. Average SDDD values for intra- and inter-age comparisons (96 female subjects)

		AGE		
		O	M	Y
AGE	O	6.3	5.9	6.4
	M		5.4	5.7
	Y			6.0

3.2 Data Analysis - Age Effects

Both inter- and intra-age-class comparisons are considered. Each comparison of one class to another involves 496 interspectral dissimilarity measures. Each figure in TABLES 3 and 4 is therefore the average of 496 values.

The intraclass dissimilarity is, on the average, slightly less than the interclass on (males: 5.69 < 5.78; females: 5.89 < 6.04), suggesting overall weak aging effects: the LTAS drawn from a given age group tend to be more similar to one another than they are to spectra drawn from other age groups.

The greater intraclass homogeneity is nevertheless most sensitive for middle-aged subjects (both males and females), and older males: in those cases, the intraclass value is less than all the interclass values involving the considered class.

3.3 Data Analysis - SES Effects

Both inter- and intra-SES-class comparisons are considered. Each comparison of one class to another involves 276 interspectral comparisons. Each figure in TABLES 5 and 6 is therefore the average of 276 values.

The intraclass dissimilarity tends, on the whole, to be less than the interclass one (males: 5.65 < 5.95; females: 5.69 <

6.14), suggesting overall SES effects greater than the aging effects.

The relationship is particularly strong in the MMC both for males and females, and also for UWC females.

4. DISCUSSION

LTAS distributions are found to contrast across social varieties of Vancouver English for some SES groups, particularly for middle-aged female UWC and MMC speakers. These results suggest that long-term voice quality settings differ systematically between the respective social groups, at least in acquired oral-reading style.

These descriptive relationships will be subject to further examination using inferential analysis to determine probabilities. The results and methodologies of both the SDDD dissimilarity measure and the initial LTAS procedure will also be compared in relation to group-by-group vowel formant class analyses [3] for age and SES classes.

5. REFERENCES

- [1] ABERCROMBIE, D. (1967), "Elements of general phonetics", Edinburgh: Edinburgh University Press.
- [2] ESLING, J. H. (1987), "Vowel shift and long-term average spectra in the Survey of Vancouver English",

TABLE 5. Average SDDD values for intra- and inter-SES-class comparisons (96 male subjects)

		SES CLASS			
		MWC	UWC	LMC	MMC
SES CLASS	MWC	5.7	5.7	5.8	5.9
	UWC		5.9	6.2	6.4
	LMC			5.9	5.7
	MMC				5.3

TABLE 6. Average SDDD values for intra- and inter-SES-class comparisons (96 female subjects)

		SES CLASS			
		MWC	UWC	LMC	MMC
SES CLASS	MWC	6.0	6.4	6.2	5.8
	UWC		6.0	6.1	6.4
	LMC			6.0	6.0
	MMC				4.7

Proceedings of the XIth International Congress of Phonetic Sciences, 4, 243-246. Tallinn: Academy of Sciences of the Estonian SSR.

[3] ESLING, J. H. (1989), "Analysis of vowel systems and voice setting in the Survey of Vancouver English", Final report to the Social Sciences and Humanities Research Council of Canada, Research grant #410-87-0334.

[4] GREGG, R. J., MURDOCH, M., HASEBE-LUDDT, E. & DE WOLF, G. (1981), "An urban dialect survey of the English spoken in Vancouver". *Papers from the Fourth International Conference on Methods in Dialectology* (H. J. Warkentyne, ed.), 41-65, Victoria: University of Victoria.

[5] GREGG, R. J., MURDOCH, M., DE WOLF, G. & HASEBE-LUDDT, E. (1985), "The Vancouver survey: Analysis and measurement". *Papers from the Fifth International Conference on Methods in Dialectology* (H. J. Warkentyne, ed.), 179-200, Victoria: University of Victoria.

[6] HARMEGNIES, B. (1988), "Contribution à la caractérisation de la qualité vocale", Doctoral dissertation, Université de Mons.

[7] HARMEGNIES, B. & LANDERCY, A. (1985), "Language features in the long-term average spectrum", *Revue de Phonétique Appliquée, 73-74-75, 69-79.*

[8] HARMEGNIES, B. & LANDERCY, A. (1986), "Comparison of spectral similarity indices for speaker recognition", *Proceedings of the 12th International Congress on Acoustics, 1, A1-4, Toronto.*

[9] HARMEGNIES, B., ESLING, J. H. & DELPLANCQ, V. (1989), "Quantitative study of the effects of setting changes on the LTAS", *Eurospeech '89: European Conference on Speech Communication and Technology* (J. P. Tubach & J. J. Mariani, eds.), 2, 139-142. Edinburgh: CEP Consultants.

[10] LAVER, J. (1980), "The phonetic description of voice quality", Cambridge: Cambridge University Press.

[11] NOLAN, F. (1983), "The phonetic bases of speaker recognition", Cambridge: Cambridge University Press.

Note

This research was supported by grants from the Social Sciences and Humanities Research Council of Canada, numbers 410-87-0334 and 410-89-0191.

OBSERVATIONS SUR LA CHUTE DU L DANS LE FRANÇAIS DE NORTH BAY (ONTARIO)

Jeff Tennant

University of Toronto, Canada.

ABSTRACT

L-deletion in French clitic pronouns and definite articles is studied here using a corpus of adolescent speakers from North Bay, Ontario, a town with a 17% minority of Francophones. The variable is found to be socially stratified, speakers of working-class background generally showing higher rates of L-deletion. Speakers with lower French language dominance tend in most cases to delete fewer Ls in object clitics and articles. Male and female speakers delete Ls in equal proportions, except in "elle", where males delete fewer.

0. INTRODUCTION

On se propose ici d'étudier l'effacement du /l/ des pronoms clitiques et des articles définis dans le français parlé par les adolescents de la minorité francophone d'une ville du Nord de l'Ontario, North Bay. Tout en analysant le jeu des facteurs sociaux et linguistiques traditionnellement considérés dans les enquêtes sociophonétiques, on s'interrogera sur le rôle que pourrait jouer le degré de dominance linguistique, dont la variation est considérable entre les individus de cette communauté franco-ontarienne, qui constitue 17% de la population de North Bay.

1. TRAVAUX ANTÉRIEURS

Le problème de la suppression du /l/ dans ce contexte morphologique a attiré l'attention de nombreux chercheurs, dont Bougatiéff & Cardinal [3], Laliberté [4], Léon & Tennant [5] et Thomas [11]. Les études empiriques sur la question, notamment Poplack & Walker [8], reprenant Sankoff &

Cedergren [9] et Santerre, Noisoux & Ostiguy [10], pour Montréal, ainsi qu'Ashby [1], pour Tours, démontrent que l'élision du /l/ des morphèmes grammaticaux est plus avancée en français canadien qu'en français européen, et qu'elle semble être socialement stratifiée.

On se propose d'introduire dans l'étude de la chute du /l/ une autre variable sociale, celle de la dominance linguistique, dont l'importance a déjà été bien démontrée par Beniak & Mougeon [2] en ce qui concerne la variation morpho-syntaxique et lexicale chez les Franco-ontariens. Chez les moins francodominants des adolescents de cette communauté, étant donné leur usage très restreint du français en dehors du milieu normatif que constitue l'école, on pourrait s'attendre à trouver un taux moins élevé d'élisions du /l/, un taux élevé d'élisions étant -- les études citées ci-dessus le prouvent -- un trait saillant du vernaculaire franco-canadien.

2. CORPUS ET MÉTHODE

Le corpus utilisé pour cette enquête a été gracieusement fourni par Raymond Mougeon. Il s'agit d'entrevues d'en moyenne trois quarts d'heure conduites en champ libre avec 36 locuteurs dans une école secondaire française de North Bay. Les sujets sont tous des élèves de cet établissement, âgés entre 15 et 18 ans. Les proportions des locuteurs masculins (19) et féminins (17) sont à peu près égales. Les sujets sont groupés dans trois catégories de classe sociale, selon le métier des parents. On trouve une description détaillée de ce corpus dans Mougeon et al. [7]. Un indice

de dominance linguistique basé sur la fréquence d'emploi du français a été calculé pour chaque locuteur. Pour plus de précisions sur cet indice, voir Mougeon & Beniak [6]. On a analysé environ 20 minutes de parole pour chaque sujet, en éliminant le début de l'interview et en choisissant au hasard des intervalles de 4 ou 5 minutes. Toutes les occurrences de la variable ont été codées dans une transcription orthographique saisie sur ordinateur, selon la nature du morphème en question et sa réalisation (LO pour effacement du /l/; L1 pour prononciation du /l/), ainsi que selon le contexte phonologique. Ce codage a permis la quantification des données à l'aide d'un logiciel de concordance textuelle, WordCruncher. Les scores ainsi obtenus pour chaque locuteur ont été ensuite regroupés selon les catégories de sexe, classe sociale (3 groupes: I Supérieure, II Moyenne, III Ouvrière), et dominance linguistique (3 groupes de 12 sujets selon l'indice: G1 .01-.49; G2 .50-.70; G3 .71-1.00).

3. RESULTATS ET DISCUSSION

3.1. Résultats globaux

Le tableau 1 montre les taux d'effacement du /l/ des pronoms et des articles dans le français de North Bay, ainsi que les résultats obtenus par Poplack & Walker [8] pour Ottawa-Hull. On constate d'emblée certaines ressemblances. Le taux d'effacement du "il" personnel est moins élevé que celui du "il" impersonnel, pour lequel le /l/ tombe de façon presque catégorique. Cela confirmerait à nouveau l'hypothèse de Sankoff & Cedergren [9] selon laquelle la quantité d'information contenue dans le morphème serait un facteur dans l'effacement du /l/. La seconde constatation que l'on peut faire, c'est que le patron observé par Poplack & Walker [8] concernant l'ordre des morphèmes par fréquence de syncope, est reproduit à North Bay: pronom sujet > pronom objet > article.

Ces résultats semblent au premier abord indiquer que la suppression du /l/ est moins fréquente à North Bay qu'à Ottawa. Cependant, vu la situation de dis-

Tableau 1.

Effacement du /l/ dans le français de North Bay, et d'Ottawa-Hull (selon Poplack & Walker [8])

	North Bay			Ottawa-Hull	
	LO	N	%		%
PRO SUJ					
il (pers)	378	416	90.9%	100.0%	
il (imp)	369	374	98.7%	100.0%	
ils	491	530	92.6%	99.0%	
elle	91	136	66.9%	84.0%	
elles	0	0	...	33.0%	
PRO OBJ					
lui	138	15	86.7%	91.0%	
les	17	55	30.9%	50.0%	
leur	6	21	28.6%	4.0%	
la	2	14	14.3%	32.0%	
le	11	117	9.4%	8.0%	
l'	4	116	3.4%	5.0%	
TOTAL	53	338	15.7%	28.2%	
ART DEF					
la	142	560	25.4%	38.0%	
les	76	510	14.9%	17.0%	
le	55	558	9.9%	7.0%	
l'	27	373	7.2%	18.0%	
TOTAL	300	2001	15.0%	19.0%	

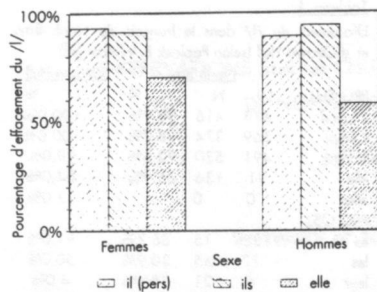
cours plus informelle dans laquelle Poplack & Walker [8] ont recueilli leur corpus, il se peut qu'il s'agisse là de divergences stylistiques, au lieu d'une véritable différence entre les réalisations de la variable dans les deux communautés.

3.2. Facteurs Sociaux

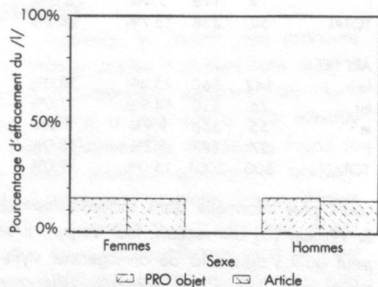
3.2.1. Sexe

Le graphique 1 montre les taux d'élision du /l/ des pronoms sujets "il" (personnel), "ils" (masc. et fém. dans ce dialecte), et "elle", selon le sexe des locuteurs. On a éliminé le pronom "il" (impersonnel) de l'analyse de la variation sociolinguistique, l'élision du /l/ étant presque catégorique pour ce morphème. On constate que les différences sont très minimes pour les pronoms masculins: "il" (femmes: 93,3%; hommes: 89,1%), et "ils" (femmes: 91,3%; hommes: 94,1%). Quant au pronom féminin "elle", on remarque un certain conservatisme chez les hommes (femmes: 70,4%; hommes: 57,9%), ce qui va à l'encontre de la tendance observée dans d'autres études.

L'élision dans les pronoms objets (femmes: 15,4%; hommes: 15,9%) et les articles définis (femmes: 15,4%; hommes: 14,6%), comme on peut le voir dans le graphique 2, semble n'avoir aucune



Graphique 1. Effacement du /l/ des pronoms sujets selon le sexe des locuteurs.

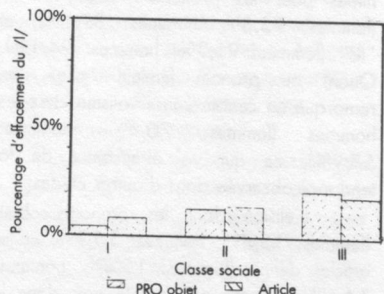


Graphique 2. Effacement du /l/ des pronoms objets et des articles selon le sexe des locuteurs.

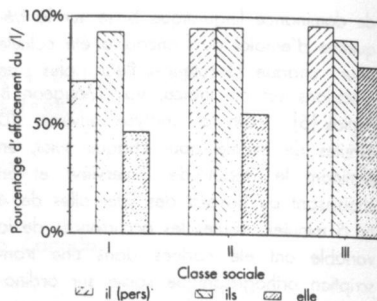
corrélation avec le sexe du locuteur dans notre corpus.

3.2.2. Classe sociale

Les données présentées dans les graphiques 3 et 4 confirment en grande partie la tendance observée dans d'autres géolectes du français en ce qui concerne la stratification sociale de la variable, la classe ouvrière faisant plus d'élisions que la classe supérieure: "il" personnel (I: 69,9%; II: 94,3%; III: 96,2%), "elle" (I: 46,7%; II: 55,3%; III: 78,4%), pronoms objets (I:



Graphique 3. Effacement du /l/ des pronoms sujets selon la classe sociale des locuteurs.

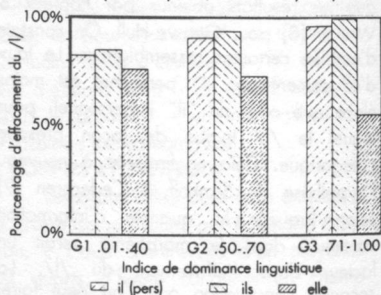


Graphique 4. Effacement du /l/ des pronoms objets et des articles selon la classe sociale des locuteurs.

4,0%; II: 12,4%; III: 21,1%), et articles (I: 8,1%; II: 13,9%; III: 18,6%). Le pronom "ils" semble cependant faire exception, mais les différences entre les groupes sont peu importantes (I: 92,5%; II: 94,8%; III: 90,4%). A North Bay comme ailleurs, il semble s'agir d'un marqueur sociolinguistique stable.

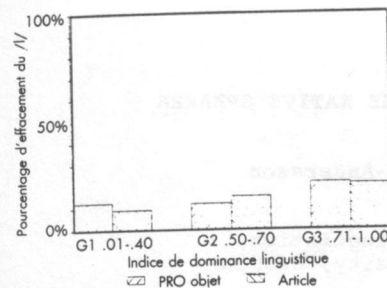
3.2.3. Dominance linguistique

Les graphiques 5 et 6 illustrent le rapport entre la chute du /l/ et l'indice de dominance linguistique. Pour "il" personnel (G1: 82,5%; G2: 90,7%; G3: 96,2%) et "ils" (G1: 85,2%; G2: 93,8; G3: 99,4%), ainsi que pour les pronoms objets (G1: 12,4%; G2: 11,6%; G3: 21,2%) et les articles (G1: 9,0%; G2: 15,0%; G3:



Graphique 5. Effacement du /l/ des pronoms sujets selon l'indice de dominance linguistique.

20,4%), le taux d'élisions est plus élevé chez les plus francodominants que chez les plus anglo-dominants. Cela semble confirmer l'hypothèse que nous avons émise ci-dessus, à savoir que les locuteurs dont le français est la langue la plus utilisée se rapprochent davantage de la norme statistique du vernaculaire. Cependant, les scores pour le



Graphique 6. Effacement du /l/ des pronoms objets et des articles selon l'indice de dominance linguistique.

pronom "elle" (G1: 76,1%; G2: 73,0%; G3: 54,7%) contredit cette observation. Cette contradiction serait peut-être attribuable à la grande variation interindividuelle que l'on a pu observer, les moyennes des réalisations à l'intérieur de chaque sous-groupe étant accompagnées d'écart types très élevés.

3.3. Facteurs phonétiques

Pour le pronom "elle", le contexte /_C est très favorable à la chute du /l/ (84,6% de suppressions) par rapport au contexte /_V (59,6%).

Quant aux pronoms objets et aux articles, le /l/ est effacé plus souvent derrière voyelle (pronoms: 19,6%; articles 21,0%) que derrière consonne (pronoms: 7,0%; articles: 12,7%).

3.4. Facteurs morphosyntaxiques

Le /l/ des articles est effacé le plus souvent dans les groupes prépositionnels, notamment suivant "dans" (39,7%), "sur" (37,9%) et "à" (30,7%).

4. REMERCIEMENTS

Ce travail a pu être réalisé grâce à une bourse doctorale du Conseil de Recherches en Sciences Humaines du Canada. Je suis également reconnaissant envers Raymond Mougéon pour avoir fourni le corpus de North Bay. Enfin, à Susana Soler qui a tellement aidé dans le travail sur le corpus, mille fois merci!

5. RÉFÉRENCES

- [1] ASHBY, W. (1984), "The Elision of /l/ in French Clitic Pronouns and Articles", *Romanitas: Studies in Romance Linguistics*, Ann Arbor: Michigan Romance Studies, 1-16.

[2] BENIAK, É. & R. MOUGEON (1989), "Recherches sociolinguistiques sur la variabilité en français ontarien", in R. Mougéon & É. Beniak, dirs., *Le français canadien parlé hors Québec*, Québec: Presses de l'Université Laval.

[3] BOUGAIEFF, A. & P. CARDINAL (1980), "La chute du /l/ dans le français populaire de Québec", *La linguistique*, 16, 91-102. [4] LALIBERTÉ, T. (1974), "L'élision du l en français québécois", *Lingua*, 33, 115-122.

[5] LÉON, P. & J. TENNANT (1990), "'Bad French' and Nice Guys": A Morphophonetic Study, *The French Review*, 63, 5, 763-778.

[6] MOUGEON, R. & É. BENIAK (1990), "Linguistic Consequences of Language Contact and Restriction: The Case of French in Ontario, Canada". Oxford: Oxford University Press.

[7] MOUGEON, R et al. (1982), "Le français parlé en situation minoritaire", vol. 1., Québec: Centre international de recherche sur le bilinguisme.

[8] POPLACK, S. & D. WALKER (1986), "Going through (l) in Canadian French", in D. Sankoff, dir., *Diversity and diachrony*, Amsterdam-Philadelphie: John Benjamins, 173-198.

[9] SANKOFF, G. & H. CEDERGREEN (1971), "Les contraintes linguistiques et sociales de l'élision du L chez les Montréalais", in M. Boudreault & F. Moehren, dirs., *Proceedings of the XIII International Congress of Romance Linguistics and Philology*, Québec: Presses de l'Université Laval, 1101-1116.

[10] SANTERRE, L., D. NOISEUX & L. OSTIGUY (1977), "La chute du /l/ dans les articles et pronoms clitics en français québécois", *The Fourth LACUS Forum* 1977, 530-538.

[11] THOMAS, A. (1990), "Normes et usages phonétiques de l'élite francophone en France et en Ontario", *Information/Communication*, 11, 8-22.

FOREIGN ACCENT AND THE NATIVE SPEAKER

Una Cunningham-Andersson

Department of Linguistics,
Stockholm University, Sweden

ABSTRACT

This paper reports the results of an experiment investigating a) native speaker reactions to common non-native pronunciations on the three dimensions of perceived importance of error, perceived friendliness of speaker and perceived educational level of speaker, and b) the effect of information about speakers' ethnic origin on the reactions of native speakers from two socially distinct, but otherwise comparable groups to non-native pronunciations.

1. INTRODUCTION

It is important to distinguish between, on the one hand, attitudes to ethnic groups, which can evidently be elicited using more or less accented speakers as stimuli (e.g. the classical matched-guise work of Lambert et al., [1] and on the other hand, attitudes to the non-native accents associated with these groups. Here we are concerned with the relationship between the way native speakers of Swedish see different immigrant groups and the way they react to different phonetic features of immigrants' pronunciation of Swedish.

2. EXPERIMENT 1

2.1 Hypotheses

In order to access phonetically conditioned attitudes to foreign accent, we can look at a number of non-native phonetic features which crop up in several different accents of Swedish. Our hypothesis is that different phonetic features of a certain non-native speaker's pronunciation cause native listeners to react in different ways (HYPOTHESIS 1). This kind of discrimination is clearly independent of the listener's attitude to the ethnic group he believes the speaker to represent.

We know a good deal about the way the Swedish population views various immigrant groups from investigations such as that reported in [2]. Given this, we would expect to find that native speakers will perceive non-native speech differently depending on what they know, or believe they know about the speakers' linguistic, and, therefore, ethnic, origin (HYPOTHESIS 2).

Westin [2] reports that blue-collar workers are, on average, less tolerant of immigrants than are other groups. We hypothesize that

native listeners with a lower educational standard (for example, those 17-19 year-olds studying to be blue-collar workers - electricians, mechanics, builders etc) will be more influenced by what they believe about a speaker's linguistic origin than will similar students on theoretical courses (directed at university admittance to subjects such as engineering) (HYPOTHESIS 3).

2.2 Material

From non-native readings of a short text, we listed a large number of NNPs, from which we chose five which occurred in a number of different accents. Five versions of each of these non-native pronunciations each taken from one of two places in the text, produced by non-native speakers, were chosen with the smallest possible total number of speakers, such that the maximum number of comparisons between native reactions to a single speaker's deviant pronunciations could be made. This gave us a total of 25 tokens from 11 speakers, with up to three tokens from each speaker.

In order to let us test hypothesis 3 we used two groups of native speakers: (a) 72 native Swedish students on three-year theoretical courses at upper secondary school (T3), and (b) 33 native Swedish students on two-year practical courses (P2). The only difference between groups P2 and T3 is assumed to be their educational, and therefore social status (cf [3]), in relation to the employment they will be expected to have at the end

of their studies. We have performed extensive preliminary experiments on T3-type listeners (c.f. [4], [5]).

Previous work on linguistic attitudes (for example, [6]) has shown the need for two basic dimensions to describe speaker characteristics elicited from linguistic stimuli. We have used three dimensions representing the perceived importance of eliminating each NNP as uttered by a particular speaker, and the perceived friendliness and educational level of each speaker when he or she uses particular NNPs.

2.3 Results

Hypothesis 1 predicted that different phonetic features of a single non-native speaker's pronunciation cause naive native listeners to react in different ways. The NNPs spoken by different speakers were often found to be judged differently from each other as regards how important it is to eliminate each NNP. This is also true of the perceived educational level of the speakers, but the perceived friendliness of a speaker is only in one case influenced by the various NNPs he or she uses. It is, however, clearly the case that non-native pronunciations from a single speaker can be judged very differently. This gives us corroboration for hypothesis one.

Our second hypothesis was that the same phonetic feature in different non-native accents will elicit different native responses. We found that the variation between judgements of dif-

ferent non-native speakers' productions of single NNP categories is usually significantly greater than the variation within speakers. This means that our hypothesis is supported. There is no significant tendency for similar non-native pronunciations to be judged in the same way.

The third hypothesis predicts that a speaker's pronunciation will be judged differently depending on what the native informants know, or believe they know about the speakers' linguistic origin. This was tested by comparing the judgements of stimuli where the same NNP from the same speaker is presented more than once with conflicting information about the speakers' linguistic backgrounds). The fact that the listeners accepted this information is a reflection of their limited capabilities of identification of foreign accents, as mentioned above. Hypothesis 3 is not corroborated for either listener group on any dimension. There are very few significant differences between the judgements of the speaker guises. This means that, while the listeners were not aware that they were hearing the same speaker more than once, they were uninfluenced by the information about the speakers' backgrounds when making their judgements.

2.4 Discussion

We have here a clear case of distinct phonetically conditioned attitudes to different non-native pronunciation features. A single speaker's NNPs can be judged differently by

naive native listeners as regards the importance of the NNPs used and, more surprisingly, how friendly (in one case) and highly educated the speaker is perceived to be. If a non-native speaker is perceived as having a lower level of education when he or she lets a final velar nasal be followed by a voiced velar stop than when he or she inserts a vowel in a consonant cluster this can have serious social consequences for the speaker. This has, therefore, implications for the teaching of Swedish as a second language, and is an important finding, since it shows that impressions of accent strength may change while hearing a speaker, and indicates that these impressions may be influenced by the systematic elimination of stigmatized non-native features from the individual's speech.

It would clearly be useful for immigrants to learn to avoid the stigmatized NNPs. Our five NNP categories can only give an indication that differences exist here. Obviously we must have more NNP categories if we are to investigate this area in more detail. The following experiment is an attempt to establish which kinds of NNPs elicit the least favourable reactions from native listeners.

3. EXPERIMENT 2

NNPs occurring in both the readings of texts (once again, "The North Wind and the Sun") and in spontaneous speech (the speaker was encouraged to tell the "story of his life") were extracted from the data base, along with as little

accompanying material as was deemed appropriate. The NNPs were divided into categories. All in all, 94 stimulus tokens were selected, falling into 26 NNP categories. 21 speakers of 13 languages were involved.

Only one listener group was used for this experiment, although they were tested in smaller groups of 20-30. This group was composed of 91 of the same kinds of upper secondary school students of technical subjects in the final year of a three-year theoretical course as the T3 group in the last experiment, although none of the students took part in both experiments. The same three judgement dimensions were used as in the first experiment: NNP importance, friendliness and education.

The second experiment showed again that there was more difference between the judgements made by a new listener group (similar to the T3 group) of the speakers than of the various NNPs. A list of the speakers and a list of the NNPs occurring in our material were compiled in the order of the judgements they elicited from the listeners. The NNPs associated with the least favourable overall impressions would, naturally, be worth avoiding for non-native speakers. These results have obvious pedagogical implications.

4. REFERENCES

[1] LAMBERT, W., R. HODGSON, R. GARDNER & S. FILLINBAUM (1960) "Evaluational reactions to spoken languages". *Journal of Abnormal and Social Psychol-*

ogy 60 (Jan) 44-51.

[2] WESTIN, C. (1984) "Majoritet om minoritet, en studie i etnisk tolerans i 80-talets Sverige". En rapport från Diskrimineringsutredningen. Stockholm: LiberFörlag.

[3] FARINGER, G. (1982) "Språk och social identifikation - en undersökning bland gymnasieelever i Stockholmsområdet". (FUMS rapport nr 106.) Uppsala.

[4] CUNNINGHAM-ANDERSSON, U. & O. ENGSTRAND (1988) "Attitudes to immigrant Swedish - A literature review and preparatory experiments", *Phonetic Experimental Research at the Institute of Linguistics, University of Stockholm (PERILUS) 8*, 103-152.

[5] CUNNINGHAM-ANDERSSON, U. & O. ENGSTRAND (1989) "Perceived strength and identity of foreign accent in Swedish", *Phonetica 46*, 138-154.

[6] LAMBERT, W., H. GILES & O. PICARD (1975) "Language attitudes in a French-American community". *International Journal of the Sociology of Language 4*, 127-152.

EVOLUTION DE L'ACCENT MERIDIONAL EN FRANCAIS NICOIS:
LES NASALES

A. Thomas

Université de Guelph, Ontario, Canada.

ABSTRACT

A mini-survey of Southern French pronunciation was carried out in the Nice area to determine to what extent and on what points of the system the local vernacular is influenced by standard French pronunciation.

Interviews with speakers from three generations, featuring both spontaneous and formal speech, provide useful information on the evolution of nasal vowels - the subject of the present paper - and other features of Niçois pronunciation.

1. INTRODUCTION

L'intérêt des dialectologues pour les français régionaux étant relativement récent, nous sommes encore mal renseignés sur les différences phonétiques entre le français standard (FS) et celui d'un grand nombre de locuteurs, dont le parler est encore marqué par les substrats régionaux. Ceci est particulièrement vrai pour le français méridional (FM), qui nous intéresse ici, et sa variante niçoise qui, à notre connaissance, n'a fait l'objet d'aucune étude phonétique. De plus, parmi les travaux existants, les préoccupations diachroniques font défaut, malgré l'intérêt évident que présente l'aspect évolutif des parlers régionaux.

C'est en partie pour combler cette lacune que nous avons constitué un mini-corpus de français niçois, où plusieurs générations sont représentées, de manière à pouvoir observer, en "synchronie dynamique" [3], l'évolution de certaines caractéristiques phonétiques du FM. Une analyse de type labovien permettra de déterminer si celles-ci se perdent en milieu niçois et, si oui, à quel rythme, et selon quelles modalités stylistiques et sociolinguistiques. On se bornera ici à l'analyse des nasales.

2. PROTOCOLE D'ENQUETE

Le manque d'espace nous limite à une présentation squelettique. Pour tous détails complémentaires, veuillez consulter [4].

2.1. Sujets

- Choisis à partir des contacts personnels de l'enquêteur, sans souci des règles de représentativité.
- 13 sujets, répartis entre trois familles et trois générations, tous originaires des Alpes-Maritimes (surtout Nice et environs immédiats).
- Age et condition sociale relativement homogènes à l'intérieur de chaque génération. 3e. génération: travailleurs manuels, 62-83 ans; 2e: petite bourgeoisie (surtout enseignants), 35-43 ans; 1ère: lycéens ou étudiants, 16-22 ans.

2.2. Interview

- Conversations d'une demi-heure en moyenne sur des sujets d'intérêt général, au domicile des sujets, qui ont été interviewés individuellement (sauf un couple pressé) et par le même enquêteur non-méridional, mais connu des familles ou présenté par une personne de confiance.
- Atmosphère très détendue.
- Lecture, "aussi naturelle que possible", de phrases et d'un article amusant de Nice-Matin. Cette épreuve était justifiée par notre intention d'obtenir deux niveaux de formalité verbale et un énoncé commun à tous les locuteurs pour faciliter les comparaisons.

2.3. Choix des variables

- Celles où le contraste est normalement le plus marqué entre FM et FS: voyelles à double timbre, E muet, /R/ final et voyelles nasales.

- Pour ces dernières, on a examiné l'épenthèse d'un appendice consonantique (1) en finale absolue ou devant voyelle (sauf cas de liaison), et devant consonne (2) dentale, (3) labiale et (4) vélaire.

- Variables sociologiques limitées au sexe et surtout à l'âge, qui recouvre en partie des distinctions sociales (cf. ci-dessus).

2.4. Traitement des données

- Analyse auditive des textes lus et d'au moins 10 mn de parole par sujet: nous avons attribué une valeur phonétique binaire à chacune des occurrences préalablement repérées sur les transcriptions (+ ou - épenthèse), les réalisations intermédiaires ayant été comptées dans la catégorie dont elles se rapprochaient le plus.

- Le nombre n de réalisations "méridionales" a ensuite été rapporté au nombre total N d'occurrences de la

variable. On a ainsi obtenu un "degré de méridionalité", c'est-à-dire un pourcentage représentant l'écart entre la prononciation du sujet et la norme du FS (0%), utilisée comme point de référence connu des Niçois et commode pour l'analyse.

3. RESULTATS

3.1. Variation phonétique

Tableau 1. Taux d'épenthèse en fonction du contexte

Contexte	Parole sp.		Lecture	
	N	%	N	%
V, ##	586	14	169	17
C den.	1177	19	1001	18
C lab.	572	21	338	16
C vél.	283	23	234	21

Sur 4360 occurrences de voyelles nasales dans l'ensemble du corpus, seulement 18% sont suivies d'un appendice consonantique, ce qui paraît bien peu en regard des observations de Detrich [1], qui en trouve 47% (d'après nos déductions) dans la lecture des Lettres de mon moulin par Fernandel. Mais cette comparaison est douteuse, puisqu'il s'agit là d'un artiste âgé, probablement peu représentatif du parler local, limité au style de la lecture, et lisant un texte à forte connotation méridionale. On ne peut donc rien en tirer sur d'éventuelles différences régionales entre Marseille et Nice.

Quand on considère séparément les 4 contextes définis plus haut, on constate que les pourcentages s'écartent peu de la moyenne (14 < n < 23%), surtout en lecture. On pourrait donc avancer l'hypothèse suivante, en incorporant les résultats de Detrich [1]: la probabilité

d'épenthèse serait indépendante du lieu d'articulation de la consonne suivante (nos données), mais dépendante de sa nature (dures plutôt que douces) et de sa position par rapport au mot considéré (interne plutôt qu'externe; Detrich).

3.2. Variation stylistique

Le tableau 1 montre clairement que les locuteurs interviewés n'ont aucunement modifié leur réalisation des nasales en passant de la parole spontanée à la lecture. Il en est de même, d'ailleurs, pour les autres variables méridionales analysées à partir de ce corpus. Cette uniformité contraste avec les travaux de Le Douaron ([2]; sujets des Bouches-du-Rhône), qui suggère que cette variable "est conditionnée par le degré de formalité de la situation de parole", contrairement au E muet.

Plusieurs explications possibles viennent à l'esprit: 1) l'insuffisance des données; 2) une distinction régionale Nice-Marseille; 3) un degré de formalité similaire en lecture et en parole spontanée dans notre corpus; 4) "l'oro-nasalité en syllabe atone serait la composante méridionale que les locuteurs auraient le plus de mal à contrôler" ([2], ce qui rendrait son hypercorrection (dans le sens labovien du terme) difficile en contexte formel; 5) l'absence de besoin d'hypercorrection, d'ailleurs confirmé par les commentaires des sujets eux-mêmes, qui acceptent aussi bien leur prononciation régionale que celle du FS.

Le choix entre ces explications ne pourra se faire qu'au prix de nouvelles enquêtes, plus vastes et mieux contrôlées.

3.3. Variation sociophonétique

Les différences relatives au degré de formalité et à l'environnement phonétique de la variable étant faibles (cf. ci-dessus), on a rassemblé les pourcentages partiels en un seul pourcentage global d'épenthèse pour chaque sujet.

3.3.1. LE SEXE

Etant donné la petite taille du corpus et afin d'éliminer autant de facteurs que possible autres que le sexe des locuteurs, on n'a retenu ici que les quatre couples disponibles, également divisés entre la génération des parents et celle des grands-parents.

Les résultats (voir tableau 2, ci-dessous) indiquent que les hommes réalisent l'épenthèse au moins deux fois plus souvent que les femmes, qui se situent beaucoup plus près de la norme nationale que leurs maris, illustrant ainsi une tendance souvent observée en sociolinguistique labovienne.

3.3.2. L'AGE

L'écart entre maris et femmes étant assez important, on a dû limiter les comparaisons diachroniques aux sujets d'un même sexe, ce qui réduit grandement la valeur des résultats. Deux mesures ont été faites comparant les pourcentages obtenus par les pères et leurs fils ou les mères et leurs filles (8 sujets) et d'autre part, par les grands-pères et leurs petits-fils (4 sujets; il n'y a pas de fille parmi les "jeunes"). Cela a permis d'observer en synchronie dynamique, sur une puis deux générations, l'évolution phonétique dans les trois familles interviewées.

Tableau 2. Taux d'épenthèse en fonction du sexe et de l'âge des sujets.

Variables	%	Ecart (%)
Sexe F	18	
Sexe M	39	+ 21
Enfants	8	
Parents	28	+ 20
Enfants	3	
Gds-parents	48	+ 45

Dans les limites de validité des chiffres proposés, qui ne reposent dans le dernier cas que sur 4 sujets, les résultats suggèrent une nette diminution de l'épenthèse, qui semble s'effectuer au rythme minimum de 20% par génération. L'évolution est particulièrement frappante quand on compare les grands-parents, chez qui l'épenthèse est fréquente, à leurs petits-enfants, qui la connaissent à peine.

Notons au passage que cette diminution n'est probablement pas attribuable aux différences sociales existant entre nos sujets, puisqu'on observe des différences sensibles entre les deux premières générations, qui sont pourtant de même condition sociale.

4. CONCLUSION

Notre étude de l'épenthèse consonantique nasale a permis de documenter (du moins pour quelques sujets de la région niçoise et si l'on accepte que la synchronie dynamique est un substitut acceptable de l'étude proprement diachronique) une rapide évolution vers les formes non-épenthétiques du FS, particulièrement chez les sujets féminins. Cette rapidité s'explique peut-être par le fait que les

facteurs externes (omniprésence du FS dans les médias, chez les migrants internes et les touristes, ajoutée à une régression du nissart, où l'épenthèse est la norme) convergent ici avec la dynamique interne du FM, qui suit le même chemin que le FS, mais avec quelques siècles de retard (chute des consonnes finales, et en particulier des nasales, depuis le latin vulgaire).

Ces conclusions - d'ailleurs similaires à celles obtenues pour E muet (cf. [4]) - sont évidemment sujettes à caution, étant donné le nombre réduit de sujets et les critères de sélection appliqués ici. Nous les proposons simplement comme piste à suivre vers une meilleure connaissance du français méridional.

REFERENCES

- [1] DETRICH, D. (1979), "Nasal Consonant Epenthesis in 'Southern French'", Current Issues in Linguistic Theory 9, 521-529.
 [2] LE DOUARON, M. (1985), "Etude des parlers méridionaux: analyse factorielle", Travaux de l'Institut de phonétique d'Aix 10, 245-285.
 [3] MARTINET, A. (1975), "Diachronie et synchronie dynamiques", Evolution des langues et reconstruction, Paris, 5-10.
 [4] THOMAS, A. (1991), "Evolution du E muet en français niçois", Information, Communication 11, Toronto.

IDENTIFYING FOREIGN LANGUAGES

Z. S. Bond and Joann Fokes

Ohio University, Athens, Ohio U.S.A.

ABSTRACT

This experiment examined listener abilities to identify the language in which a message is spoken. Short samples in five languages were presented to listeners for identification. All groups of listeners identified the samples at better than chance levels. The respective native languages and English received the highest identification. Confusions among samples varied according to native language.

1. INTRODUCTION

People who work in environments where they commonly hear foreign languages claim that they develop the ability to identify languages without understanding any of them. Surprisingly, whether people indeed have this ability has never been investigated in spite of the fact that anecdotal accounts are common and claim considerable sophistication.

House and Neuburg [7] examined the possibility of identifying languages from a statistical distribution of segment types. This work dealt with feasibility rather than human performance.

The vast literature comparing the phonetic structures of languages has dealt

with similarities and differences rather than with information which identifies a particular language. The consonant and vowel inventories have been investigated from the point of view of interference with language learning [5]. Languages have been compared according to their phonetic implementation of a linguistic process [4], the relative timing of syllables [3] or their overall use of fundamental frequency [1,6]. Considerable effort has been devoted defining the rhythmic patterns of various languages, [2]. Although some of these differences are undoubtedly responsible, none of the work directly addresses the question of how listeners use phonetic properties to identify languages.

The purpose of this experiment was to determine how well listeners of various language backgrounds are able to identify spoken samples of languages which they do not speak.

2. METHOD

2.1. Materials.

Two native speakers of Chinese, Japanese, Spanish, Arabic and English recorded short paragraphs taken from newspapers. These are languages commonly spoken by students at Ohio University.

All the listeners would have had some exposure to them.

To prepare samples for presentation, the speech was digitized and four two-second samples per speaker were excerpted. The samples were fluent without hesitations or long pauses. After normalization to the same peak amplitudes, two samples for each speaker were digitally mixed with noise at a S/N ratio of 3 dB. The noise condition was included to examine the contribution of vowel and particularly consonant inventories to the identification of the languages.

The samples were recorded in random order for a listening test consisting of 40 test items (5 languages, 2 speakers of each, 2 speech samples from each speaker, 2 listening conditions).

2.2 Subjects.

Seven groups of listeners were tested. First, 14 native English speakers, undergraduate students at Ohio University, limited in experience with foreign languages. Second, 13 native English instructors in the Ohio Intensive English Program. These listeners are very familiar with speakers of other languages and each had spent at least one year in a non-English speaking environment. Third, ten native speakers of each of the other languages used in the test: Arabic, Chinese, Japanese, and Spanish. Finally, ten native speakers of languages other than the sample languages, that is, Korean (6), Bahasa Malaysia (3), and Buhusu, a Bantu language of Kenya (1). For these listeners, all the sample languages are foreign, though they know English well.

2.3. Procedure.

The listeners were tested in a quiet room, in small groups. They were asked to identify the languages in a forced choice format.

3. RESULTS

3.1. Identification.

The percent correct identifications of the all four samples of the language are given in Fig. 1. The data are combined for all listeners. All listeners identified English samples at very nearly 100% in quiet. Arabic, Chinese, and Spanish were identified at slightly lower rates, while Japanese was identified least accurately. The pattern of correct identification in noise corresponded exactly to the pattern obtained in quiet, simply with more errors present.

3.2. Language background.

The identification scores of each listener group are given in Fig. 2. As might be expected, the teachers experienced with languages had the highest scores, in quiet. The Japanese listeners performed best in noise while the Spanish listeners had the lowest scores. Overall, all language groups identified the languages at above chance rates. This was most clearly true when the samples were presented in quiet.

Table 1. gives identification scores for all five languages and all listener groups. Each group identified its respective native language and English at very high rates. In the noise condition, scores for all groups and languages were depressed.

The ranges of scores were relatively consistent for listeners from different backgrounds. In all groups, some listeners made perfect,

or nearly perfect, scores in quiet. In all groups, other listeners made relatively low scores. The lowest individual score was made by a Spanish listener, 8 out of 20 items correct in quiet. The ranges of scores in noise were depressed for all groups of listeners.

Table 1. Percent correct identifications of languages by listeners from different backgrounds. The top row gives identification scores in quiet, the bottom row gives scores in noise.

	EN	AR	CH	JP	SP
EN(s)	100	88	91	78	73
	79	59	34	31	46
EN(t)	100	92	79	69	88
	73	58	56	33	56
AR	98	98	65	50	93
	58	90	43	25	53
CH	100	65	100	85	65
	68	53	93	48	53
JP	93	73	98	85	73
	80	50	78	90	58
SP	88	63	45	45	95
	50	38	20	10	63
OTHR	95	78	90	80	58
	68	64	78	63	40

3.3. Confusion patterns.

The most obvious confusion pattern affected Chinese and Japanese. Listeners who were not Asians tended to confuse these two languages, as if they were operating with a broad category Oriental Language. Asian listeners, including those from Korea and Malaysia, seldom confused the two. The Spanish listeners in particular, had difficulty identifying these two languages. The Asian listeners, in turn, tended

to confuse Spanish and Arabic.

4. DISCUSSION

4.1. Limitations.

There are two limitations of the experiment. The first is a lack of control of the amount of experience the different listener groups have had with languages. The language backgrounds of the listeners are confounded with experience hearing various languages. At this time, we do not know how much and what kind of exposure to languages allows listeners to identify them.

The second problem concerns confounding of the language samples with speaker characteristics. Each language was represented by only two speakers. Although all the samples used were different, it is possible that listeners relied on speaker characteristics in making language identifications. A listener may have adopted a strategy of identifying, for example, a relatively high pitched voice as a Chinese speaker or a fast rate of speech as Japanese.

4.2. Conclusion.

The conclusion is that listeners are able to identify languages which they do not know. Since noise decreased the identification scores rather than altering the patterns, it is possible to infer that listeners are relying on suprasegmental properties of languages as much as, or more than, consonant and vowel inventories.

Listener experience with various foreign languages was a major factor in their ability to identify languages. Asian listeners with experience with Asian languages identified Chinese and

Japanese accurately. Arabs and Spanish listeners from South America have had little experience with Asian languages and tended to confuse them.

Some listeners from each group were very good at the task while others made many misidentifications. Whether the differences in scores are a result of individual talent or experience is, at this time, unknown.

How do listeners identify a language? They may proceed by a process of elimination: 'I don't understand it, so it's neither English nor my native language. It must be _____.' Alternatively, they may have developed prototypical auditory patterns which characterize languages.

5. REFERENCES

- [1] BECKMAN, M. (1986), Stress and non-stress accent, Dordrecht: Foris.
- [2] DAUER, R. M. (1983), Stress-timing and syllable-timing reanalyzed, Journal of Phonetics, 11, 51-62.
- [3] DELATTRE, P. (1966), A comparison of syllable length conditioning among languages, International Review of Applied Linguistics, 4, 184-196.
- [4] DELATTRE, P. (1969), An acoustic and articulatory study of vowel reduction in four languages International Review of Applied Linguistics, 7, 295-325.
- [5] DIPIETRO, R. J. (1971). Language Structures in Contrast. Rowley, Mass.: Newbury House.
- [6] EADY, S.J. (1982), Differences in the F0 patterns of speech: Tone language versus stress language, Language and Speech, 25, 29-42.
- [7] HOUSE, A. S., and NEUBURG, E. (1977), Toward automatic identification of

the language of an utterance. Preliminary methodological considerations, Journal of the Acoustical Society of America, 62, 708-713.

Fig. 1. Identification of languages by all listeners.

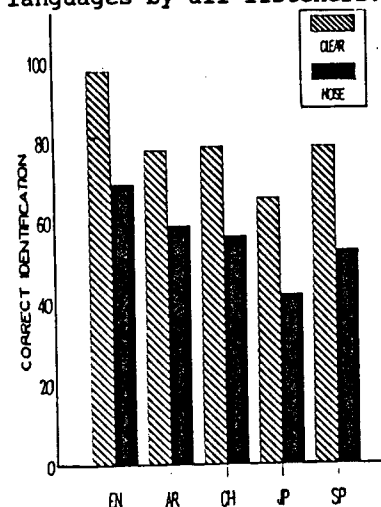
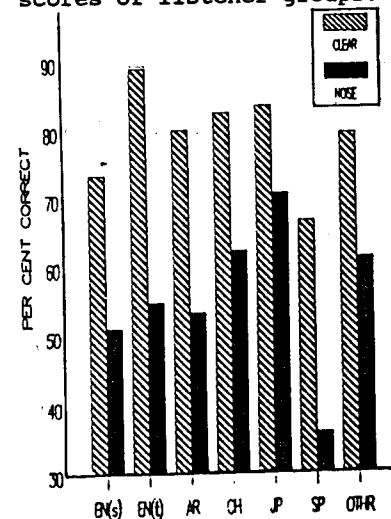


Fig. 2. Identification scores of listener groups.



COMPREHENSION OF VOCALIZATIONS ACROSS SPECIES

Reijo Aulanko*, Lea Lelnonen, Ilkka Linnankoski and Maija Laakso

*Department of Phonetics, University of Helsinki, Finland
Department of Physiology, University of Helsinki, Finland

ABSTRACT

This is a study of how well naive human listeners can interpret the vocalizations of stump-tail macaques (*Macaca arctoides*). The monkey vocalizations were recorded in different behavioural situations. The human listeners were asked to classify 18 vocalizations into one of 7 semantic categories. The listeners were quite unanimous in their judgements, which indicates that they may have based their interpretations on some kind of common "feature analysis" of the vocalizations. The interpretations of the listeners were also mostly "correct", i.e., the listeners were able to infer the situation in which the monkey had produced the sound. These results may be taken to suggest a possible common basis for the vocal behaviour of all primates.

1. INTRODUCTION

It is well known from everyday life that people and their pet animals can understand each other (or that they at least seem to reach a consensus on certain issues). The owner of a cat or a dog should find it easy to make very accurate interpretations (according to personal judgement) of the behaviour of the pet, e.g. of its vocalizations. Similarly, an animal sometimes reacts to the speech of its human companion as if it understands the human language. These cases are not, however, indications of *language* comprehension in the strict sense. The reactions of an animal are determined primarily by all kinds of non-verbal cues, and the most important phonetic aspects of the human speech are its *prosodic* characteristics — rather than the purely phonological structure of the utterance.

To put it simply, comprehension of a vocal message is an interpretation or understanding of the "internal state" of the sender. The correctness of the interpretation can be inferred from the reaction of the receiver. Humans can react verbally, but in the case of other species we have to deduce the interpretation of the message only on the basis of other kinds of overt (non-verbal) behaviour.

Because of their common evolutionary history, the basic mechanisms of sound production are similar in all mammals. There are similarities in the vocal apparatus as well as in the neural control of behaviour. Vocalizations of non-human primates are taken to be mainly reflections of their emotional-motivational state. In human speech, indications of such 'internal' states are often conveyed by prosodic or paralinguistic features. There may be enough acoustic similarity in the emotional-motivational vocalizations of human and non-human primates for comprehension across species.

In human speech, the various emotional and motivational states are reflected primarily in the general voice quality and the prosodic characteristics of speech, i.e. pitch, rhythm, and loudness (e.g. [4, 7]). These auditory characteristics normally co-occur with different kinds of facial expressions and body movements, but the auditory cues are usually sufficient for the identification of the speaker's emotional state.

Human beings are used to inferring the emotional state of a speaker from the acoustic characteristics of his/her speech. An interesting question would be how well these "inference rules" can be applied to the vocalizations of another species.

2. AIM

In this study [3], we explored the ability of the representatives of one species to interpret the vocalizations of another species. More specifically, we tried to determine how well naive human listeners can interpret the vocalizations of another primate species, viz. stump-tail macaques (*Macaca arctoides*). "Interpreting" is here defined as identifying the emotional-motivational state of the monkey during the production of different sounds.

The ultimate aim in studies like this is to resolve the question of a possible common control of emotional-motivational vocal behaviour in mammals. In other words, we are looking for universals in communicative behaviour.

3. RESEARCH MATERIAL

Sounds. Recordings of the macaque vocalizations were made in many different behavioural situations at the Department of Physiology, University of Helsinki, in the colony of stump-tail macaques (*Macaca arctoides*) at present consisting of 12 monkeys (Marantz CP430 tape recorder, AKG C 568 EB microphone). On the basis of the situation and the total behaviour of the monkey, the sounds used in this study were taken to represent seven different categories of psychological states: (1) aggression, (2) fear, (3) sexual arousal, (4) dominance, (5) submission, (6) contentment, (7) calling / informing (contacting). The criteria used in this classification were based on, e.g., the posture and facial expressions of the monkeys, as they are generally used in primate behavioural studies [1, 2].

The vocalization sequences were digitized and tapes for the listening test were prepared, where the vocalizations occurred in a random order. The vocalizations in the test material were analyzed acoustically using sound spectrograms and computerized FFT spectra (Fig. 1). The acoustic characteristics of the vocalizations are described elsewhere [3].

Listening test. Eighteen sound sequences ("whole vocalizations") were selected from all the recorded material for the listening test. The 18 sounds represented different behavioural situations.

A total of seventy-five subjects (50 women and 25 men) participated in the listening test. They were 19–62 years of age, most of them students (of medicine,

dentistry, and anthropology), but there were also some speech therapists, medical doctors, technicians, and nurses. The subjects were not familiar with the vocalizations of the *Macaca arctoides*, but 43 of the 75 listeners had daily contacts with domestic or pet animals.

In a forced-choice test, the subjects were asked to classify each vocalization into one of the seven response categories, each of which was described by (the Finnish equivalents of) the following adjectives:

1. angry, cross, raging
2. frightened, timid, terrified
3. ecstatic, excited, orgasmic
4. commanding, threatening, domineering
5. submissive, pleading, begging
6. satisfied, satiated, delighted
7. calling, informing, addressing

The subjects were given two minutes to become acquainted with the response classes by thinking about each adjective momentarily.

The sounds were presented in a random order. Five of the 18 sounds were included twice in the test tape in order to find out the replicability of the subjects' classifications. (Thus, there was a total of 23 sounds to be judged.) The subjects heard a sound sample twice before a 10-second response interval during which they had to write down the number of the response class that best characterized the sound.

4. RESULTS

4.1. Listener agreement

The responses were not distributed randomly, i.e., the subjects were quite unanimous about the "meaning" of most of the monkey vocalizations (Table 1). The most variable responses were elicited by the vocalizations produced by aggressive monkeys, whereas the listeners were most unanimous in their responses to the "dominance roar" of the leading male.

4.2. "Correct" interpretations

On the average, 60 per cent of the listeners' interpretations were "correct". A response was defined as correct when it corresponded to the original behavioural classification of the monkeys' vocalizations. Somehow the listeners could infer the situation where the sound had been produced.

It has to be noted that certain sounds were originally taken to reflect at least two behavioural classes simultaneously. For example, one of the vocalizations was produced by a female macaque in a situation where her non-vocal behaviour indicated both aggression and fear. These alternatives have not been taken into account in Table 1, where only the responses falling on the diagonal are considered "correct", on the basis of the primary characteristic of the monkey's behaviour.

There were no general differences between the interpretations of men and women, or between those of younger and older listeners, although some individual vocalizations were classified differently. In contrast, daily contact with animals had a significant effect: those listeners who owned animals had more correct answers than those who did not have pets at home (61.5 % vs. 56.8 %).

The sounds that were best identified were a vocalization of a female monkey associated with pleasure (85 % of the subjects had the correct classification) and a dominance roar of a dominating male monkey (84 %). All meaning categories included vocalizations that were classified correctly by more than half of the listeners. Most of the subjects gave the right answer to 13 different vocalizations. One vocalization (threat grunts of a female monkey) was misclassified by all subjects. The distributions of the classifications of the five vocalizations that were presented twice remained stable.

5. DISCUSSION

The rather high general agreement among listeners shows that human listeners do tend to interpret monkey vocalizations that they have never heard before on the basis of some common ideas about the effects of different emotions on the sound production of another primate species.

The high proportion of "correct" interpretations shows that there are acoustic characteristics in the vocalizations which enable naive human listeners to infer the emotional-motivational state of the vocalizing macaque.

The most plausible explanation of the ability to interpret these monkey vocalizations is that the effects of different emotional and motivational states produce rather similar effects both in humans and

in macaques. The possible similarity in the acoustic cues of different affective states in humans and monkeys is treated in more detail elsewhere [3].

The present results suggest that there is a common reference / interpreting scheme with regard to the effects of emotional states on vocal behaviour, according to which the humans interpret all the animal vocalizations they encounter.

Listeners who had daily contact with animals as pets gave more correct interpretations than the others, which suggests that part of the ability to comprehend another species is acquired by experience. However, a proportion of correct responses well above chance level was reached even by those subjects who did not have daily contact with animals. This proves that such close contacts are not necessary for a certain ability to interpret correctly the sounds of another primate species.

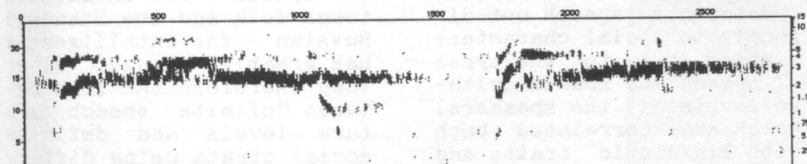
Evidence from this study lends support for the hypothesis that there is a common basis for the recognition of vocalizations between primate species.

6. REFERENCES

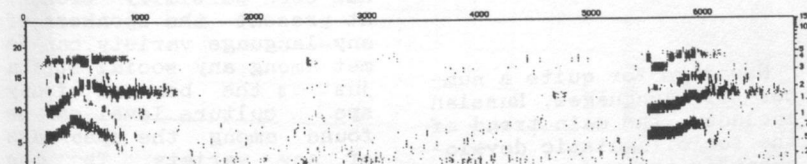
- [1] BERTRAND, M. (1969), "The behavioral repertoire of the Stumptail Macaque", Basel: Karger.
- [2] CHEVALIER-SKOLNIKOFF, S. (1974), "The ontogeny of communication in the Stumptail Macaque (*Macaca arctoides*)", Basel: Karger.
- [3] LEINONEN, L., LINNANKOSKI, I., AULANKO, R. & LAAKSO, M. (in preparation), "Interspecies vocal communication of emotions: Man and Macaque."
- [4] SCHERER, K.R. (1981), "Speech and emotional states". In *Speech Evaluation in Psychiatry* (J.K. Darby, editor), 189-220. New York: Grune & Stratton.
- [5] TOIVONEN, R. (1988) Intelligent Speech Analyser (ISA) järjestelmän käyttöohje. Tampere, 22.6.1988.
- [6] TOIVONEN, R. (1990) Tiivistetty ISAn käyttöohje. Tampere, 13.6.1990.
- [7] WILLIAMS, C.E. & STEVENS, K.N. (1981), "Vocal correlates of emotional states". In *Speech Evaluation in Psychiatry* (J.K. Darby, editor), 221-240. New York: Grune & Stratton.

TABLE 1. Percentages of human interpretations of vocalizations representing different emotional-motivational states of stumptail macaques. (Proportions less than 10 % are not indicated in the table. The number of vocalizations representing each behavioural state in the test material is given in parenthesis.) The figures on the diagonal indicate the proportion of responses considered "correct" in the strictest sense.

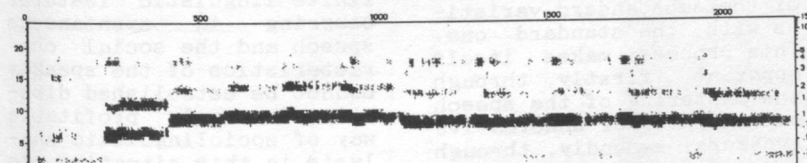
Behavioural states	Human interpretations						
	angry	frightened	ecstatic	commanding	submissive	satisfied	calling
aggression (4)	29	17	22	14			13
fear (2)	24	60					
sexual arousal (2)			46	32			
dominance (1)	11			84			
submission (4)					50	15	23
contentment (3)				10		68	12
contacting (2)		18			12		50



A. Female screams of fear



B. Female contact call



C. Female cooing

FIGURE 1. Examples of the macaque vocalizations analyzed with a psychoacoustic spectrogram program, a specialty of the Intelligent Speech Analyser system [5, 6]. Horizontal axis = time (ms), vertical axis = frequency in Bark (left) and kHz (right).

SPEECH TYPES, SPEECH CULTURE
AND THEIR SEGMENTAL CORRELATES

N. Geilman

Leningrad University, Leningrad, USSR

ABSTRACT

The experiments with various segmental traits are described, demonstrating the advantage of a new method of analysis of sociolinguistic variability. It is applicable to the languages with strongly expressed democratization tendency and consists in correlating a set of linguistic traits of spontaneous speech not directly to social characteristics, but to the types of speech and speech culture levels of the speakers, which are correlated both with linguistic traits and with social characteristics.

Nowadays for quite a number of languages, Russian included, the main trend of the sociolinguistic development is their democratisation, i.e. the rapprochement of the substandard varieties with the standard one. This process makes itself apparent firstly, through the mastering of the speech norms by former unnormative speakers, secondly, through the penetration of substandard features into the norm-bearers' speech. This trend being caused by such social processes as the mass-media spreading, the educational level uplifting, both rural

and urban population's migration, has hard consequences for the linguistic development. For Russian it manifests itself through the gradual erasing of the distinctions between dialectal, urban popular and standard (literary) speech. The situation of the beginning of the century, when local dialects were typical for peasantry, urban popular speech - for uneducated towns-folk and the Standard Russian - for intelligentsia, has greatly changed since then. Therefore the link between definite speech culture levels and definite social strata using different varieties of Russian has been partially broken. At present the speakers of any language variety can be met among any social strata just as the bearers of any speech culture level can be found among the speakers of any variety. Thus the correlation between the definite linguistic features occurring in spontaneous speech and the social characteristics of the speaker cannot be established directly. The only profitable way of sociolinguistic analysis in this situation is to correlate the social characteristics of the speakers to their speech type (ST: standard or normative, not quite normative, unnormative) and their speech

culture level (SCL: from high to low) and the latter two to the set of linguistic features, thus linking the social belongings of the speakers with definite linguistic features in the indirect way. Besides, the social structure of the linguistic groups and the linguistic structure of the social groups should be established.

It is exactly in this way that the present study was carried out. Firstly, it was shown in the course of the auditory experiments that such social characteristics as the educational level, the social status, the character of profession etc. are very poorly recognized by experts when listening to spontaneous texts: the percentage of correct recognition for 29, 44 and 70 speakers in different series was accordingly 56, 36 and 31 on the average. On the contrary, ST and SCL are unanimously ascribed by a group of listeners to 88 and 98% of speakers out of 206. Secondly, it has been found out in the course of the correlation analysis based on 69 1.5-2 min texts that out of 80 specific traits of different linguistic levels 42 correlate significantly with ST and 56 with SCL; meanwhile only 37 correlate with social position, 34 with educational level, 18 - with professional usage of public speech, 10 - with social status etc. At last, it has been demonstrated that the linguistic structure of various educational, professional and age groups differs to a considerable extent.

Having thus proved the rightfulness of methodology, the question arose which segmental features are sig-

nificant if all for the estimation of ST and SCL both by the researcher and by the experts.

First, it was decided to test the influence of the segmental layer upon the estimation of ST and SCL as compared to the other language layers. For the purpose 4 series of experiments were carried out where 18 listeners had to establish ST and SCL of the 6 informants (2 possessing high, 2 mid and 2 low SCL) by 1) listening to the isolated words cut out from the original text (segmental information preserved), 2) listening to the text in noise (prosodic information), 3) reading the written version of the text (lexico-syntactic information), 4) listening to the original spontaneous text (integral es-

1
timation). The results show that the significance of different levels depends on the SCL: with the bearers of low SCL the prosodic level is the poorest since it has the lowest SCL marks; next goes the lexico-syntactic level which is somewhat better organized judging by the estimation; the isolated words series having the highest marks, the segmental subnormal traits influence the SCL estimation least of all. The shortcomings of all the levels being summed up, the integral SCL marks are the lowest. On the contrary, with the bearers of mid and high SCL the integral estimation is the highest showing the integrity of the natural texts compared to the unnatural character of the other

1 These experiments were carried out together with N. Bogdanova and P. Skrelin.

series. The isolated words and the written texts have lower marks than the texts in noise. That testifies to non-importance of segmental and syntactic organization for SCL estimation as compared to that of prosodic level, as the integral SCL estimation grows up parallel to the rising of the prosodically based marks.

Then an attempt was made to find out the segmental features which can help the researcher to establish the linguistic belongings of the speaker. As we have demonstrated earlier [1] although the list of peculiar segmental traits present in different varieties of spoken Russian coincides, the frequency of their occurrence differs depending on ST and SCL. For further analysis four segmental features frequently occurring in spontaneous speech were chosen: 1/ spirantization of stops (/č/ taken separately), 2/pronunciation of /č/ as not enough palatalized [c'], 3/ vocalization of consonants /l, l', r, r', v, v'/, 4/stronger /a/-vowel reduction than prescribed by the norm: pronunciation of [ɔ] instead of [ʌ] in the 1st pre-stressed syllable, at the beginning and in the end of the words. 4 texts (about 11.5 th phonemes) were transcribed, two speakers (a linguist and a worker) being the natives of Moscow and thus the bearers of the norm and two others (a worker and a journalist) - the natives of a small Northern city. The two of them (the workers) - the bearers of low SCL, while the other two - high SCL-speakers. The results show that all the chosen traits are typical for spontaneous speech as they occur in all the 4 texts. But there

exists a clear cut tendency to their different distribution in the texts: two features (№1, 3) are more typical for high SCL-speakers (the linguist and the journalist) while the other two (№2,4) are more frequent with the low SCL-speakers. As for ST, the traits connected with high SCL are also linked with the norm, as they are more frequent in the texts spoken by the Moskovites, both with high and with low speech culture. That testifies to the fact that SCL is still somewhat higher with the normative speakers. Still such traits as spirantization and vocalization are more dependent on the normative distinctions while the other two are influenced more by SCL differences. Although the further research is desirable where both a range of traits studied and a number of speakers would be increased, these results show that there exist definite segmental traits whose number of occurrence lets us to distinguish between different ST and SCL by the research analysis of spontaneous texts.

On the next stage of the investigation it was decided to study whether four of the described traits are essential for auditory analysis, i.e. whether their presence in speech influences the estimation of ST and SCL by the experts. For this purpose an experiment was staged where isolated words cut out of 6 texts (2 high SCL, normative ST, 2 mid SCL, non-fully normative and 2 low SCL unnormative) were listened to by a group of 10 experts in 5 series. Into the 1st series the words containing no peculiar pronunciation traits were included. The other 4

contained words with vocalized sonorants, spirantized stops, /č/ not enough palatalized, the cases of stronger /a/-vowel reduction.

The results of auditory analysis confirmed the results described above, that the SCL, when being estimated by the isolated words, is almost equal to that of the original text with low SCL bearers and far too lower with mid and high SCL bearers. That confirms the closer link of various linguistic levels with the speakers of high SCL. Besides, the results show that the presence of spirantized stops and vocalized sonorants in the series, being typical for normative ST and high SCL almost does not influence the estimation of the first and even rises that of the latter with the speakers of low SCL. As for /č/ not enough palatalized and stronger /a/-vowel reduction, being both of dialectal origin, they act quite in a different way. The presence of the first turns the former normative speakers (by 1st series estimation) into the unnormative but only slightly lowers their SCL marks, while the presence of the second trait influences more SCL marks lowering them and only slightly decreases ST marks. This demonstrates that /č/ not enough palatalized is connected in the mind of the experts with substandard (dialectal) varieties the speakers of which are not compulsory of low SCL. The stronger reduction is, on the contrary, linked with low SCL, typical for urban popular speech, which, being widely spread nowadays, is not estimated as unnormative.

On the whole, all the four pronunciation traits are used by experts in the process of estimating both ST and SCL.

To draw a conclusion it is necessary to underline that for many languages with blotted out social differentiation the sociolinguistic variation can not be described in any other way but indirectly through such general linguistic characteristic of the speakers as the type of their speech and the level of their speech culture, which have quite a definite set of linguistic correlates on all the levels of linguistic analysis, the segmental one included. The existence of these linguistic realities having been proved by auditory analysis results, their correlates can be used both by researchers and by experts to determine the linguistic belongings of the speakers in the course of sociolinguistic analysis, thus indirectly correlating concrete linguistic traits to the social characteristics of speakers.

As for the significance of segmental level for the estimation of speech, it is different by various ST and SCL.

REFERENCES

[1] GEILMAN, N. (1987). Variability of Phonemes in Spoken Russian. Proc. XI ICPHS. V.3, pp.161-164.

1 The author wants to thank her student O. Yugaï for her assistance in the experiments.

MORAIC NASAL AND TONAL MANIFESTATION IN OSAKA JAPANESE: IMPLICATIONS FOR THE REPRESENTATION OF MORA

Y. Nagano-Madsen

Department of Linguistics and Phonetics
Lund, Sweden.

ABSTRACT

A moraic nasal and a CV mora were compared as regards their tonal manifestation. A moraic nasal in accented and post-accented position was found to have stronger energy than a moraic nasal in word final position. A four-mora word composed of either /CVNVCV/ or /CVCVCVCV/, where /N/ represents a moraic nasal, had almost identical duration and F0 configuration within the same accent type. However, a boundary between /CV/ and /N/ was found to be more ambiguous than that between /CV/ and /CV/ both in spectral pattern and in the timing of the onset of F0 change. In slow speech, the second mora (both CV and N) tended to be prolonged regardless of the accent type.

1. INTRODUCTION

Previous work on the perception of mora and pitch accent in Japanese has shown that in a bi-moraic word /ama/, accented either on the first or the second mora, a shift of the original fundamental frequency (henceforth F0) contour caused a change in linguistic and paralinguistic categories within a certain range [1]. In the same experiment, however, listeners' response was markedly different for bi-moraic words composed with a moraic nasal. In words like /aN/ and /heN/, the acceptable range of shift for the F0 contour was much greater.

One possible explanation, phonetically oriented, is that a moraic nasal may be significantly different from a CV mora in its pitch manifestation. Since a moraic nasal in Japanese only occurs in syllable

final position and preceded by a vowel, it may form a coherent unit of accentuation with the vowel in its acoustic manifestation.

Further possibility may arise from the difference in articulatory type. The Japanese moraic nasal is known to vary greatly in its exact phonetic nature. In word final position, as it is in /aN/, it may become more or less a nasalized vowel. Since it has been reported that the timing of vowel articulation and phonatory control is less constantly maintained compared to that of consonant articulation and phonatory control, the difference might arise from a difference in articulatory types [2].

A phonologically oriented explanation may be that the two types of test words differ in their higher constituent. /ama/ is counted as bi-moraic as well as bi-syllabic while /aN/ is mono-syllabic. We may expect that the relative timing of mora and F0 is less important within a syllable.

Very little is known about the exact acoustic characteristics of the moraic nasal in relation to pitch manifestation. In Standard Japanese, only a syllabic-mora, a mora that can form a syllable by itself, bears pitch accent. In Osaka Japanese, however, even non-syllabic moras such as vowel mora and nasal mora can bear pitch accent. I follow Kubozono, adopting the terms "syllabic" and "non-syllabic" moras [5]. This is one of the reasons why the independence of the mora is claimed more strongly for Osaka Japanese. In the present paper we report the results of a pilot study that has

compared the moraic nasal and the CV mora both in accented and post-accented positions using Osaka speakers.

2. EXPERIMENT

The following words served as test words. They are all four-moraic words with or without moraic nasals (/N/). In this position, the first /N/ is said to be realized as [m] assimilating to the following consonant.

/koŋbaN/	"tonight"
/káNbaN/	"signboard"
/kaNpaN/	"deck"
/komádori/	"kind of bird"
/kámatano/	"Kamata's"

These words were embedded in a carrier sentence "sorewa ___ desu (it is ___)" and read 4-5 times by two native speakers of Osaka Japanese at two speaking rates (slow and fast). The recorded data was digitized and processed by Mac Speech Lab and LUPP (Lund Prosodic Parser) using a Macintosh II.

3. RESULTS

3.1. F0 and time dimensions

Typical F0 contours for /koŋbaN/ and /komádori/ of a male speaker are shown in Fig. 1. When the utterances were lined up at the onset of the vowel [o] in /ko/, the F0 contours for the two words were found to be almost identical within the same speaking rate. In fast speech, the entire F0 except for the utterance final position, was raised to a higher pitch range. Within the same speaking rate, the two F0 contours tended to have the same duration, use the same pitch range and have the same timing of F0 rise and fall. Similar observations were made for /káNbaN/ and /kámatano/. There was a tendency for the timing of the F0 rise in /koŋbaN/ to come slightly earlier than that of /komádori/ in fast speech. In slow speech, there was a tendency for the second mora to be prolonged regardless of the accent type. Note that the duration of the second [a] in /kámatano/ is longer than that of the first [a] which is accented (Fig. 2).

3.2. Acoustic characteristics of the moraic nasal

Usually the vowel that precedes a moraic nasal was found to be longer than the vowel that precedes a CV mora irrespective of pitch patterns. This may imply that even when the /N/ is realized as [m] due to the following consonant, the unmarked quality of the moraic nasal which is supposed to be articulated at the uvular region is still there. The difference in vowel duration was most eminent and consistent before an accented moraic nasal. When a moraic nasal was accented (H) or when it appeared immediately after an accented mora as in /káN(baN)/, it was articulated with more energy than in word final position (see Fig. 2).

3.3. Relative timing of F0 and articulatory event

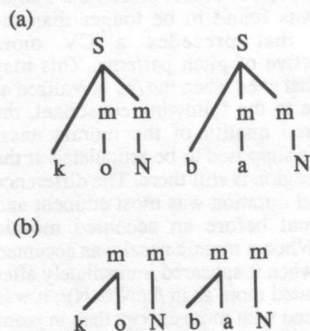
The relative timing of the onset of the F0 rise/fall and articulatory events was examined both from narrow band spectrograms and F0 plotting. As for the CV moras, there was a strong tendency for the onset of F0 rise and fall to occur around the CV-CV boundary both at slow and fast speech.

As for the moraic nasals, the situation was less consistent. There was a tendency for a moraic nasal to use separate switch points depending on whether it was accented or post-accented. When it was accented, the onset of the F0 rise usually started during the preceding vowel. When it was in post-accent position, the onset of F0 fall started around the [V-m] boundary. For the male speakers, the onset of the fall tended to go into the vowel as well. On the other hand, none of our data included the instance in which the onset of F0 change started considerably after the onset of [m]. When the moraic nasal is accented, and when it is spoken with fast speaking rate, the onset of the F0 rise started very early in the vowel, sometimes right from the onset.

4. DISCUSSION

In the current theories of phonology, two fundamentally different approaches have been proposed for the representation of the mora. One is the offspring of metrical phonology in which mora emerges from the branching syllable structure as in (a) [3]. The other approach, proposed by

Hyman, takes a mora as a prior necessary step to syllabification as in (b) [4].



Recent analysis of speech error and language game showed that neither syllable boundaries nor the notion of rhyme played an apparent role in Japanese[5][6]. The results of these studies also indicated that the nature of the mora in Japanese is like the one proposed by Hyman in which an onset and a nucleus are represented as exclusively forming a coherent single unit [4].

The results of the present study are also favourable for such representation since a CV mora and a moraic nasal showed close similarity in their tonal manifestation. There was a strong tendency for the four-mora test words to have the same duration and same F0 configuration regardless of their segmental and syllabic compositions within the same accent type. It seems that the mora is the most obvious unit by which Japanese utterances are regulated. The observation that the CV-CV boundary rather than the C-V boundary tended to be used as the switch point for F0 control, may be additional evidence for making a CV mora a coherent unit.

However, drawing clear boundary between the moraic nasal and the preceding mora may be less easy in some instances. While for a CV mora, the onset of F0 change tended to be timed with the CV-CV boundary, the timing of the moraic nasal and the F0 onset was less

consistent. This was most evident in fast speech when the moraic nasal was accented. In this position the onset of the vowel while the CV mora respected their boundary. It may imply that the association of mora and tone is less important within a syllable. Alternatively, it may mean that it is difficult to manifest pitch accent on the moraic nasal when time is shortened.

Another observation was the tendency for the second mora (both CV and N) to be prolonged in slow speech regardless of the accent type. This indicates that each mora is not proportionately prolonged in slow speech but rather that there is some kind of temporal organization at a higher level which takes place regardless of pitch accent. Further experiments are in preparation to test some of the findings in the present study.

REFERENCE

- [1] NAGANO-MADSEN, Y. & ERIKSSON, L. (1989), "The location of F0 turning point as a cue to mora boundary", *STR-QPSR 1/1989*, 41-45, Department of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm.
- [2] SAWASHIMA, M., HIROSE, H., & HONDA, K. (1980) "Relative timing of articulatory and phonatory controls in Japanese word accent: a study on nonsense two-mora words", *Ann. Bull. RILP*, 14, 139-147, University of Tokyo.
- [3] HAYES, B. (1989) "Compensatory Lengthening in moraic phonology", *Linguistic Inquiry* 20, 253-306.
- [4] HYMAN, L. (1985), "A theory of phonological weight", Foris: Dordrecht.
- [5] KUBOZONO, H. (1989), "The mora and syllable structure in Japanese: evidence from speech errors", *Language & Speech*, 32(3), 249-278.
- [6] KATADA, F. (1990) "On the representation of moras: evidence from a language game", *Linguistic Inquiry*, 21, 641-646.

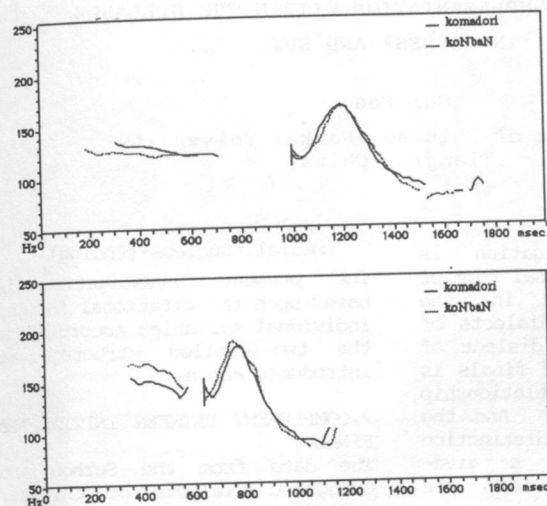


Fig. 1. F0 contours of /koŋbaN/ and /komádori/ in slow (above) and fast (below) speech.

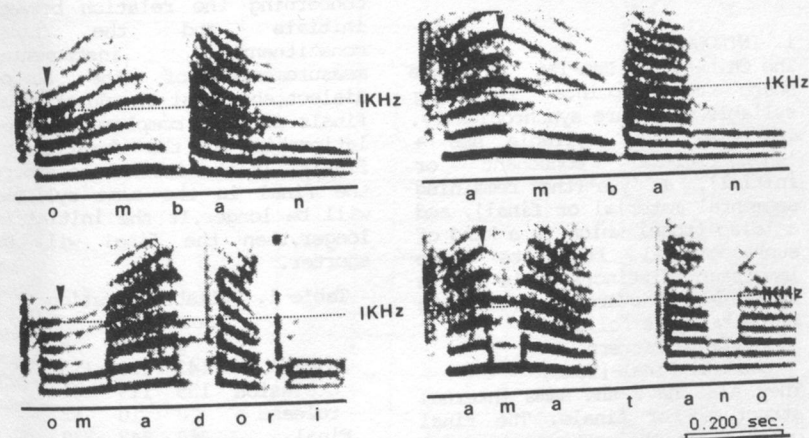


Fig. 2. Narrow band spectrograms of [(k)óm̄ban], [(k)omádori], [(k)ám̄ban], and [(k)ámatano] in slow speech by the female speaker. The arrow indicates the onset of F0 change.

**DURATIONAL COMPLEMENTATION WITHIN THE SYLLABLE
IN CHINESE AND SUI**

Shi Feng

Department of Chinese, Nankai University
Tianjin, China

ABSTRACT

The present investigation is based upon the citational form of individual syllables in the Suzhou and Guangzhou dialects of Chinese, and Zhonghe dialect of Sui. The duration of finals is in a complementary relationship to that of initials. And the long-short vowel distinction manifests itself in a give-and-take relationship with the ending within the final. The duration of stop endings is confirmed to be a part of the whole final although the closure is silent.

1. INTRODUCTION

The Chinese and Kam-Tai languages share many points regarding syllable structure synchronically. In both, each syllable has a 'sheng'(initial consonant or initial), a 'yun'(the remaining segmental material or final), and a 'diao'(tone) which is a kind of suprasegmental but its more important distinction is as a syllable feature. This is illustrated as follows:

Syllable Pattern
Initial-Final-[Tone]

They also have the same internal structure for finals. The final contains an obligatory main vowel as its nucleus. Sometimes a high vowel appears in front of the main one, serving as a medial. It is also called the 'final head'. Sometimes there is a consonant or a high vowel following the nucleus, serving as the final ending. Thus the typical structure of the final is as follows:

Final Structure

(Medial)-Nucleus-(Ending)

The present investigation is based upon the citational form of individual syllables according to the two-levelled structure we introduced above.

2.COMPLEMENT BETWEEN INITIAL AND FINAL

The data from the Suzhou and Guangzhou dialects of Chinese, and the Zhonghe dialect of Sui were selected to investigate the relationship between the different components in a syllable.

Concerning the relation between initials and the final constituent, instrumental measurements of the Suzhou dialect show that the duration of finals is in a complementary relationship to the duration of initials. If the initial is short, the final in the same syllable will be longer. If the initial is longer, then the final will be shorter.

Table 1. Syllable Duration (Suzhou)

	p	t	k
Initial	143	129	113
occlusion	135	119	101
release	8	10	12
Final	262	243	279
Syllable	405	372	392
	p'	t'	k'
Initial	204	179	160
occlusion	98	104	68
release	106	75	92
Final	226	185	211
Syllable	430	364	371ms

It is evident in Table 1 that aspirated stops are much longer

than unaspirated as regards initial duration whereas the finals following aspirated stops are much shorter than those following unaspirated. (Shi 1983) The same conclusion was also drawn from Pekingese. (Feng 1985)

It should be pointed out that initial and final are not equal in the complementation. The initial is the passive factor and the final the active. The duration of an initial is relatively stable and the duration of a final is easily variable. The cross-match test shows a final may alter its duration after different initials while an initial duration would only change a little before different finals. The difference between various initial durations is mainly due to their mode of articulation.

3. REPLACEABLE DURATION OF MEDIAL AND ENDING

Although there is a phoneme of zero initial in phonological analysis, it is in fact a glottal stop or a slightly voiced fricative at the beginning of the syllable. In the structure of finals, only the nucleus is indispensable. Both the medial and the ending are optional for some of the finals. And the nucleus will fill up the vacancy by extending its duration when the medial and/or the ending are absent in a final. Thus we call medial and ending replaceable and nucleus obligatory. As can be seen in Table 2, the duration of finals containing nucleus alone is approximately the same as those of finals involving a nucleus and a nasal ending.

Table 2. Duration of Finals (Suzhou)

	t	t'	k	k'
Syllable	378	359	387	379
Final	245	180	274	219
(=nucleus)	---	---	---	---
Syllable	369	365	394	356
Final	242	190	284	204
(=nucleus	91	61	72	68
+ending)	151	129	212	136ms

4. THE LONG-SHORT DISTINCTION OF VOWELS

In general, there is a long-short vowel distinction in Cantonese as well as Kam-Tai languages. However this long-short vowel distinction does not result in a durational distinction of syllables as a whole. The duration of main vowel, as the nucleus in a syllable, is quite different from the duration of syllable. We can balance the nucleus and the ending in duration in the same final. In the finals with a long vowel serving as nucleus, the ending is short; in those with a short vowel as nucleus, the ending is long. Therefore the duration may be either long or short for a main vowel in a final. However, in general, the duration of the two types of finals tends to be the same. (Ma & Luo 1962) The ending can play the role of adjustment in the duration of the whole syllable. The following measurements are from 10 pairs of syllables containing durational distinctions in the main vowel /a/ of Cantonese and Sui.

Table 3. Duration of Finals

	Cantonese		Sui	
	%	ms	%	ms
Vowel(L)	70	169	61	244
Ending	40	106	34	136
Final	110	275	95	380
Vowel(S)	46	116	33	130
Ending	46	116	72	284
Final	92	232	105	414

Comparing the average duration of the long vowels and the short ones, the long is in the ratio of 3:2 to the short in Cantonese, and the ratio in Sui is 2:1. The complement of the ending to the nucleus is obvious in Sui, but it is not so evident in Cantonese.

5. SYLLABLES WITH STOP ENDING

There is another kind of syllable in Cantonese and Sui, the entering tone syllable, which ends with a stop consonant such as /p/, /t/, or /k/. The long-short vowel distinction of these syllables in duration is as follows:

	Cantonese		Sui	
	%	ms	%	ms
Long Vowel	71	171	83	329
Short Vowel	44	112	25	126

The durational distinction is generally the same as those with unstopped endings in Cantonese, while in Sui the long vowels are much longer and the short even shorter. But what is the difference between the stop endings?

In general, the pronunciation of a consonant is divided into three steps: start, hold and release. It is difficult to measure the duration of stops because they do not release in the ending. As a substitution for this, the silence intervals from three informants' bisyllabic utterances were measured. This involved each entering tone syllable followed by a syllable with a voiced initial consonant in Cantonese. The following is the result:

Speaker	A	B	C	Average
Closure(with L)	106	124	107	112ms
Closure(with S)	123	153	109	128ms

Here are some individual variations in different informants. The closure duration following a short vowel is longer than that following a long vowel for A and B, but not C. However if we add the closure duration to the nucleus duration respectively, then the result will be roughly similar to that with an unstopped ending.

e.g. Final(L) 112+171=283 275ms
 Final(S) 128+112=240 232ms

Thus the closure duration of the stop ending should be considered as a part of the final duration although it is silent. The entering tone syllables are, thus, those that are interrupted with a period of silence. It is unreplaceable in the final duration.

Concerning durational relationships within the final, the long-short vowel distinction manifests itself in a give-and-take relationship with the ending. The final ending will be long if the main vowel is short and vice versa. In this way they reduce the

difference between the two kinds of finals in duration. Therefore we can say that, in citational forms, syllables of all types tend to have roughly the same duration, while their internal constituents vary in duration in a complementary way.

References:

- [1] FANT, G. (1969), Stops in CV-syllables, *STL-QPSR*, 4, 1-25.
- [2] FENG LONG (1985), The duration of initial, final and tone in Pekingese, in LIN TAO (ed.) *Experimental Works on Pekingese*, Beijing Univ. Press, 131-195.
- [3] LADEFOGED, P. (1975) *A Course in Phonetics*, Harcourt Brace Jovanovich, Inc..
- [4] LEHISTE, I. (1970) *Suprasegmentals*, M.I.T. Press.
- [5] MA, X-L. & LUO, J-G. (1962) The vowel duration in the Sino Tibetan languages, *ZGYW*, 5, 193-211.
- [6] OHALA, J. (1977) Production of tone, in FROMKIN, V. (ed.) *Tone: A Linguistic Survey*, Academic Press.
- [7] SHI FENG (1983), The acoustic features of voiced plosives in Suzhou dialect, *YYYJ*, 1, 49-83.
- [8] SHI FENG (1990) *Tone paradigms in Kam-Tai languages*, Doctoral dissertation of Nankai Univ.
- [9] SHI FENG (1991) *Tones and Stops*, Beijing Univ. Press.

PHONOLOGICAL INTERPRETATION OF F₀ VARIATIONS IN A BANTU LANGUAGE: KINYARWANDA.

Thierry Chambon

Institut de Phonétique, Aix-en-Provence, France.

ABSTRACT

In order to interpret what is transcribed in the various studies of tone in Bantu languages by finding some points of reference in the acoustic signal, and to investigate the interaction between lexical tonal patterns and phrase intonational patterns, a recorded corpus of Kinyarwanda has been processed with a signal edition software, with which we can better interpret the relationship between the patterns, using a sophisticated method to measure and manipulate their respective features.

Pitch variations in so-called tone languages have generally been represented by associating segmental or autosegmental tonal features with particular syllables or moras. By modeling F₀ curves of Kinyarwanda, we examine which acoustic variations, among the various representations of the melody of this language proposed by linguists, are considered pertinent. Then, we can suggest a method to reconstitute the structure of its prosodic constituents and to sketch a production model of F₀ variations.

1. PROCEDURE

1.1 Corpus

We selected a set of nominal words with dissyllabic radicals representing all of the syllabic and tonal patterns found in the previous literature on this topic. Their morphology is structured as follows:

i-bi-múga (the disabled persons)

a-ba-gaanga (the doctors)

pre-prefix + noun class prefix + radical

These words were combined into utterances following the pattern :

N1 + beéretse + N2 + N3

(N1 showed N2 to N3).

1.2 Informants and Recordings.

The language of informants was representative of the Kinyarwanda spoken in the southern part of Rwanda. These informants were recorded in an anechoic chamber with a professional recorder and microphone. Utterances were typed on cards following the current Kinyarwanda spelling, namely without any vocalic quantity or accent marks, and presented to informants in a random order, except for a set of utterances composed only of radicals bearing no lexical H tone, to which signs of pauses (#) were added. The informants were told to group the words according to these marks (which sounded very natural to them, since these pauses never cut any tonal unit – see 2 –, but rather organized the sentence into intonative units as they had naturally spoken them before).

1.3 Operating and Modelling of Data.

These recordings were transferred from their analogic support to a digital one, namely the mass-memory of a MassComp 16 byte minicomputer, with a 10 kHz sampling rate. Using VES, signal edition software designed by R. Espesser & O. Balfourier, we edited oscillograms and sonagrams representing each sentence, segmented and labelled them with syllable markers, and validated the segmentation by listening to each labelled segment. VES can also be used to construct F₀ curves. We edited each curve into MO-MEL (melody modeling software) and drew a superimposed F₀ curve by interpolating a quadratic spline function between target points located on F₀ peaks

and valleys. This procedure, described in [2] enabled us to separate from the raw curve the (prosodic) intentional variations and the non-intentional variations, which result from articulation and coarticulation constraints – cf.[1]. Electromyographic studies indicated that in two different words showing the same stress structure but different segments (e.g. French [papa] and [mamā]), the vocal fold tensor muscles are preprogrammed in the same manner. Thus, the curve represents the phonologically relevant variations and gives an indication of the vocal fold control, where the inflection points correspond to the maximum muscle activity. To validate these abstract F₀ curves, utterances were recorded and played back to the informants. They could not hear any difference between the recorded and the original signals.

2. EXTRACTING THE PROSODIC CONSTITUENTS

We labelled these F₀ targets, assigning H (High) to F₀ peaks, L (Low) to valleys, and D to downstepped peaks – cf.[2]. These labels should not be interpreted like the ones in traditional tonological descriptions as they are not tone features associated with any syllable or segment, but relative pitches that describe contours of prosodic constituents disregarding their segmental structure. According to [3], prosodic structure can be represented using the following constituents: Syllable, Tonal Unit (T.U.) and Intonative Unit (I.U.). In accordance with the previous studies on Bantu languages, we will treat the syllable structure as: V, CV (one mora) or CVV (two moras): u-mu-saambi (an old mat).

2.1 Tonal Unit

We assume for now that the T.U. is equivalent to the prosodic word, namely the unit that bears a lexical tone pattern (e.g. the radical and its extensions for short words). Then, the utterances in our corpus contain four T.U.: [abagabo]₁ [beéretse]₂ [abagore]₃ [ibimúga]₄. We first grouped the words according to the seven possible tonal and moraic patterns found in previous descriptions – [4], [6] and others. By comparing them with the acoustic realizations, we tried to find a correspondence between the various

transcriptions and discover the acoustic feature that was considered pertinent in each study (see 2.3).

2.2 Intonative Unit

Pauses are the most obvious clue in the identification of I.U., as the speakers group the utterance in one, two or three T.U. clusters. Moreover, in Kinyarwanda, a word final vowel is deleted when the following word begins with a vowel:

beéretse + abagabo = beérets'abagabo (they showed the men)

beéretse + abagaanga + ibimúga = beérets'abagaang'ibimúga

(they showed the doctors the disabled people).

We assumed that the lack of application of this rule between two T.U. implies that there is an I.U. boundary. In order to verify this and to find the possible I.U. patterns in our corpus, we proceed as follows:

1) determine the pattern of I.U. bounded by pauses;

2) check that this pattern applies to any unit not bounded by a final vowel deletion (or any relevant segmental rule for other languages);

3) use this template to determine possible I.U. boundaries in the contexts in which the final vowel deletion can not apply: abageenzi + beéretse... (the travellers showed...).

To segment the I.U. when there is no pause, we need to know both its tonal template and its interaction with the U.T. templates. T.U. with the [LL] pattern (e.g. abagabo) never bear any F₀ peaks in the final position or when the final vowel deletion rule applies (fig.1). But an F₀ peak always appears on the last syllable before a pause or the lack of application of the final vowel deletion rule (fig.2). From this, we induced that there are two patterns for I.U. in this corpus: final [LL], and continuative [LH].

2.3 Interaction between Levels of Representations

Any T.U. in an utterance is necessarily supported by an I.U. We observed the realization of each word in our corpus in the context of both types of I.U. The final contour is dependent on three parameters: the underlying (lexical) pattern of the radical, the I.U. pattern, and some possible

interactional and contextual rules (downstep). Our observations led us to extract only three possible T.U. patterns for this corpus: [LL], [HL], [LH]. The most interesting interactions observed are the ones involving an H in I.U. and T.U. simultaneously:

I.U.	[L H]	[L H]
T.U.	[LH]	[HL]
linearization	[LHH]	[LHLH]
Downstep1	[LHD]	[LHLD]
Downstep2 (optional)	[L H D]	

This downstep is generally linked to the rising of the intermediate L. The regression of this F_0 valley in relation to the distance between the two surrounding peaks (d) gives a negative correlation ($r = -0.609$, $a = 0.01$) and its value tends to meet the one of the second peak (to disappear) when (d) tends towards 150 ms. This explains the optional step of the rule and means that downstep seems to be triggered by a physiological rather than phonological constraint.

Without any H-interaction, we have the following linearizations:

I.U.	[L L]	[L L]	[L L]	[L H]
T.U.	[LL]	[LH]	[HL]	[LL]
lin.	[L L]	[LHL]	[LHL]	[LLH]

3. IDENTIFYING THE LEXICAL TONAL PATTERNS.

The difference between the [LH] and [HL] tonal patterns, which are both produced as [LHL] on isolated words, is difficult to evaluate because, according to our own and all previous research, there is no minimal pair illustrating such an opposition in Kinyarwanda. We grouped the words according to: a) the pattern [HL] or [LH] of the radical; b) the glissando on the first syllable of the radical (rising or falling); and we analysed in both cases the variance of the distance from the F_0 peak to the beginning of the word and of the radical. Results are significant in both cases. The difference based on the first syllable glissando seems to be the most

attested ($F = 51$, $p = 0.0001$), and this explains why this has been the relevant clue in the previous studies. But, if this opposition seems to be the marked one for perception, which is predictable since glissando on vowels is the relevant feature for tone perception (cf. [7]), it could be just a consequence of the word pattern variation which may be more relevant for a production model. To find out which is the relevant domain for this opposition, we plan to collect a corpus of words beginning with the same class prefixes (a-ba vs i-bi) and possibly with the same radical initial syllable, in order to avoid the variations of the intrinsic and co-intrinsic durations of segments. If we compare series of words grouped according to prefix duration and bearing the same tonal template, the variations of the position of the peak will reflect whether this one is directly linked to a syllable (or mora) or not.

4. CONCLUSION

Using this representation, we can extract from the acoustic shape what should be transcribed as lexical value variation and as intonative variation. Thus, it could be used for transcription of tone in field investigations. Furthermore, the intra/inter speaker variability and the adequacy of this model for our corpus lead us to question the association between a tone and a mora, and to assume that the melodic tier is independent of the segmental tier, so that we may not need association lines in phonological representations of tone or accent in this kind of language.

REFERENCES

- [1] DI CRISTO, A. (1978), "De la microprosodie à l'intonosyntaxe", Th. d'Etat, Université de Provence, Aix.
- [2] HIRST, D. (1983), "Structures and categories in Prosodic Representations", *Prosody, Model and Measurements*, Cutler & Ladd, Springer, Heidelberg, 93-109.
- [3] HIRST, D. (1988), "Tonal Units as Constituents of Prosodic Structure: the Evidence from French and English Intonation", *Autosegmental Studies on Pitch Accent*, Van der Hulst & Smith, Foris, Dordrecht, 151-165.

[4] JOUANNET, F. (1985), "Prosodie et Phonologie non linéaire", GERLA-SELAF, Paris.

[5] JOUANNET, F. (1989), "Modèles en tonologie (Kirundi et Kinyarwanda)", Editions du CNRS, Paris.

[6] KIMENYI, A. (1976), "Tone Anticipation in Kinyarwanda", *Studies in Bantu tonology*, SCOPIL 3, Hyman (ed.) 169-181.

[7] ROSSI, M. (1971), "Le Seuil de Glissando ou Seuil de Perception des Variations Tonales pour les Sons de la Parole", *Phonetica*, 23, 1-33.

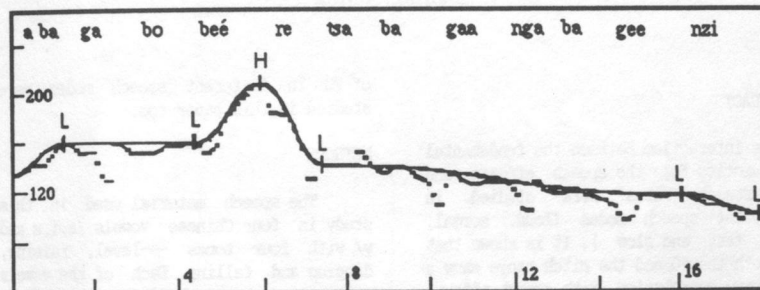


Fig. 1: Detected F_0 (dotted line) and modelled F_0 (continuous line). All the words, except the verb, bear the [LL] tonal template and are produced within one I.U.

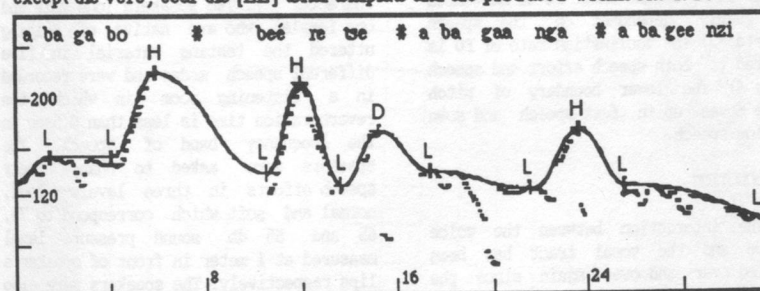


Fig. 2: same sentence as above produced by 4 I.U. bounded by pauses. An F_0 peak, downstepped if preceded by another one, appears at the end of each I.U.

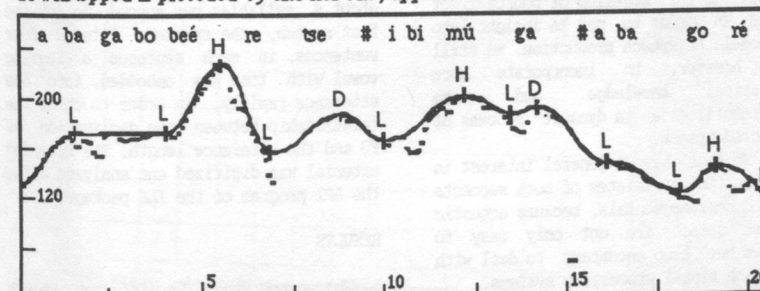


Fig. 3: sentence produced by 3 I.U.: a L rises in relation to the distance between the two surrounding peaks while the second one is downstepped. Note the anticipation of the H on /abagoré/, triggered by the [LL] template of the I.U.

INTERACTION BETWEEN SUPRASEGMENTAL FEATURES

Jialu Zheng

Institute of Acoustics, Academia Sinica
Beijing, China

ABSTRACT

The interaction between the fundamental frequencies F_0 , the speech efforts and the speech rates were studied in different speech modes (loud, normal, soft, fast and slow). It is shown that 1) Both the F_0 and the pitch range show a positive correlation with speech effort, but a rather very weak correlation in speech rate; 2) The onset F_0 of sentences are mainly dependent on the speech efforts; 3) The declination rate of F_0 is related to both speech effort and speech rate; 4) The lower boundary of pitch range moves up in fast speech and down in slow speech.

INTRODUCTION

The interaction between the voice source and the vocal tract has been studied over and over again since the beginning of this century. And some recent work [1] has laid the theoretical foundations for modeling of source-tract system. In order to get an insight into the model of speech production, we still need, however, to incorporate more theoretical knowledge and more experimental data in dynamic process of connected speech.

We thus have a general interest in the acoustic correlates of both segments [2] and suprasegmentals, because acoustic data of speech are not only easy to acquire but also necessary to deal with in speech signal processing systems.

This paper aims at examining the interrelation between the fundamental frequency F_0 , the speech effort and the speech rate, which are more important in suprasegmental level. And the declination

of F_0 in different speech modes were studied in this paper too.

METHODS

The speech material used in this study is four Chinese vowels /a, i, u and y/ with four tones—level, raising, dipping and falling. Each of the vowels with tone was embedded in a frame sentence "wǒ dú zì" (I utter the character _). Two speakers (one male and one female) who are native of Beijing uttered the testing material in five different speech modes and were recorded in a listening room in which the reverberation time is less than 0.5sec in the frequency band of speech. The speakers were asked to change their speech efforts in three levels—loud, normal and soft which correspond to 75, 65 and 55 db sound pressure level measured at 1 meter in front of speaker's lips respectively. The speakers were also asked to speed up and to slow down the speaking rate. The normal speech rate is about 4 syllables per second. As for the fast speech, the speakers uttered four sentences, in each sentence a testing vowel with tone was embedded, into one utterance rapidly, in order to check the relationship between the declination of F_0 and the utterance length. The recorded material was digitized and analyzed using the API program of the ILS package.

RESULTS

1. The pitch range in different speech modes

The pitch range is determined between the highest and lowest fundamental frequencies of the utterance. The highest

F_{0max} and the lowest F_{0min} of the testing vowels with tones and of the whole frame sentences were measured individually. And the average pitch range of a vowel over four tones $\Delta F_0 = F_{0max} - F_{0min}$ (the upper figures), and the relative pitch range F_{0max}/F_{0min} (the lower figures) in different speech modes are listed in Table 1. F_{0max} and F_{0min} stand for the average highest and lowest F_0 values of a certain vowel over four different tones respectively. The pitch limit which was

defined as the range from the ceiling to the floor of the voice which can be reached in different speech modes by the speakers is much larger than the average pitch range and not given in Table 1.

It was a pity that the API program did not work well and no data were given in Table 1,3,5 for female loud speech.

From Table 1. it can be seen that 1) The pitch range is moved up and expanded as increasing the speech effort; 2) The pitch range is somewhat shrunken as

Table 1. The average pitch range of testing vowels in different speech modes in both $F_{0max} - F_{0min}$ (Hz) and F_{0max}/F_{0min} .

Vowel	Loud	Normal	Soft	Fast	Slow	Ave.
m	244-101	171-94	143-94	171-104	172-90	181-96
a	2.42	1.82	1.52	1.64	2.12	1.91
f		299-148	218-128	327-173	271-139	278-147
		2.02	1.70	1.89	1.95	1.89
m	280-105	199-102	143-97	183-113	193-95	199-103
i	2.67	1.95	1.45	1.62	2.03	1.93
f		330-150	222-128	329-204	287-146	293-157
		2.20	1.73	1.61	1.97	1.86
m	291-116	193-96	141-97	182-110	209-96	203-104
u	2.51	2.01	1.45	1.65	2.18	1.95
f		340-149	231-122	348-179	294-139	304-148
		2.28	1.29	1.94	2.12	2.05
m	288-122	194-100	141-95	190-113	196-93	202-105
y	2.36	1.94	1.48	1.68	2.11	1.92
f		315-150	234-126	362-187	294-148	301-154
		2.10	1.86	1.94	1.99	1.95
m	275-111	189-98	141-96	181-110	193-91	196-102
Ave.	2.49	1.93	1.48	1.65	2.11	1.92
f		321-149	226-126	341-186	286-143	294-152
		2.15	1.79	1.84	2.00	1.93

Note: m stands for male and f for female and Ave. for average.

Table 2. The average tone duration of vowels /a/ and /i/ v and the average frame sentence duration s in different speech modes (ms).

Tone	Loud		Normal		Soft		Fast		Slow		Average	
	v	s	v	s	v	s	v	s	v	s	v	s
m	272	1062	310	1111	298	1127	179	589	352	1530	282	1084
1 f	269	1168	250	1076	215	1053	144	599	365	1466	249	1072
m	301	1059	298	1050	317	1095	196	615	448	1654	312	1095
2 f	314	1175	256	1114	253	1047	157	589	397	1536	275	1092
m	282	1040	288	1082	295	1079	167	634	455	1596	297	1086
3 f	317	1187	262	1097	256	1040	125	525	362	1581	264	1086
m	256	973	266	998	266	1031	154	627	375	1501	263	1026
4 f	263	1123	250	1066	205	1028	147	570	352	1456	243	1048

Note: 1 stands for level tone, 2 for raising, 3 for dipping and 4 for falling.

increasing speech rate; 3) The pitch limit of a speaker is about double what the speech range in normal voice; 4) The intrinsic pitch of vowels are well kept up in any speech mode; 5) The pitch range of male voice and of female voice are nearly the same in relative scale and semitone is a good measure of pitch for developing speech processing systems.

2. The duration of testing vowels and frame sentences in different speech modes

The durations of the testing vowels and the frame sentences were measured on speech waveforms. The results show that the duration ratio of testing vowel to frame sentence is nearly invariant in different speech modes. The average duration of low vowel /a/ and high vowel /i/ with the same tone type is called the tone duration and listed in Table 2.

It is worthy to be noted that the rank order of the tone durations from long to short is raising, dipping, level and falling and it is well kept up in any speech mode. In addition, the same relation exhibited in sentence level too. The fact that tone duration shows some intrinsic feature is identical with the results which were obtained from a statistical analysis of a large vocabulary[3].

3. The declination of F0 in different speech modes

The declination of F0 is easy to determine using the frame sentence designed in this paper, because of the sentence initial syllable with dipping tone and the final syllable with falling tone. Both dipping and falling tones can provide a local minimum in pitch contours. Then the baseline can be easy drawn by connecting the two minima. What is an exception to this is that when a vowel with dipping tone is embedded in the frame sentence, then the sentence final F0 may be higher than the local minimum of preceding dipping tone. That is due to the sharpening rule running on sentence intonation. In this case the baseline passes through the two minima of dipping tones.

Table 3. shows the F0 declination value $\Delta F0$ and the sentence final fundamental frequency F0f in different speech modes. And the F0 declination value of the four sentence utterances and

Table 3. The F0 declination value $\Delta F0$ and the sentence final fundamental frequency F0f for different vowels in different speech modes (Hz).

Vow.	Loud		Normal		Soft		Fast		Slow	
	$\Delta F0$	F0f	$\Delta F0$	F0f	$\Delta F0$	F0f	$\Delta F0$	F0f	$\Delta F0$	F0f
a m	51	103	19	95	14	88	26	96	27	82
f			37	152	21	136	42	142	27	146
i m	63	105	27	102	8	96	18	90	25	95
f			49	159	25	147	27	146	23	145
u m	51	116	23	96	7	98	28	92	23	95
f			52	150	28	139	46	140	29	139
y m	46	124	20	100	7	96	28	90	23	94
f			42	151	27	150	59	144	15	149
Av.m	53	112	22	98	13	95	26	92	25	92
f			45	153	25	143	43	143	23	145

Note: Av. stands for average.

Table 4. The F0 declination value $\Delta F04$ of the four sentence utterance and $\Delta F01$ of the last sentence and the utterance final fundamental frequency F0f for different vowels in fast speech (Hz).

Vowel	a		i		u		y		Ave.	
	m	f	m	f	m	f	m	f		
$\Delta F04$	40	64	40	61	58	100	51	81	44	77
$\Delta F01$	26	42	18	27	28	46	28	59	25	44
F0f	96	142	90	146	92	140	90	144	92	143

of the last sentences and the utterance final F0 were shown in Table 4.

From Table 3. and Table 4. it can be seen that 1) The sentence final F0 value is invariant even for the four sentence utterance of fast speech; 2) The F0 declination values are speech mode dependent, the more effort has been made the more F0 declination value appears; 3) There is no distinct difference between male and female in relative value of F0 declination.

4. The onset F0 value of sentences in different speech modes

Table 5. shows the measurement results of the onset F0 of sentences in which different vowels were embedded in different speech modes.

Table 5. indicates that the onset F0 of sentences showed an obvious correlation with speech efforts but it is

Table 5. The onset F0 value of the sentences (Hz).

vowel	loud		Normal		Soft		Fast		Slow		Average	
	m	f	m	f	m	f	m	f	m	f	m	f
a	154		123	212	115	185	135	221	124	198	130	204
i	168		138	224	120	182	135	209	136	191	139	201
u	167		128	221	117	174	130	229	138	199	136	206
y	170		133	216	125	176	126	225	136	195	138	203
Ave.	165		131	218	119	179	132	221	134	196	136	204

independent of speech rates for both male and female. That means the onset F0 of sentences are mainly controlled by the subglottal pressure and laryngeal tension.

The average onset F0 value of sentence for a certain vowel over different speech modes were listed in last two columns in Table 5.

DISCUSSION

From Table 1. and 5. it is indicated that the speech efforts showed a good positive correlation with both the pitch ranges and the onset F0 of sentences, whereas speech rates did not exhibit a significant relation. The F0max of the average pitch range for normal, fast and slow speech are nearly the same, because the speech level of them are equal. It appears that, speaker can manipulate the voice source and the vocal tract independently. The F0max of pitch range and the onset F0 of sentence are mainly controlled by subglottal pressure incorporated with laryngeal tension.

As for the F0 declination and the physiological process underlying it, the $\Delta F0$ and F0f in different speech modes showed that the sentence final F0 is invariant but the sentence initial F0 is speech effort dependent and speech rate independent. So the declination rate is related to sentence length. In the case of four sentence utterance in fast speech, however, the F0 declined at different rates in different parts of the utterance. F0 declination value of the last sentence is above 50 percent of the total $\Delta F0$ of the four sentence utterance. That means speaker can in some degree control the subglottal pressure to match the syntactic structures of the utterances. So that the F0 declination may be a passive phenomenon mixed with some active speaker control process.

It is worth notice that the intrinsic

pitch of vowels and the intrinsic duration of tones are well kept up in different speech modes and showed somewhat effect on sentence level.

CONCLUSIONS

1. The fundamental frequency and the pitch range of speech show a strongly positive correlation with speech effort, but rather a very weak correlation in speech rate.
2. The onset F0 of sentences are mainly dependent on speech efforts and independent of speech rates.
3. The sentence final F0 is invariant in different speech modes, and the F0 declination rate is closely related to speech efforts and sentence lengths.

ACKNOWLEDGMENTS

The author would like to thank Prof. A. Fourcin for his encouragement and collaboration in study of voice characteristics. This work was supported by the National Project 863 of China.

REFERENCES

- [1] Fant, G. (1982), "Preliminaries to analysis of the human voice source", STL-QPSR 4/1982, 1-27.
- [2] Zhang, Jialu (1987), "The intrinsic fundamental frequency of vowels and the effect of speech modes on formants", Proceedings of 11th ICPHS (Tallinn), Sy 3.5. 1-4.
- [3] Zhang, Ning (1986), "On the relation between the duration and the phonetic structure of syllables", Master thesis, Institute of Acoustics, Academia Sinica.
- [4] Carole, E. and et al. (1983), "Is declination actively Controlled?", in "Vocal fold physiology" edited by Titze, I.R. and Scherer, R.C., Colorado: The Denver Center for the Performing Arts, Inc.

DOWNSTEP ET DOWNSTEP

Annie RIALLAND

CNRS, URA 1027
Paris III

ABSTRACT

Downstep, a lowering of successive high tones or tonal accents, is often conditioned by intervening low tones. This paper discusses a second type of downstep in which intervening low tones are absent. We will show that by recognizing this second type, we can improve the analysis of certain prosodic systems, in particular the challenging example of Tonga (a Bantu language). We will also show that this type of downstep plays a role in other types of prosodic systems such as that of French.

Le concept de downstep est habituellement associé à celui de ton bas. En fait, il existe un deuxième type de downstep que nous avons qualifié d'"intégrateur" et qui rabaisse également des tons hauts ou des accents tonals mais sans ton bas intermédiaire. Il s'applique à des tons hauts ou des accents appartenant à un domaine donné et marque leur intégration à ce domaine. L'identification de ce deuxième downstep est récente dans la recherche africaniste et sa reconnaissance peut améliorer l'analyse de systèmes prosodiques de nombreuses langues, surtout de langues à accent tonal. Nous montrerons, dans cette communica-

tion, comment la substitution de cette deuxième forme de downstep à la première permet de mieux comprendre des faits prosodiques du tonga, langue bantoue qui a déjà donné lieu à de nombreuses analyses. Une fois ce downstep "intégrateur" identifié dans des langues à accent tonal, nous verrons que le même downstep peut être reconnu dans des langues sans ton ni accent mais à intonation comme le français.

1. UNE REANALYSE DU TONGA

La première analyse de langue africaine faisant usage de la notion de downstep sans ton bas est sans doute celle du somali par Hyman (1981). Dans le domaine bantou, cette entité a été d'abord introduite par Odden dans son étude du kishambaa (1982) et utilisée ensuite par E. Leeung pour le Llogoori (1986). Nous verrons ici comment elle peut transformer l'approche des faits tonga.

Le tonga a donné lieu à de nombreux travaux et notre analyse sera confrontée avec celle des auteurs précédents, en particulier Goldsmith (1984) et Pulleyblank (1983) qui eux-mêmes doivent beaucoup aux auteurs plus anciens, en particulier à H. Carter

(1962, 1971, 1972) et Meeussen (1963). Nous considérerons d'abord des formes verbales (formes fortes et faibles du passé récent), ensuite des formes nominales.

1.1. Formes verbales "faibles" et "fortes"

Les deux formes présentent des différences prosodiques et correspondent à des focalisations différentes. Exemples:
forme forte du verbe

ndà ká tólà nyàmà
"je PRENDS de la viande"

forme faible

ndà ká tólá nyàma
"je prends de la VIANDE"

Tous les auteurs s'accordent pour reconnaître des syllabes déterminantes selon le terme de Meeussen (1968) mais les analyses sont soit tonales, soit accentuelles, ou encore composites.

1.2. Comparaison entre les analyses

Pour comparer notre analyse avec celle des auteurs précédents (Goldsmith 1984 et Pulleyblank 1983), nous utiliserons la deuxième phrase citée précédemment, avec comme objet focalisé, le nom nyama, en lui-même dépourvu d'accent ou de ton haut.

a) Analyse de Goldsmith (1984)

Goldsmith pose des accents marqués par une mélodie HB et propose une dérivation qui d'accentuelle devient tonale. La règle de formation du downstep est tonale: le downstep résulte du déliage d'un ton bas comme suit:

* * *
nda ka tola nyama
| | |
H B H B H B
(* notant l'accent)

b) Analyse de Pulleyblank (1983)

Pulleyblank rejette tout niveau accentuel et propose une analyse entièrement tonale. Le downstep est également engendré par un ton bas devenu flottant en cours de dérivation.

nda ka tola nyama
| | | |
B H B H B B

c) Notre analyse

Notre analyse est complètement accentuelle et la forme faible avec son complément d'objet sont traités comme un *seul groupe accentuel*.

Entre les accents tonals venus en contact après l'application de règles accentuelles, le downstep se forme indiquant l'intégration des accents tonals au groupe accentuel.

* *
nda ka tola nyama
| |
H H

réalisé donc avec downstep entre les deux accents.

1.3. downsteps en cascade

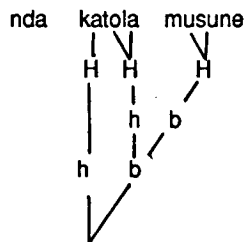
Si un groupe accentuel comporte plusieurs accents, chaque accent sera rabaisé par un downstep: on peut donc trouver des cascades de downsteps comme dans l'exemple suivant:

ndà ká tólá Imúsúnè
"j'ai pris un BOEUF"

qui correspond à un groupe accentuel unique. Après application de diverses règles, trois accents tonals, viennent en contact.

* * *
nda ka tola musune
| | |
H H H

Les downsteps intègrent les trois tons hauts à l'intérieur de ce groupe accentuel et les organisent selon une hiérarchie qui peut être représentée par l'arbre suivant :



(où h (haut) et b (bas) sont relatifs l'un par rapport à l'autre : b placé après h signifie "plus bas" que ce qui précède).

1.4 La rencontre de deux groupes accentuels

Un verbe à la forme forte constitue, en lui-même un groupe accentuel et l'objet qui le suit forme un second groupe accentuel comme dans l'exemple suivant :

ndàpá músúnè "j'AI DONNE un boeuf"

Les deux accents successifs n'appartenant plus à la même unité accentuelle ne sont plus hiérarchisés par le downstep : ils sont réalisés sur la même hauteur.

1.5. Analyse de syntagmes nominaux

Les composés forment aussi une seule unité accentuelle, qu'ils proviennent de reduplication comme dans les exemples suivants :

í cǐ ífúmó fúmò "un matin"

í mú ífúbá fúbà "un fou, un sot"

ou qu'ils soient des "complex" selon le terme de Carter :

í kúbó!kókwámùsánkwa
"le bras d'un garçon"

í búl mwázibwàhòmbè

"une maladie du bétail"

Chacun de ces syntagmes constitue un "complex" et une unité accentuelle unique qui forme un seul domaine pour le downstep.

Faute de place, nous ne pouvons ici montrer que l'analyse avec downstep intégrateur fonctionne pour l'ensemble du système du tonga, mais le lecteur pourra se reporter à A. Rialland, 1988.

1.6. Conclusion

L'analyse du tonga se trouve donc fortement simplifiée par l'utilisation de la notion d' downstep hiérarchisateur. Elle se trouve aussi éclairée : la formation du downstep cesse d'être aléatoire mais prend un sens puisqu'elle devient un processus intégrateur permettant de former des unités plus larges.

2. LE DOWNSTEP INTEGRA - TEUR DANS D'AUTRES LANGUES

Cette hiérarchisation des accents par le downstep dans un groupe donné se rencontre dans d'autres langues, entre autres le somali et le japonais.

En somali par exemple selon Hyman (1981), tous les accents tonals inclus entre certaines frontières sont abaissés les uns par rapport aux autres.

Ex: wiil - ka % ma dilayo
13 3 1 24 3

"le garçon ne (le) frappe pas"
où % représentela frontière et les chiffres, les hauteurs respectives des syllabes.

En japonais, langue également à accent tonal et à plateaux tonals. le downstep intègre les accents tonals des groupes intermédiaires (en angl. intermediate phrases), c'est-à-dire des groupes qui se

situent entre le mot et la phrase (Beckman et Pierrehumbert, 1988).

On retrouve donc des groupes d'une dimension comparable à ceux que caractérise le downstep en tonga ou en somali et le découpage de ces groupes est, dans les trois langues sensibles au focalisations.

Ce downstep intégrateur se rencontre également dans des langues à "stress" telles que l'anglais ou même sans accent comme le français (Nous considérons le français comme langue avec allongement final de constituant mais sans accent). Il a été décrit en anglais sous les noms de downstep ou de catathesis (Pierrehumbert et Beckman, 1988, entre autres). Il peut être aussi reconnu en français où le downstep s'applique non plus à des accents mais à des tons hauts, marqueurs de continuation, indique leur appartenance à une même unité prosodique. Comme dans les langues à accent tonal, la continuité du downstep est remise en cause par les focalisations.

3. CONCLUSION

Nous avons montré dans cette communication que la reconnaissance du downstep "intégrateur" - que nous avons substitué au downstep dû à un ton bas posé par les analyses précédentes- permettait d'améliorer l'analyse des faits du tonga et de les éclairer. Dans cette langue à accent tonal, le downstep a pris un sens puisqu'il est devenu le processus intégrant les accents appartenant à une même unité accentuelle. Cette forme de downstep, encore insuffisamment reconnu dans les langues africaines, peut aussi être rapprochée d'autres downstep ou

catathesis à valeur intégrative que l'on peut dégager aussi bien dans des langues à accent tonal (somali ou japonais) à accent non tonal (anglais) que dans des langues sans accent comme le français.

4. BIBLIOGRAPHIE :

- (1) CARTER H., 1962, *Notes on the Tonal System of Northern Rhodesian Plateau Tonga*, Her Majesty s' Stationary Office, Londres
- 1971, *Morphotonology of Zambian Tonga: Some Developments of Meeussen's system I*, *African Languages Studies* n°12 pp. 1-30
- 1972, *Morphotonology of Zambian Tonga: Some Developments of Meeussen's system II*, *African Languages Studies* n°13 pp. 1-30
- (2) GOLDSMITH J., 1984, *Tone and Accent in Tonga dans Autosegmental Studies in Bantu Tones*, Clements G. et Goldsmith N. ed, Foris, Dordrecht
- (3) HYMAN L., 1981, *Tonal Accent in Somali*, *Studies in African Languages* n°12, pp 169-203
- (4) LEUNG E., 1986, *The Tonal Phonology of Llogoori*, M.A. thesis, Cornell University
- (5) MEEUSSEN A., 1963, *Morphotonology of the Tonga Verb*, *Journal of African Languages* vol 2 pp. 72-92
- (6) PIERREHUMBERT J. et BECKMAN M., 1988, *Japanese Tone Structure*, Linguistic Inquiry Monograph 17, M.I.T. Press, Cambridge USA et Londres
- (7) PULLEYBLANK D., 1986, *Tone in Lexical Phonology*, collection: *Studies in Natural and Linguistics Theory*, Reidel, Dordrecht
- (8) RIALLAND A., 1988, *Systèmes prosodiques africains: fondements empirique pour un modèle multi-linéaire*, thèse d'état, Université de Nice

DEPHONOLOGIZATION OF SYLLABIC INTONATION
IN LITHUANIAN URBAN SOCIALECTS

L. Grumadienė

Institute of the Lithuanian Language,
Vilnius, Lithuania

ABSTRACT

Under the influence of present-day vowel length dephologization the tendency of dephologization of syllabic intonation has arisen in Lithuanian urban socialects. The basis of syllabic intonation realization is being lost. As to the vocalic and mixed diphthongs they undergo essential qualitative changes.

1. INTRODUCTION

The tendency of dephologization of vowel length in Lithuanian urban socialects is rather obvious, though it is not typical to any Lithuanian dialect /A/. The results of the experimental investigation show that only 4-24% of unstressed long vowels in the word final position are being realized in the fluent speech. For instance, in the words: vyrų 'men' Gen. pl., kėpa 'dune' Acc. sng., kandyš 'moth' Nom. pl., lėlė 'doll' Acc. sng. etc. The same phenomena at the beginning and in the middle of a word equals to 4-17%. For instance, ažuolynas 'oak-wood' Nom. sng., rūkyti 'to smoke', rašymas 'writing' Nom. sng. etc.

The most specific feature of present-day urban Lithuanian is the lengthening of

short stressed final vowels which was found in 9-24% cases. This may be viewed as phonetic specification of approaching phonological change, which is possible in typological field of long and short vowel opposition. For instance, vaikų 'children' Acc. pl., maži. 'small' Nom. pl., namė. 'in the house' Loc. sng. etc. The substitution V → V. / [-stressed] as far as now is not found in any Lithuanian dialect, nevertheless it is predicted from the phonological structure of the Lithuanian language. Columnal morphemic accentuation becomes more and more accepted. The reason of this may be both the factor of analogy and uniformity of syllabic intonation /3/. For instance, ryšiai 'relation' Nom. pl.; Standard form - ryšiai; idomus 'interesting'; idomus; spjuviu 'spit' Instr. sng.; spjuviu; sėslius 'settled' Acc. pl.; sėslius; pilka 'grey' Nom. sng.; pilka; mūses 'fly' Acc. pl.; mūses; etc. Though, there were only 11-18% of the above mentioned cases, it must be noted, that such changes take place in spite of Saussure-Fortunatov law, which is based on the opposition of syllabic intonation. One should say that

hypercorrection becomes a rather common phenomena. The extraneous factor in this change is speakers or their parents origin from different dialects, who came into closer contact as communication improved, especially in the second half of our century. We must not ignore the fact that language is a system of auditory signs, and that language learners acquire the phonology of their language by ear/1/. It would seem that, when a learner becomes aware of differences between his speech and the speech of the others, there are two ways in which he can adjust his grammar. He can revise his analysis of the linguistic units in question so that his grammar will naturally produce the desired output. Or he can devise ad-hoc rules to cover up the inadequacy of his analysis. The results are difficult to foresee. The changes which undergo in Lithuanian urban socialects are greatly resulted by the intertwining of representatives of various dialects. For instance, the syllabic intonation of western Lithuanian dialects is of dynamic origin, while that of eastern ones is of quantitative origin/2, 6/. As a result of the tendency of dephologization of length of vowels emerged the tendency of dephologization of syllabic intonation. The main feature of vowel length opposition of some Lithuanian dialects is tenseness, while of the other ones - duration. The intertwining of representatives of different dialects in urban sphere allows some of them to accept the pronunciation V. (half-long

stressed vowel) instead of V̄ (long circumflex vowel) and V̇ (long acute vowel) as a normal one. In this way the basis of syllabic intonation realization is being lost. This is obviously seen in stressed monophthongs. As to the vocalic and mixed diphthongs they undergo essential qualitative changes, which greatly differ under the influence of dialects. Such variety leads to rule generalization.

That's why the realization of syllabic intonation is more distinctive in the syllables with diphthongal nucleus. Though, essential qualitative changes appear in such diphthongs, they may greatly differ as in some dialects the first component of diphthongs is more important while for other dialects - the second component.

2. SPECTRAL ANALYSIS

The spectral peculiarities of acute and circumflex diphthongs /ai/, /au/, /ei/ have been investigated. The experimental corpus consisted of 8 similar word pairs in which the diphthongs under the investigation are between voiceless consonants: kaišo - paišo, taiko - kaiko, kaušo - kaušo, keikia - peikia, auk - auk, taiko - paiko, keik - peik, keikė - peikė. The spectral analysis was made according to the computer program compiled by prof. V. Undžėnas at Vilnius University. The results obtained show that: a/ the quality of 1st component of stressed acute and circumflex diphthongs differs significantly, b/ the durational difference of such diphthongs is less significant.

F_1 and F_2 characteristics of first components of acute diphthongs are more like to F_1 and F_2 corresponding monophthongs / ɛ /, with circumflex being more different (tables 1-3). For instance, F_1 and F_2 of the 1st component of / a^* / in the word 'kaĩto' (table 1) is between / ɔ / ($F_1=400$ Hz, $F_2=900$ Hz; compare with final / ɔ /, where $F_1=400$ Hz, $F_2=1200$ Hz) and / e / ($F_1=600$ Hz, $F_2=1800$ Hz). "Normal" / a^* / has $F_1=800$ Hz, $F_2=1200$ Hz; compare with the first component of the word 'taĩko' - $F_1=850$ Hz, $F_2=1500$ Hz. F_1 and F_2 of first component of / e^* / (with circumflex) in the word 'peĩkia' resembles to those of the monophthongs / e /; / au / (with circumflex) in the word 'kaũšo' - resembles to / a /.

The comparison of the results of present investigation with the ones of previous Lithuanian dialectal investigations carried by other linguists show that qualitative difference of the first component of acute and circumflex diphthongs is not a unique phenomena in Lithuanian. There are some changes of the same kind in eastern dialects where syllabic intonation of quantitative origin. This is found especially in peripheral dialects. For instance, the pronunciation of 'kiaũlė' is [k'qũ.ĩ.ĩ.], but of the word 'kiaũras' - [k'zũras], or like that from Anykšćiai - a word 'kãrtis' they pronounce [kãr'tis], but a word 'vaĩgas' - [vaĩ.gas], while in a word 'laũkas' the pronunciation is [lãũkas], in 'laũžas' it is [lãũžas]. The changes in western dialects are of different character: the vowels of the

second component of stressed diphthongs in western dialects are narrower than those of eastern, for instance, [u] in the word 'jãutis' in western dialects it is pronounced like [jãutis] while the pronunciation with / a^* /, [a*] is met only in eastern dialects /5/. Because of this speakers from eastern dialects living in urban sphere may take acute syllables for circumflex ones. The opposition of syllabic intonation is realised not only by tone or intensity modulation changes, but by the place of stress contrast too. The fact that acute diphthongs are longer than circumflex ones in the urban sociallect may be explained by the qualitative characteristics of stressed diphthongs.

3. CONCLUSION

The tendency of columnal stress, the qualitative changes of stressed diphthongs and hypercorrection show the tendency of dephonologization of syllabic intonation opposition in the Lithuanian urban sociallect. The dephonologization of long and short vowel opposition is the main condition for the above mentioned dephonologization. Due to it the quantitative word stress is being formed. There is no opposition of stressed short syllables in Lithuanian. It is typical only for stressed long syllables. Though at present time this opposition is disappearing in the long syllables in the Lithuanian urban sociallects, what leads to the disappearing of syllabic intonation, the opposition of them.

4. REFERENCES

- 1/ ANDERSEN, H. (1978), "Abductive and deductive changes", Readings in historical phonology/ed. by Ph. Baldi, R.N. Werth, Pennsylvania Univ. Press, 313-348.
- 2/ GIRDENIS, A. (1974), "Prozodinės priegaidžių ypatybės šiaurės žemaičių tarmėje", Eksperimentinė ir praktinė fonetika, 160-198.
- 3/ GRUMADIENĖ, L., STUNDŽIA B. "Dinamika opozicijų slogovych intonacijų v fonologičeskich sistemach dialektnoj i gorodskoj reči", Proceedings of Ith ICPHs, 5, 95-98.
- 4/ GRUMADIENĖ, L. (1989), "Ilgųjų ir trumpųjų balsių priešpriešos nykimas miestiečių lietuvių kalboje", Baltistica, III(2) Addition, 292-296.
- 5/ JAVNIS, K. (1908-1916), "Grammatika litovskago jazyka. Litovskij original i russkij perevod", Petrograd.
- 6/ KAZLAUSKAS, J. (1968), "Lietuvių kalbos istorinė gramatika", Vilnius, 13 etc.
- 7/ PAKERYS, A. (1986), "Lietuvių bendrinės kalbos fonetika", Vilnius; PAKERYS, A. "Lietuvių bendrinės kalbos prozodija" (1982), Vilnius.

TABLE of F_1 and F_2 in Standard Lithuanian (according to /7/):

/i*/	$F_2=2300$ Hz $F_1=300$ Hz
/e*/	$F_2=2100$ Hz $F_1=350$ Hz
/e/	$F_2=1800$ Hz $F_1=600$ Hz
/a/	$F_2=1200$ Hz $F_1=800$ Hz
/o*/	$F_2=900$ Hz $F_1=400$ Hz
/u*/	$F_2=600$ Hz $F_1=300$ Hz

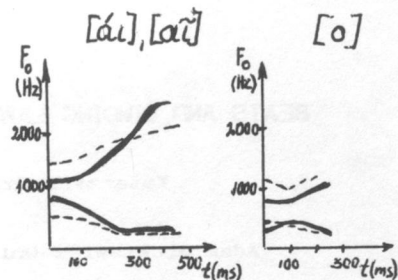


Fig. 1 Schematic spectrogram of 'taiko' - - - - -
'kaĩto' - - - - -

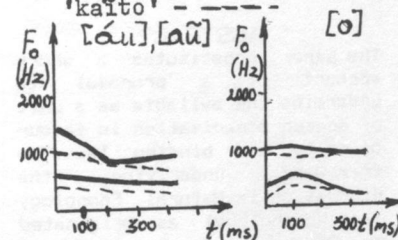


Fig. 2 Schematic spectrogram of 'kaũšo' - - - - -
'kaũšo' - - - - -

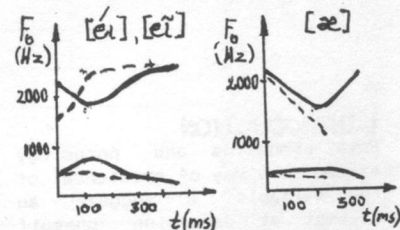


Fig. 3 Schematic spectrogram of 'keĩkia' - - - - -
'peĩkia' - - - - -

BEATS AND BINDING LAWS INSTEAD OF THE SYLLABLE

Katarzyna Dziubalska-Kořaczyk

Adam Mickiewicz University, Poznań, Poland

ABSTRACT

The paper constitutes a short account of a proposal to undermine the syllable as a unit of speech organization in favour of beats and binding laws. The framework underlying the discussion is Natural Phonology (and Morphology) as originated by Stampe and Donegan and developed by Dressler. A brief critical discussion of the syllable is conducted and followed by a presentation of the paper hypothesis.

1. INTRODUCTION

Both phonetics and phonology assume some way of existence of the syllable. I suggest an attempt at detaching oneself from a tradition, cherished for centuries, to acknowledge some form of the syllable as useful and indispensable in phonetic/phonological description. When looked at from outside and with a sufficient distance to obtain objectivity, what used to be called a syllable may turn out to be an unnecessary and mistaken complication of the already necessarily complex description of the speech chain. The problem as set above does not qualify for a paper-size discussion. Basically, then, I will shortly present here my counter-proposal to the syllable.

2. THE FRAMEWORK

The discussion is conducted in the framework of Natural Phonology (and Morphology) (cf. [3]).

Thus, firstly, the criteria and explanations I propose with reference to the segmental, prosodic, semantic/semiotic and lexical/morphological levels of language are of a functional nature. Secondly, whenever terms like "principle", "law" or "rule" are mentioned, they are to be understood as universal or language-specific preferences, and not (!) absolute generalizations. I constrain the existence of the latter (one consequence of which is avoiding the notion of exception) to certain language-specific "intensifications" of universal preferences.

3. THE SYLLABLE ?

If recognized as an identifiable entity, the syllable needs to possess some unity, constituent structure and boundaries.

As to the unity of the syllable, there exists phonetic evidence for a certain stability of consonantal transitions to and from vowels rather than for a stability of the whole (?) syllable (cf. e.g. [5]). Speech error evidence seems to demonstrate a greater cohesiveness of a VC sequence as opposed to a CV one (cf. [2]), while CV is, at the same time, generally acclaimed to be a basic syllable structure.

No matter a great variety of types of constituent structures

posited for the syllable, constituents tend to get organized according to the scale of sonority. However, the requirement for particular sonority slopes appears to be often violated by the languages of the world. To retain the syllable, "rescue strategies" are then introduced, e.g. Rubach and Booij (cf. [7]) would assume that an edge consonant (word-edge) does not count for a sonority slope. Doesn't this move make sonority useless? To Sievers (cf. [8]) consonants violating the expected gradation of "Schallfülle" formed the so called "Nebensilben" dominated, however, by "Hauptsilben". Two kinds of the syllable having different status - isn't it a complication? As a unit, the syllable needs to possess determinable boundaries. Boundary placement, or, in other words, division into syllables (of words or longer stretches of speech), however, turns out not to be a straightforward procedure. Available hints come from, basically, two very different sources: first, speakers' ability to "syllabify", second, the application of some phonological processes in the "syllable domain". But are they really hints for "syllable boundary placement"? As for the former source, what speakers are able to do is to distinguish in the flow of speech those sounds which are more prominent against the less prominent background, and the chunks that arise in this way are listed in the form parallel to counting. The problem with the latter source concerns the circularity of argumentation it introduces which entails arbitrariness of boundary placement: one and the same process may both condition and be conditioned by the syllable boundary (e.g. a tense vs. lax vowel opposition in

English, or "syllable final" devoicing in German). Reliably enough, both speakers and phonological processes have access to words, on one hand, and to feet, on the other. Access to words is guaranteed by the existence of a lexicon; access to feet - by the fact that it is impossible not to act rhythmically (cf. [1]). A functional unit of phonology which is smaller than a word, and which shows its accessibility better than a syllable, is a beat.

4. HYPOTHESIS

I suggest that the notions of a beat, word and foot as well as morpheme suffice to make it possible for the functions of the syllable to be accounted for without maintaining it as a unit. A basic speech skeleton consists of regularly recurring beats. Beats are primary, (preferably) vocalic figures against the consonantal ground. They are preferably vocalic due to the saliency potential inherent to vowels, although consonants might take over a beat function in a number of circumstances (cf. below). Inter-relationships between beats and pre-beat and post-beat consonants are specified by a set of binding laws which look both at a "micro-level" - constituted by a single beat and consonants surrounding it, and at a "macro-level" - governed by rhythm. Consonants clustering between beats coexist according to the preferred order as well. A universal preference for isochrony is rooted in universal principles of human behaviour which are reflected in one statement: it is impossible not to act rhythmically (cf. 3. above). In speech, an underlying organizational principle predicts a default tendency for equal time intervals between beats. The latter tendency is realized in different degrees

and modified versions to give a variety of typological and language-specific distinctions among particular tongues. From this derives a continuum of language types whose one end is occupied by the so called "iso-syllabic" languages - i.e. the ones in which, in the extreme case, all beats are regularly distributed timewise; and the other end is occupied by the so called "iso-accentual" languages - only stressed beats count for rhythm. Parallel to the typological hierarchisation there exists a language-specific differentiation as to how particular languages realize a universal preference for even beat distribution.

Universally, the inter-relationships between vowels and consonants in a speech chain are based on the following criteria: sonority, segmental strength, perceptual salience, ease of articulation, and symmetry in binding consonants to vowels in the speech chain. The latter is meant to signify a proportional in numbers grouping of consonants around beats which supports an ideally regular beat distribution timewise. This criterion, however, is easily overridden by other preferences. By means of the above criteria one can account for the universally preferred structure of a foot i.e. a C₁V₁C₂V₂ with a trochaic rhythmic pattern (cf. [4] for details). In a one-beat content word there is a preference for a CVC structure or for a CVV one (i.e. a consonant followed by a long vowel or a diphthong) by means of which stress on this only beat is conveyed (at least partly, beside a potential change in pitch and loudness). These structures are traditionally called "heavy syllables". Thus, what used to be called a "heavy syllable" is the preferred structure of a

minimal content word. A "light syllable" stands for "less than that" i.e. a single beat structure not able to satisfy the above minimal content word requirements.

It is the number of vowels that is indicative of the number of beats in the first place. There are two other sub-cases, however. Firstly, the sequences V:C and VCC(C)o..n, although they involve one vowel, count for more than one beat, i.e. they form a category in between a one-beat structure and a two-beat structure. Secondly, a consonant may take up a beat function.

5. CONSONANTAL BEATS

Preferably, a consonantal beat is separated from the nearest vowel by a consonant of a low sonority (or, at least, lower than that of the consonantal beat itself).

Universally, consonantal beats are assigned post-lexically: they function as a real-time resolution of a rhythmical conflict. Thus, for instance, if a vowel is elided in fast/casual speech, one of the neighbouring consonants may take over a beat function (e.g. Eng. ['hæpɪ], Pol. ['fʃs(t)kɔ], or, otherwise, a cluster that results from the reduction may get simplified (e.g. Pol. ['fʃstko] → ['fsko]). Those clusters are originally, i.e. immediately before vowel elision, disfavoured by universal word phonotactics as well as, often, by language-specific phonotactics. If such a phonotactically disfavoured cluster is legalized in a given language (pre-)lexically (e.g. Pol. ['mgwa], ['rtɛŋtɕ], ['mɛɕ] or ['nastɛmpstf]), a post-lexical resolution is either to weaken a sonorous element (a potential consonantal beat) or to reduce a cluster of consonants (if their sonority is levelled). Language-specifically, however, consonantal beats might arise in

a different fashion i.e. they either become lexicalized as a result of a generalization of a post-lexical rule (cf. Eng. ['lɪtɪ]) or they are lexically assigned in the first place (cf. Czech [kr̩moval] or Serbo-Croatian [krkaɫ]). In other words, consonantal beat assignment is a process which has reached different levels of application in particular languages, ranging from full lexicalization to phonostylistics.

6. INTER-BEAT SEQUENCES

There is a preferred order of consonants from one beat to another with respect to their sonority value. Specifically, what is favoured is a constant fall in sonority starting just after a beat and finishing just before another one (which constitutes a rise). This general preference can be most obviously overridden by morphology (a break in the sonority fall enhances morphological transparency), but also, language-specifically, within a morpheme.

Apart from the preference concerning the inter-beat consonants themselves, there are certain regularities concerning the way in which the consonants tend to bind to beats. These bindings derive from the criteria discussed in 4. above, as well as from the just mentioned preference. And, thus, in a VCV sequence, a C is preferably bound to the following V. This mirrors word and foot-initial binding, but notice also that in a VCV the consonantal sonority fall is impossible - there is only a rise on the second V thanks to the preceding C, which draws them together. In a VCCV, consonants bind to a respective preceding and following V (cf. symmetry in 4. above), unless sonority or stress-assignment criteria intervene (e.g. more

consonants are bound to a stressed beat). If there are more than two consonants in an inter-beat cluster, a default binding is as above, i.e. one C is bound to the following V and the remaining consonants are bound according to the symmetry, stressed-beat and sonority slopes principles. The default binding, however, is subject to a number of potential modifications of a language-specific and/or post-lexical (phonostylistic) nature.

Thus, generally, beat-counting constitutes the basic organizational principle of the speech chain, while binding laws should be understood as a set of universal potentials invoked in a language-specific way by particular languages. The reader is referred to [4] for a more comprehensive treatment of the issue.

7. REFERENCES (SELECTED)

- [1] ALLEN, G.D. (1975), "Speech rhythm: its relation to performance universals and articulatory timing", *Journal of Phonetics* 3.2, 75-87.
- [2] BERG, T. (1989), "Intersegmental cohesiveness", *Folia Linguistica* 23, 245-280.
- [3] DRESSLER, W.U. (1985), "Explaining Natural Phonology", *Phonology Yearbook* 1, 29-50.
- [4] DZIUBALSKA-KOŁACZYK, K. (1991), "A natural principle of sound structure organization: the syllable undermined", forthc., in Hurch, B. (ed.), Mouton.
- [5] FUJIMURA, O. (1981), "Temporal organization of articulatory movements..", *Phonetica* 38, 66-83.
- [6] HYMAN, L.M. (1985), "A Theory of Phonological Weight", *Dordrecht: Foris*.
- [7] RUBACH, J., BOOIJ, G. (1990), "Syllable structure assignment in Polish", *Phonology* 7, 121-158.
- [8] SIEVERS, E. (1901), "Grundzüge der Phonetik", Leipzig: Breitkopf and Härtel.

SYLLABLE TONEMES IN LATVIAN

D. Markusa

Latvian University, Riga, Latvia

ABSTRACT

The present contribution attempts to verify and specify the analogous phonological opposition of the two syllable tonemes occurring in the North-East Vidzeme and Latgale variants of the High Latvian dialect. The unification process of syllable tonemes which is under way in the Latvian Standard language has also been analysed.

1. INTRODUCTION

In Latvian the initial syllable bears the stress, as a rule. It is only in some cases that the stress may rest on any other syllable. The syllable toneme is an independent prosodic feature in Latvian. It functions irrespective of word-stress. In some cases the syllable toneme has a semantic function, for example, luōks 'spring onion', luōks 'a bow or shaft-bow', luōks (-gs) 'window'.

2. LATVIAN STANDARD LANGUAGE

In the Latvian Standard language three types of syllable toneme are conventionally distinguished: falling (˘), broken (glottalized) (^) and drawling (˜), yet the use of two tonemes: drawling and non-drawling is compulsory. It is left for a speaker to choose

either one syllable toneme (basically falling) or distinguish between the falling and broken toneme within the limits of non-falling syllable tonemes. Sometimes the choice of these tonemes has semantic function, for example, when distinguishing the adverb kā (kā tu roc? 'How do you dig?') from the pronoun kā (kā tev nav? 'What do you lack?') Latvian linguistics lacks experimental research concerning both the systems of syllable toneme in the Standard language and dialects, and the processes proceeding in them. At present the u n i f i c a t i o n of syllable tonemes can be observed in the Standard Latvian language. Formerly the attention to this phenomenon was drawn by J. Endzelins [3], V. Dambē [2], S. Rāge [6]. The present research attests that the unification proceeds in several directions: I. Under the influence of the Low Latvian dialect, called Lejzemiēku, the drawling syllable toneme, as a hypernormal feature, may be heard in some words, for example, maizē 'bread', tauta 'people', ieļa 'street'. S. Rāge [6] and V. Dambē [2] had also noted the occasional wrong use of the drawling syllable toneme, in-

stead of the falling one, by High Latvians under the influence of Low Latvian dialects. II. In Present-day Standard Latvian the use of the falling toneme has notably increased. In the speech of professional linguists, too, one can often hear the wrong usage of gaitē 'cord', gūkatē 'to brush', siņucē 'onion', taupa '(he) saves'. (Some cases, for example, taupa, may also serve as a reverse reaction to the hypernormal pronunciation of tauta). Basically two reasons account for this: 1) the impact of the High Latvian dialect (In this dialect the falling toneme is substituted for the drawling one. A great number of Latvian intellectuals are born High Latvians); 2) the impact of the Russian language: a) the drawling syllable toneme that is marked by the duration of vowels is pronounced shorter. Therefore it is quite credible that duration marks the drawling toneme only in the syllables containing open and half-open vowels; b) The specific features of a broken toneme tend to disappear - there is a notable decrease of glottalization. The latter is more level, and the increase of duration of the broken tonemes is occasionally noted. Besides, the falling toneme is specified by stability in the Latvian Standard language. In general, within the system of three syllable tonemes which comprises two tonemes of level character, e. i. falling and drawling, the falling toneme is effected more precisely (usually by the level falling duration of the intensity and the fundamental pitch)

than in the area where two tonemes are used. In the latter they must differ, for example, only from the broken toneme which, in its turn, is marked by a cut intensity and the fundamental pitch changes. We have already pointed out the variations in the intensity of the falling tonemes and the fundamental pitch direction in the two tonemes area [4]. In 1923 A. Abele wrote [1] that when word-stress in the Latvian was shifted to the initial syllable, the drawling toneme turned into the falling one under the impact of the East Slavonic neighbours. In the Present-day Latvian language the accentuation of the initial syllable has become fixed. You need not go East in search of the impact of the Russian language. Consequently, the shifting process of toneme is recurring.

3. DIALECTS

The syllable tonemes in dialects appear to be more stable. The area of the very singular High Latvian dialect with two syllable tonemes (˘ and ^) in use, embraces the territory of two Latvian regions: Latgale and North-East Vidzeme. We have made experimental measurements of the syllable tonemes used in Latgale, namely, in the subdialects of Baltinava, Berzgaile, Silajani and Preiļi. The obtained data have been compared to the characteristics of the syllable tonemes used in the formerly explored sub-dialects of North-East Vidzeme, namely, Ziemeļi [4, 5], Aluksne, Jaunlaicene, Jaunroze, Karva, Vec-laicene [5]. The sub-dialects under discussion are still used in daily commu-

nication and farming. I have analysed tonemes in the syllables containing long monophthongs and diphthongs in both isolated words and phrases. It may be concluded that the characteristic features of the syllable tonemes in isolated words, as well as in phrases, are identical irrespective of the fact that the duration of produced speech sounds in isolated words of the sub-dialects under consideration at an average 1,3 times exceeds the duration of the speech sounds fixed in phrases. The experiment concerned with the investigated Latgale sub-dialects permits to conclude the usage of two phonologically distinctive types of syllable tonemes: 1) the *level* toneme (conventionally called falling and marked by \searrow). This type of toneme is specified by level intensity and fundamental pitch changes and longer duration. The intensity and fundamental pitch direction can be specified as level falling, level rising - falling, or level rising; 2) the *acute* toneme (conventionally called broken and marked with \wedge). This type of toneme is specified by the acute intensity and fundamental pitch changes and shorter duration. The intensity and fundamental pitch direction can be described as acute falling, acute falling-rising, or acute rising-falling. In the Berzgaile sub-dialect the oscillogram of the monophthong \underline{e} produced by an informant shows a particular saw-shaped design with a rising-falling -rising-falling direction. This type (\wedge) of toneme is specified by glottalization, i.e. a decrease in the regula-

rity of the vocal cords vibrations. Yet it is impossible to state the starting point of glottalization only by oscillograms. Both types of toneme are contrasted to each other by the presence or absence of a specific prosodic distinctive feature - an *acute* or *level* characteristics of the intensity and fundamental pitch changes. Yet it is indisputable that intensity changes is a more precise toneme indicator, for example, the monophthongs \underline{i} , \underline{u} , \underline{e} , \underline{a} , produced in phrases, differ in the intensity of the second part of monophthongs; in the absolute distinction between the intensity of the first and second part of a monophthong; they differ in the distance to the upper limit of intensity maximum; in the distance between the maximum and minimum intensity; in the range of the rise and fall; in the rapidity of the intensity of rise and fall. The above mentioned indicators have a credibility rate of toneme distinction which is $> 97,0\%$ - $99,9\%$ (by Student's criterion). According to the fundamental pitch changes the same monophthongs differ only in the average fundamental pitch and that of the second part of the monophthong. Besides these indicators are only relative toneme distinctive features (85,3% and 86,4% distinctive credibility by Student's criterion). Other measurements produce similar results. Consequently, in Latgale sub-dialects which produce level and acute toneme opposition dynamics is the basic distinctive feature of these tonemes. I have arrived at similar conclu-

sions from my former research when investigating level and acute toneme oppositions in North - Vidzeme sub-dialects I4,5I. So this permits generalization that in the High Latvian Non-selonian sub-dialects, where two syllable tonemes prevail, two types of toneme - *level* and *acute* are phonologically distinctive in both North-East Vidzeme and Latgale. Dynamics serves as the basic distinctive feature between them. The High Latvian dialect comprises another group of sub-dialects, the so called Selonian. However, I temporarily lack a sufficient number of experimental tests to offer the complete specification of these tonemes. I may only share the hypothesis advanced by A.Sarkanis I7I to the effect that in Selonian sub-dialects, the musical or melodious moment predominates as a distinctive feature of tonemes. It would be essential to subject to experimental test also some other phenomena of the High Latvian dialect, for example, the usage of the rising syllable toneme, characteristic of the Selonian sub-dialects, in the speech of the Non-selonian (e.g. Izvalta) native population. The syllable toneme, which occurs in the result of contraction and reminds of the drawing toneme and some other syllable toneme variations, also deserve experimental research.

REFERENCES

- [1] ĀBELE, A. (1923), "Par stieptās intonācijas pārēju krītošā", FBR, 3, Rīga, 40-42.
 [2] DAMBE, V. (1974), "Intonācijas", LVKJ, 10, Rīga, 198-207.
 [3] ENDZELĪNS, J. (1974), "Weiteres zu den Lettischen Intonationen", Darbu izlase, 2, Rīga, 513-527.
 [4] MARKUS, D. (1987), "Types of syllable toneme in the Ziemeri Variant of High Latvian Dialect", Proceedings XI th ICPHS, 5, 107-110.
 [5] МАРКУС, Д. (1989), "Слоговые акценты в зимерском и в соседних с ним говорах верхнелатвийского диалекта", Верхнелатвийский диалект, Рига: ЛГУ им. П. Стучки, 68-80.
 [6] RAĢE, S. (1975), "Par zilbes intonāciju un latviešu literārās valodas normu", LVKJ, 11, Rīga, 99-120.
 [7] САРКАНИС, А. (1989), "Слоговые интонации селонского говора Калдабруня", Верхнелатвийский диалект, 81-99.

MODELIZATION OF INTONATION PATTERNS IN SPANISH FOR AUTOMATIC RECOGNITION

Juan María Garrido and Francesc Gudayol

Laboratori de Fonètica, Universitat Autònoma,
Barcelona, Spain

This paper describes a simple method to derive stylized representations from raw F0 contours. The results of a study on F0 contours in Spanish sentences using this method are also presented. Finally, application of obtained models to automatic speech recognition is discussed.

1. INTRODUCTION

Modelization of prosodic information for automatic synthesis and recognition systems can be made in several ways. An usual approach has been the analysis of fundamental frequency (F0), duration and amplitude of syllabic nuclei in sentences, in order to detect stressed syllables and clause boundaries. Intonative information can also be extracted from a representation of different F0 levels in the syllables.

Such an approach has demonstrated to be very effective in synthesis systems (as the one developed by Pierrehumbert [3]). But in recognition systems, although important results have been achieved, some questions, as the automatic detection of syllabic nucleus, have not still been entirely solved. Results reported by Vaissière [4] and Mertens [1] are good examples of this fact.

This paper describes a simple method of representation of F0 contours (section 1), and a study of melodic patterns in Spanish sentences using stylized contours obtained with this procedure (section 3). This study

has been achieved as a primary step in the development of an automatic pitch contour recognizer for Spanish. The use of the obtained results in automatic recognition is evaluated in section 4.

2. STYLIZATION PROCESS

Process of stylization included various stages:

- 1) Extraction of F0 values.
- 2) Obtention of stylized representations, by saving only the values corresponding to time and F0 at beginning, end and inflection points of each sentence. Inflection points can be defined as points in the F0 contour where slope changes its sign (from positive to negative, or *vice versa*). Straight lines traced between these points form the stylized representations for each contour.
- 3) Finally, simple time and frequency normalizations were performed, in order to avoid variations due to intrinsic F0 and speech rate for different speakers. In frequency normalization, initial F0 values of each contour were set to 0, and values corresponding to the remaining points were made relative to it. Time normalization sets to 0 and 1 time values corresponding to beginning and end of F0 trace respectively; other time values in the representation are made relative to these reference points.

This stylization procedure is very straightforward and can be easily

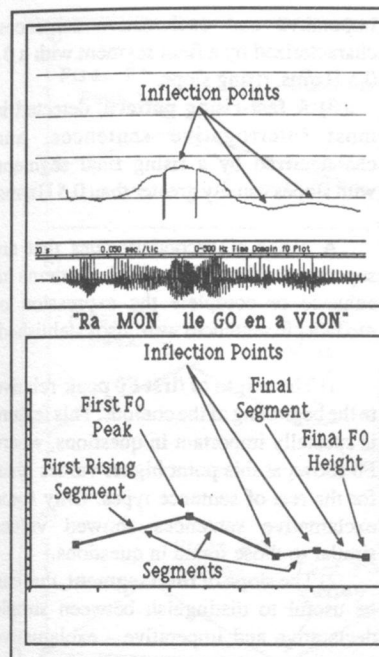


FIGURE 1: An example of original F0 trace, and its correspondent stylized representation

automatized if desired, although it has been applied manually in this study. An example of stylized representation, and its corresponding F0 contour, is shown in Figure 1.

Some remarks can be made about this method:

- 1) A whole-sentence approach has been preferred to a syllable approach in this analysis procedure. The F0 trace is divided into several segments, defined as the straight lines that link two contiguous inflection points. Obtained F0 peaks can correspond or not to stressed syllables.
- 2) Only F0 variations are analyzed. Duration and amplitude of syllables are considered to be important for transmission of stress information, but they seem secondary cues in intonation analysis.
- 3) Micromelodic variations are not considered in this method. Segments

showing increases or drops in F0 less than 10 Hz between its beginning and its end are not taken account of.

3. EXPERIMENTAL PROCEDURE

The described procedure was applied to stylize a set of F0 contours extracted from simple Spanish sentences. The analysis of the obtained representations carried to the definition of some typical F0 patterns used in Spanish for expression of sentential modality.

3.1. Design and Recording of Corpus

A set of 54 sentences was constructed representing six different modality types, adapted from the traditional classification of Navarro Tomás [2]: a) enunciative; b) interrogative type I; c) interrogative type II; d) interrogative type III; e) exclamative; and f) imperative. Each subset included different tokens of sentences with changes in number and position of stressed syllables.

Sentences contained each only one intonative group, and were formed exclusively by voiced sounds, to provide the analysis of complete F0 contours. They were embedded into brief dialogues to facilitate the production of a more natural intonation while reading. Dialogues were read by 4 Spanish speakers, 2 men and 2 women, in a sound isolated booth, and recorded on high-quality audio tape.

3.2. Analysis Procedure

Sentences to be analyzed were then low-pass filtered, digitized at a sample rate of 10 Khz, and stored. F0 contours calculations were performed by means of a pitch detector based on an auto-correlation technique, available in MacSpeech Lab II, a commercial speech analysis software for Macintosh.

The procedure described in section 2 was applied to each F0 contour in order to obtain the stylized representations.

Previous studies [5] showed that initial and final parts of F0 contours contain most intonative information. To take account of this fact, slopes of initial and final segments for each sentence were calculated.

3.3. Statistic Analysis

Two sub-groups of sentences were formed for each modality type, according whether the slope of final segment was rising or falling. A simple statistic analysis was then performed to extract mean values of the slope of first and last sentence segment for each sub-group. Means were also calculated for F0 values at the first peak (that is, at the end of the first rising segment), and at the end of the contours. The results of this analysis are shown in Table A.

3.4. Pattern Classification

According to these results, the obtained F0 stylized contours were classified into three basic patterns (see figure 2):

1) A falling pattern, for all declarative and some interrogative, imperative and exclamative sentences, characterized by the presence of a final segment with falling slope.

2) A slow-rising pattern, for some

imperative and exclamative sentences, characterized by a final segment with a 0-0,5 Hz/ms. rising slope.

3) A fast-rising pattern, detected in most interrogative sentences, and characterized by a rising final segment, with slopes usually greater than 0,5 Hz/ms.

A series of secondary cues that are superimposed to these basic patterns to enhance or complete the expression of modality in sentences were also established:

1) The height of first F0 peak relative to the beginning of the contour. This feature is specially important in questions, where F0 shows at this point higher values than for the rest of sentence types. Only some exclamative sentences showed values similar to those found in questions.

2) The slope of final segment, that can be useful to distinguish between simple declarative and imperative - exclamative sentences: the second ones usually show a more abrupt slope (less than -1 Hz/ms.) than the first one.

3) The use of some special F0 contour forms, as the "circumflex final", or the "wave-like contour" (see Figure 3), usually to reinforce the expressive message in exclamative and imperative sentences.

4) Variations in F0 range of whole

TABLE A: Mean values of Slope (in Hz/ms.) and F0 Height (in Hz relative to the initial F0 value) in main trace segments, calculated for each sentence type

Sentence type	Final Slope	FIRST RISING SEGMENT		FINAL SEGMENT	
		Slope	F0 Height	Slope	F0 Height
Enunciative	-	0,38	59	-0,27	-29
Int. Type I	+	0,65	96	0,87	171
Int. Type II	+	0,57	123	1,06	146
	-	0,75	80	-0,52	-42
Int. Type III	+	0,70	110	0,79	157
Exclamative	+	0,72	119	0,48	13
	-	0,51	93	-0,49	-35
Imperative	+	0,36	43	0,40	15
	-	0,61	59	-0,42	-38

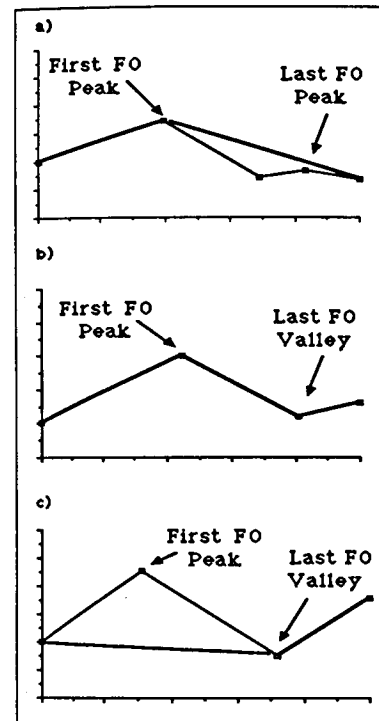


FIGURE 2: Typical F0 patterns in Spanish: a) Falling; b) Slow Rising; c) Fast Rising.

sentence, associated to the expression of emotions by speakers.

4. APPLICATION OF RESULTS TO AUTOMATIC RECOGNITION

This stylization method could be adapted to automatic implementation. A smoothing procedure would be introduced, to eliminate variations due to F0 detection errors and to interpolate segments where F0 is not present, and a different time normalization procedure would be used, since the one applied at this study is not useful with varying length sentences.

The obtained models are being applied to automatic labelling of sentence types. Slope and relative height of initial and final segments of stylized contours seem sufficient cues for the identification of

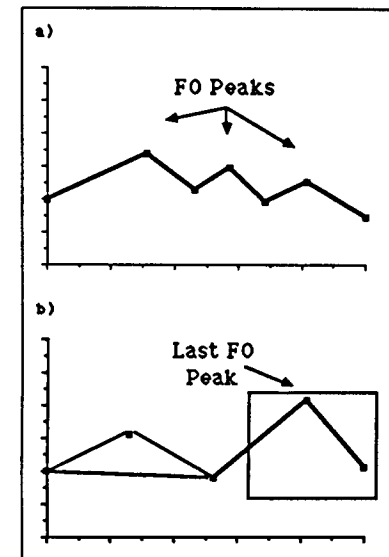


FIGURE 3: Special F0 patterns: a) "Wave-like" Contour; b) Circumflex Final.

different types.

Further research to verify the perceptual salience of this schematic models and its validity in continuous speech sentences has still to be carried out.

This work has been supported by a grant of the CIRIT from the Catalan Government.

REFERENCES

- [1] MERTENS, P. (1989), "Automatic Recognition of Intonation in French and Dutch", in TUBACH, J.P. - MARIANI, J.J. (Eds.), *Eurospeech 89*, Vol. I, Paris, 46-50.
- [2] NAVARRO TOMAS, T. (1948), *Manual de entonación española*, Madrid: Guadarrama.
- [3] PIERREHUMBERT, J. (1987), *The Phonology and Phonetics of English Intonation*, Bloomington: Indiana University Linguistics Club.
- [4] VAISSIÈRE, J. (1989), "On Automatic Extraction of Prosodic Information for Automatic Speech Recognition Systems", in TUBACH, J.P. - MARIANI, J.J. (Eds.), *Eurospeech 89*, Vol. I, Paris, 202-205.
- [5] TAKEFUTA, Y. (1975), "Method of Acoustic Analysis of Intonation", in SINGH, S. (Ed.), *Measurement Procedures in Speech, Hearing and Language*, Baltimore: University Park Press, 173-186.

INTONATION AND AMBIGUITY

M^a Carmen Fernández Leal

Universidad de La Laguna

ABSTRACT

The conditions for an expression to be ambiguous are the result of a rule violation. There is no violation of the nuclear stress rule (NSR) but its substitution for the compound stress (CSR), as a result of transformations that affect the semantic representation of a sentence. The ways of cancelling ambiguity can be regarded as having a semantic status, when there is a connection with the alteration of a context, or a change in the semantic representation.

1. INTRODUCTION

1.1. The nature of ambiguity

The twofold aspects of ambiguity, consisting of the diversity of semantic description and singularity of formal expression, is the starting point in the understanding of this semantic device. Form and meaning are to be considered in an analysis of ambiguity, as well as a disassociation of the space relation, of what goes with what, that affects word order and agreement.

1.2. Types of ambiguity

The conditions for an expression to be ambiguous are the result of the violation of a syntactic or semantic rule, and such conditions establish the kind of ambiguity.

1.2.1. Syntactic type

The violation of a syntactic rule gives rise to a syntactic type of

ambiguity, and it could affect different syntactic aspects.

A. The space definition of a word concerns its grammatical category, because of its capacity to be ascribed to more than one category, when the adequate lexical meaning is provided: (1) **They are visiting friends** has two different readings, according to the double syntactic function of **visiting** as a present participle or an adjective: (a) **they are people who are visiting their friends** and (b) **they are people who visit their friends**.

B. The delimitation of a word space is related to a conflict of agreement, due to the lack of inflections: (2) **Old men and women should be out of danger**. Interpretation (a) is based on the restrictive agreement of an adjective with the nearest noun, when there are two joined in coordination: **Old men as well as women should be out of danger**, and interpretation (b) concerns the agreement of an adjective with the two nouns: **Old men and old women should be out of danger**.

C. The space assignment to a word can affect the modal aspect, that is restricted to modal operators like quantifiers, with a modality *de dicto* or *de re*, regarding the scope of influence: (3) **Ruth wants a woman to stay**. The two readings are the result of the application of a modality *de dicto*: (a) **Ruth wants any woman to stay**, or a modality *de re*: (b) **Ruth wants a particular**

woman to stay. In (a) a woman is in the scope of **want**, and in (b) **wants** is in the scope of a **woman**.

1.2.2. **Lexical type**
The violation of a semantic rule brings about a lexical type of ambiguity. The cause of ambiguity is the cooccurrence in a syntactic context of two types of meanings that refer to the same lexeme: (4) **Mary is going to have a baby**. The dual interpretation of the sentence is due to the substitution of a conceptual meaning for an associative one: (a) **Mary expects a baby**, and (b) **Mary intends to have a baby**.

1.2.3. **Lexico-syntactic type**
The border line between syntactic and lexical ambiguity is not clearly delimited, and a violation of a syntactic rule merges with the violation of a semantic one. In this type of ambiguity a double grammatical category, and a double meaning, is attached to the same lexical item: (5) **We saw her duck**.

As a general rule syntactic and semantic ambiguity are based on space arrangement, consisting of the different scope of influence of a word, or of two words sharing the same space.

1.2.4. **Lexico-phonetic type**
It's a variety of the lexical type of ambiguity, and it's perceived only acoustically. A semantic and phonetic aspect merge in the violation of the rule, in the sense that a one to one correspondence between sound and meaning should be kept: (6) **That kind of piece over there seems to be adequate**.

2. PROCEDURE

2.1. Causes of ambiguity and their occurrence.

There is an interdependence among the causes of ambiguity, because syntactic ambiguity relies on lexical meaning, and lexical ambiguity needs the help of a specific syntactic arrangement (it

is only in surface structure that ambiguity takes place). The lexico-phonetic variety is also dependent on lexical meaning and on a syntactic pattern. As a result of these interconnections, it is inferred that ambiguity is a syntactically dependent phenomenon.

The adjustment of two senses in a syntactic structure creates, basically, a problem of word order, not being able to reflect the dual space arrangement that takes place in a semantic representation. The other source of ambiguity concerns the selection restrictions that apply to the components of a lexeme.

The word order cause of ambiguity is manifested through lexical words as in (1), (2) and (5), or with the help of grammatical words as in (3). This cause of ambiguity implies a major occurrence. Pure lexical ambiguity caused by means of a selection restriction of components is second in importance, with the predicate being the main cause of ambiguity. The occurrence of the lexico-phonetic type of ambiguity is reduced when it doesn't concern a purely lexical aspect. If we consider ambiguity on the double aspect of syntactic and semantic concern, syntactic ambiguity is more current than semantic, when the source is not purely lexical.

2.2. **The nuclear stress rule**
The violation of the nuclear stress rule (NSR) that takes place in some syntactic types of ambiguity, and in lexical ones, cannot be considered as a condition for a prosodic type of ambiguity, and it is not the only sign that distinguishes an ambiguous expression from its invalidation.

On the other hand, the violation of the NSR is only apparent, because if we take embedding, in the form of a relative clause, as a transformation of the surface

structure, the violation of the NSR does not occur, and it can be considered the consequence of the application of the compound stress rule (CSR). The transformations on the different ambiguous sentences, taken as examples, are: (1) **They that are visiting friends**. There is no transformation in (2); (5) and (6), because there is no violation of the NSR. (3) **Mary that to have a baby is going**; (4) **Ruth that to stay wants a woman**.

The superficial violation of the NSR is the result of a rewriting of the application of the compound rule to the embedded transformation: (1) **They (that are) visiting friends**; (3) **A going to have a baby woman ***; (4) **A woman to stay wanted ***. The violation of the NSR is explained through the application of the compound rule, that is lexically based, once the embedding transformation, which is syntactically rooted, takes place; we have two main resorts on which the semantic phenomenon relies: semantics and syntax.

An alternative in the application of the CSR and the NSR, in the form of a compound word or a phrase ('bluebell and bluebell'), is a way to invalidate ambiguity, so that we can either confer the ability to invalidate ambiguity to the CSR or to the NSR, and the phrase can also be interpreted as a relative clause: **the bell that is blue**.

2.3. Other means to invalidate ambiguity

Although the apparent violation of the nuclear stress rule occurs in (1), (3) and (4), it doesn't take place in (2), (5) and (6). If the external stress placement is taken as a sign of the invalidation of ambiguity, another means of acquiring it has to be indicated. The specific way to cancel ambiguity in (2) is using a pause: **Old men I and women should be out of danger**, and the same applies to (5): **We saw her I duck**. The acoustic confusion

originated in (6) is solved by making use of the context.

A pause seems to be a generalized way to make clear which sense is meant, because even in (1), (3) and (4) it helps to invalidate ambiguity: (1) **They are I visiting friends**; (3) **Ruth wants I a woman to stay**; (4) **Mary is going I to have a baby**.

Apart from the paralinguistic devices, there are non-prosodic ways to cancel ambiguity, and they can be classified into syntactic and semantic means.

The syntactic means are expressed in the form of embedding and downgrading predication, as it can be the function of the ambiguous expression like a noun clause, or the introduction of a relative clause that determines the subject or object of the ambiguous sentence. Another syntactic device is the addition of a prepositional phrase, which completes the meaning of the predicate. Lexically, the substitution of one grammatical word for another (articles) can also cancel ambiguity.

The most common semantic way to invalidate ambiguity is context. The substitution of the content words, when the new ones don't constitute part of their selection restrictions, or the ones of the predicate, can be added as a semantic device.

The distinction between a syntactic type of invalidation and a semantic one is relative, because the selection restrictions of the lexemes determine the success or failure of a syntactic type of invalidation.

If we consider the two senses of an ambiguous expression as marked and unmarked members, the invalidation is more frequent on the marked member.

2.4. Further classification

The application of the different syntactic and lexical resorts for the cancelation of ambiguity is the reason for a subsidiary classification, according to the capacity of an ambiguous expres-

sion to take the different tests. An open ambiguous expression is easily invalidated, and a close one takes more effort. The words open and close have to be considered in relative terms. The degree of invalidation of ambiguity has no correspondence with the general classification of syntactic and lexical types of ambiguity, so (1) is a close syntactic ambiguous sentence, and (4) a close lexical one. When ambiguity is based on sound perception, it is considered as open.

2.5. Semantic status

The non-prosodic means to invalidate ambiguity, consist mainly in the addition of information that concerns word meaning; the procedures are lexical, adapted to certain types of syntactic structures. They are assigned a semantic status, because there is context variation, bringing about a change in meaning.

2.5.1. Semantic status of intonation

Intonation and pause are prosodic ways to dispel ambiguity, and the question is whether they have a semantic status, or whether they are rhetorical procedures that have a rhetorical interpretation. Assuming the fact that the existence of ambiguity requires the violation of a rule that gives rise to a type of ambiguity, and that there isn't a real violation of the nuclear stress rule, there is no reason to make it responsible for the distinction between the two senses of an ambiguous expression. The surface distinction is due to an application of the compound rule after an embedded transformation. Intonation has a semantic status if we consider that there is a transformative process that applies to deep structure, having a semantic representation, but not if we make it responsible for a change in meaning, concerning the unmarked member of the ambiguous expression, because the external differences between

the two senses are the result of the application of the compound and the nuclear stress rules. On the other hand the acoustic perception of intonation doesn't seem to be clearly distinctive to the listener, and he has to have other means to distinguish one sense from the other. These means are the intuitive knowledge that the listener has of the violation of a syntactic or semantic rule, apart from the fact that sentences are inserted in a context, and don't occur in isolation.

If we approach intonation as a means of focussing new information, it's dubious that, in the same context, **visiting** is old information in (1) (a), and new in (1) (b). It's more likely to be thought of as in contrast to something that has already been said. We can take the lack of distinction between new and old information as a proof of the non-violation of the nuclear stress rule.

A pause is a prosodic device that can be used to determine the sense of an ambiguous expression, on the form of a space mark, used as a rhetorical device that invalidate ambiguity, mainly in cases of conflicts of arrangements, but it cannot give rise to a change of meaning.

2.6. Connections between conditions and cancelations of ambiguity

There is no strict connection between the conditions that make an expression ambiguous and the means to cancel ambiguity; among these are included non-lexical semantic means, the context, and prosodic devices as stress and pausing. Lexico-phonetic means are excluded, though there is a fundamental correspondence between the syntactic and lexical aspects.

EXPLOITING THE SECONDARY ACCENT IN A PROSODIC MODEL FOR FRENCH SYNTHESIS

V. Padeloup

Institut de Phonétique, Université de Provence, 13621, Aix-en-Provence.
permanent address: 8 rue du Château-Landon 75010 Paris, France

ABSTRACT

Unlike most other existing French speech synthesis systems where prosodic organization depends mainly on syntactic structure, we adopt an approach where the phonotactic criteria (syllable count, length of stress group etc.) are taken into account. We hypothesize that secondary accent takes a prominent part in stress structuration.

Based on acoustic and perceptual analysis of read corpus, we present:
- a general description of secondary accent which includes: distribution rules, phonotactic constraints which rule its occurrence, and acoustic description;
- the linguistic and phonotactic criteria for prosodic structure analysis;
- a model for sentence prosodic organization.

1. INTRODUCTION

Traditionally, French language is described as having an accent at the end of lexical words (primary accent). However some sporadic works done in the last two decades mention the presence of a secondary accent which is not on the last syllable of lexical words and not related to the rhetoric or enunciative accent (focalization) [1, 2, 3, 4].

Prosodic and perceptual analysis of 400 read utterances (5 speakers, 40 sentences, 2 repetitions) has been carried out [6].

2. SECONDARY ACCENT

Secondary accent should be taken into consideration in French accentual structure: the analysis of 400 read utterances shows that, in polysyllabic words, almost one accent out of three is a secondary accent [5, 6].

Secondary accent distribution is the following:

- on the first or second syllable (often first syllable starting with a consonant) of a word; this word is not necessarily marked by final primary accent (stressed syllables are coded I, unstressed syllables -);

un char mant gar çon
- I - - I

- on the antepenultimate of a word ending with a primary accent;

une dose in fi ni té si male
- I - - - I - I

- at the boundary of a morpheme in a polymorphemic word:

c'est an ti / cons ti tu tio nnel
- - I - - - - I

The secondary accent is located mainly at the beginning of a word or of a group.

The occurrence of a secondary accent in a word depends on various constraints [5]:

- the accentual context (among others the number of unstressed syllables between secondary accent and the preceding or following accent);
- the word position in the sentence;
- phonetic nature of word first segmental unit (consonantic or vocalic);
- the number of syllables of the word;
- accentual strategy or speaker's regional and individual characteristics.

We hypothesize that secondary accent have a *regulatory function in the production of stress*. Thus the stress pattern in a given sentence would be consistent with the *biological and psychological standards of the production of rhythm* [7, 8]: the secondary accent generally occurs so that a series of unstressed syllables never exceed the

count of 4, and is distributed to an average of one accent (primary or secondary) per 3 syllables.

In the following sentence, if solely the final primary accents occurs, the series of 5 unstressed syllables would break the perceptive-motor unit [7]. Secondary accent on the word "désintoxiqués" would allow avoiding this lengthy series of unstressed syllables:

	il	voy	ait	des	dé	sin	to	xi	qués
primary acc.	-	-	I	-	-	-	-	-	I
secondary acc.	-	I	-	I	-	-	-	-	I
secondary acc.	-	I	-	-	-	I	-	-	I

The secondary accent is acoustically distinct from primary accent. The secondary accent is best described by a rising pitch movement (with medium amplitude) and a short syllabic lengthening (generally non significant); the primary accent being best described either by a rising or falling pitch movement (with medium or large amplitude) and by a variable syllabic lengthening (generally significant). Focalization accent differs from secondary accent by a larger amplitude of the pitch movement and a stronger intensity.

3. CRITERIA FOR PROSODIC STRUCTURE ANALYSIS

Almost all prosodic models developed during the last two decades are originally based upon a syntactic analysis to generate accentuation or intonation: prosody is considered to be congruent to syntax [9, 10].

Recently, in work on prosodic models, some importance has been given to phonotactic (also called eurhythmic) phenomena such as: principles of accentual alternation (strong/weak) and syllabic balancing, constraints on the number of consecutive unstressed syllables [11, 12, 13, 4, 14, 5]. However, these phonotactic constraints are generally considered only after having derived the prosodic structure from the syntactic structure, therefore compensating for the weaknesses of these models [11, 13].

The main phonotactic constraints are the following [5, 15]:

- *size of stress groups*: a stress group - i.e. in French, one stressed syllable preceded by one or more unstressed syllables - is composed of few syllables: generally about 2 to 5, on an average of 3.

- *number of consecutive unstressed syllables*: as a rule, we try to avoid more than 4 consecutive unstressed syllables (exception occur in cases of: bracketing, rapid speaking rate...). The stress clash rule is complementary to this one.

- *rule of the first accent in a sentence*: this one is realized as soon as possible, usually on one of the first stressable syllable (often the first syllable starting with a consonant in the sentence first word).

- *phonotactic function of secondary accent*: refer to section 2 above.

- *rhythmic structure of the whole utterance*: accentual or syllabic duration alternation principles [12, 13]; recurrence of stress groups, intonative patterns and temporal sequences principles [7, 16, 5].

4. MODEL FOR SENTENCE PROSODIC ORGANIZATION

This model has been tested in text-to-speech synthesis (INRS-Télécom, Montréal). This speech synthesis system was dealing with a small amount of syntactic information: no syntactic analyzer but a small grammatical words dictionary.

We define a three-level prosodic structure (for more details refer to [6, 16]):

- *rhythmic sequences*: major intonation groups;
- *rhythmic words*: minor intonation groups;
- *accentual groups*.

The demarcation of prosodic groups and the stress structuration are defined with linguistic and phonotactic criteria. For example, intonation units (such as rhythmic words or sequences) can be for phonotactic reasons:

- either, if they are too small, regrouped to constitute a larger unit [17];
- or, if they are too long, divided in two to constitute smaller units.

We present an example illustrating the steps required to generate acceptable prosodic structures for two sentences having the same syntactic structure but different syllables count.

(A) *Determination of rhythmic sequences and rhythmic words*

(A1) *Sentence segmentation into linguistic rhythmic sequences (/ /)*

Firstly, in text-to-speech synthesis, we used punctuation information to delimit linguistic rhythmic sequences. Then, after

linguistic and phonotactic rhythmic words have been defined (refer to (A2)), we reconsider this segmentation (refer to (A3)). As no punctuation occurs inside the following sentences, each of them consists only of one rhythmic sequence.

1 / La souris est vue par un affreux chat /.
2 / Le pélican est dévoré par un gigantesque rhinocéros /.

(A2) Sentence segmentation into rhythmic words ([])

(A2a) Demarcation before the grammatical words

1 / La sou ris [est vue [par un affreux chat /.
2 / Le pé li can [est dé vo ré
[par un gi gan tes que rhi no cé ros /.

(A2b) Grouping of two (or more) small groups

Conditions :
- one is composed of 1* or 2* syllables;
- the grouping doesn't exceed the count of 7* syllables.

1 / La souris (I) est vue [par un affreux chat /.
1 2
1 2 3 4 5 syllables

(A2c) partition of large group

Conditions :
- one is composed of more than 6* syllables;
- the new group resulting from the partition must be composed of at least 3* syllables. (* These coefficients can be varied according with the speaking rate or the speakers individual characteristics).

2 / Le pé li can [est dé vo ré
[par un gi gan tes que [rhi no cé ros /.
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 [1 2 3 4

Results from the segmentation into linguistic ([]) and phonotactic ([|]) rhythmic words :

1 / [La souris est vue] [par un affreux chat] /.
2 / [Le pé li can] [est dé vo ré]
[par un gi gan tes que] [rhi no cé ros] /.

(A3) Linguistic rhythmic sequences partition into phonotactic rhythmic sequences (| |)

We never regroup together two linguistic rhythmic sequences.

Conditions :
- the new demarcation can only occurs between two linguistic rhythmic words;

- each new group resulting from the partition must be composed of at least two rhythmic words (linguistic or phonotactic).
2 / [Le pé li can] [est dé vo ré] /
1 2
[[par un gi gan tes que] [rhi no cé ros]] /.
1 2

(B) Automatic positioning of accents

The intonation where the accentual rules are applied is the rhythmic word (phonotactic or linguistic). The accentual rules define the distribution and the occurrence of primary and secondary accents. The secondary accent occurrence inside the rhythmic word is based upon the phonotactic constraints described in sections 2-3 above.

The stress clash rule is valid inside rhythmic sequences : each time an accent occurs, the preceding and following syllables inside the same rhythmic sequence cannot be accentuated.

Firstly, all unstressed syllables (grammatical words, antepenultimate of lexical words composed of at least 3 syllables) are defined.

(B1) Primary accents location

The last stressable syllable of each rhythmic word gets a primary accent.

1 / [La souris est vue] [par un affreux chat] /.
- . . . - I - . . . - I
2 / [Le pé li can] [est dé vo ré] /
- . . . - I - . . . - I
[[par un gi gan tes que] [rhi no cé ros]] /.
- . . . - I - . . . - I

(B2) Secondary accents location

(B2a) Application of the rule of the sentence first accent

First stressable syllable of the first rhythmic word gets an accent.

1 / [La souris est vue] ...
- I - - I
2 / [Le pé li can] ...
- I - I

(B2b) Secondary accent rules

These aim at effecting a secondary accent in conformity with the distributional rules so that :
- a series of unstressed syllables never exceed the count of 4*;
- the ratio of the total number of accents (primary and secondary) to the number of unstressed syllables in the sentence is included inside a range of 1/2 to 1/4 (average 1/3 : one accent per 3 syllables).

- we approximate to the followed breakdown of the total number of secondary accent :

- 80% of secondary accents on the word first syllable;
- 20% of secondary accents on the antepenultimate of a word.

If we achieve different results in respect with the preceding rules (different occurrences or distributions for secondary accent), we choose at random or according to regional or individual characteristics.

1 / [La souris est vue] [par un affreux chat] /.
- I - - I - - I - I - I
/ [Le pé li can] [est dé vo ré] /
2a - I - I - - - I
2b - I - I - - I - I
2c - I - I - - - I
2d - I - I - - I - I
[[par un gi gan tes que] [rhi no cé ros]] /.
2a - - I - - I - - - I
2b - - I - I - - - I
2c - - I - - I - - I - I
2d - - I - - I - - I - I

The accentual structure of sentence 2d has the characteristics of the journalistic speech style (systematic presence of a secondary accent at the beginning of a word). The accentual structure of sentence 2a will better match the characteristics of slow speaking rate - reading for example - (lower ratio of number of accents per number of unstressed syllables).

This prosodic model takes into consideration both linguistic and phonotactic constraints and not only, as in the past, linguistic constraints. Therefore two sentences having the same syntactic structure but composed of a different number of syllables will not be given the same prosodic structure.

5. CONCLUSION

Prosodic structure seems to be the result of a compromise between universal non-linguistic constraints (of biological and psychological type) and linguistic constraints relating to each language.

According to our hypothesis, in stress production, secondary accent has a phonotactic and linguistic function (demarcation of units), whereas primary accent has only a linguistic function.

However further research on other aspects of phonotactic constraints in prosodic structuration is required.

ACKNOWLEDGMENTS

This research was supported by a research grant (Ministère de la Recherche et de la Technologie) and by a training grant (Réseau Francophone des Industries de la Langue). I would also like to thank Doctor D. O'Shaughnessy we gave me the opportunity to work in the Montreal INRS-telecom laboratory.

REFERENCES

- [1] Fónagy, I. (1979) L'accent français : accent probabilitaire, *Studia Phonetica*, 15, Didier.
- [2] Verluuyten, S. P. (1984) Phonetic Reality of Linguistic Structures : the Case of (Secondary) Stress in French, *Proc. of the Tenth International Congress of Phonetic Sciences*, Utrecht, Van den Brucke, Cohen eds.
- [3] Hirst, D. J.; Di Cristo, A. (1984) French intonation : A Parametric Approach, *Die Neueren Sprachen*, 83:5
- [4] Rossi, M. (1985) L'intonation et l'organisation de l'énoncé, *Phonetica*, 42.
- [5] Padeloup, V. (1990) Modèle de règles rythmiques du français appliqué à la synthèse de la parole, *Thèse de Doctorat nouveau régime*, Université d'Aix-en-Provence, Aix-Marseille I.
- [6] Padeloup, V. (1988) Essai d'analyse du système accentuel du français: distribution de l'accent secondaire, *17e J. E. P.*, Nancy.
- [7] Fraisse, P. (1974) Psychologie du rythme, PUF, Paris.
- [8] Fraisse, P. (1967) Psychologie des rythmes humains, colloque "Les rythmes", *Journal Français d'Oto-Rhino-Laryngologie*, sup. n° 7, SIMEP.
- [9] Martin, Ph. (1975) Eléments pour une théorie de l'intonation, *Rapp. Inst. Phoné. Bruxelles*, 9 (1).
- [10] Di Cristo, A. et Rossi, M. (1977) Propositions pour un modèle d'analyse de l'intonation, *8e J. E. P.*, Aix-en-Pce.
- [11] Liberman, H.; Prince, A. (1977) On Stress and Linguistic Rhythm, *Linguistic Inquiry*, 8.
- [12] Bruce, G. (1983) On rhythmic alternation, *Working papers of Lund*.
- [13] Dell, F. (1984) L'accentuation dans les phrases en français, *Forme sonore du langage*, Hermann, Paris.
- [14] Martin, Ph. (1986) Structure prosodique et structure rythmique pour la synthèse, *15e J. E. P.*, Aix-en-Pce.
- [15] Padeloup, V. (1990) Multi-style Prosodic Model for French Text-to-speech Synthesis, *Workshop on Speech Synthesis*, Autrans.
- [16] Wioland, F. (1984) Organisation temporelle des structures rythmiques du français parlé, *Bulletin des rencontres régionales de linguistique*, Lausanne.
- [17] Vaissière, J. (1980) La structuration acoustique de la phrase française, *Ann. Scv. Norm. Sup. Pisa*, III (10).

THE ACOUSTIC CHARACTERISTICS OF BOUNDARIES USED IN UTTERING TELEPHONE NUMBERS IN MANDARIN CHINESE, JAPANESE AND ENGLISH

Y. Tsukuma and J. Azuma

Ritsumeikan University, Kyoto, Japan
Kenmei Women's Junior College, Hyogo, Japan.

ABSTRACT

Through the acoustic analyses on boundaries in telephone number utterances (hereafter, T.N.U.) in Mandarin Chinese, Osaka Japanese and American English (eg.18-8333,188-333, where the hyphens denote boundaries), a study was made on some prosodic features which appeared at the boundaries in T.N.U. An observation was also made to determine which prosodic parameters serve as the perceptual factors for the boundaries in T.N.U. by manipulating the F0 parameter and the duration parameter including pauses both of which are considered primary prosodic features. From the experimental results, it is concluded that the speakers of each language employ both the general and language-specific prosodic features to mark the required boundaries in T.N.U. This seems to be based on the unique prosodic characteristics of each language.

1. INTRODUCTION

Prosodic features indicating a major syntactic boundary seem to be language-specific [1][2][3].

In Japanese, the most important prosodic feature to mark a syntactic boundary is the F0 contour resetting at the boundary. Whereas in Mandarin Chinese (also known as Modern Standard Chinese, the term "Chinese" is employed in this paper) and English, pause and preboundary lengthening of a syllable before the syntactic boundary.

The main objective of this study is to investigate if the same prosodic

features marking the syntactic boundary are also employed to mark the boundaries found in a sequence of numbers such as T.N.U., which has stable experimental conditions without semantic and syntactic influences.

2. TEST GROUPINGS OF NUMBERS

TABLE 1. Two Groupings of Numbers Used in the Experiments

Language Examined	A	B
Mandarin Chinese	18-8333	188-333
Osaka Japanese	47-5333	475-333
American English	89-8333	898-333

Chinese monosyllabic words have four lexical tones and Japanese two-mora words have four pitch patterns in their phonological inventory. Both the numbers "yao" for 1 and "ba" for 8 in Chinese carry a high-level tone. In Japanese both the numbers "yon" for 4 and "nana" for 7 carry a high-low pitch pattern, and "goo" for 5 and "san" for 3 carry a high-high pitch pattern.

The grouping of two and four is called A, while that of three and three is called B for the sake of convenience. Thus, the hyphens denote boundaries in the sequence of numbers.

3. EXPERIMENTAL PROCEDURES

The test numbers written in Arabic numerals were uttered by the native speakers of their respective languages in declarative intonation.

TABLE 2. Details of the Subjects

Language Examined	Sex	Age	Birth Place
Mandarin Chinese	♀	37	Beijing
Osaka Japanese	♂	36	Osaka
American English	♂	48	New York

Five tokens were selected from ten repetitions of A and B of each grouping and analysed with the acoustic analysis system using NEC-PC9801RX2. The typical acoustic traces of T.N.U. are shown next.

4. AUDIO SIGNAL AND F0 CONTOURS

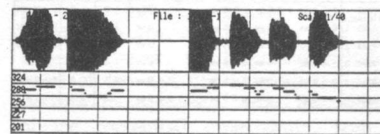


FIG. 1. Grouping A in Chinese

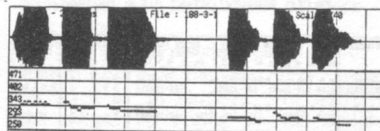


FIG. 2. Grouping B in Chinese

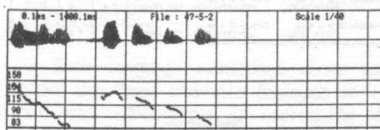


FIG. 3. Grouping A in Japanese

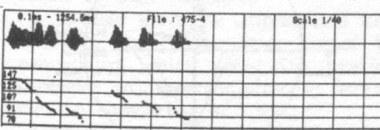


FIG. 4. Grouping B in Japanese



FIG. 5. Grouping A in English

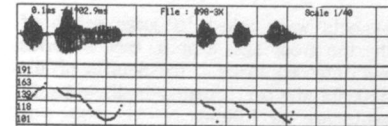


FIG. 6. Grouping B in English

5. PARCOR-SYNTHESIZED STIMULI WITH PAUSE MANIPULATION

In order to investigate to what extent the pause, the stretched duration and the F0 contour resetting at the boundary can contribute to the correct location of the boundaries in T.N.U., a perceptual experiment was carried out.

Using the editing function of the acoustic analysis system, a pause found in groupings A was completely removed in every language. The stretched vowel duration before the boundary as a result of preboundary lengthening was also cut by 100ms and 200ms in Chinese and English. On the other hand, pauses of 200ms and 400ms were inserted between the numbers where there was no pause originally in every language.

6. PARCOR-SYNTHESIZED STIMULI WITH F0 AND PAUSE MANIPULATION

It is known as a physiological fact that resetting the phrase component takes place after a pause or syntactic boundary. Thus, the F0 parameter was manipulated in every language studied. The F0 values after the boundaries in the groupings A were lowered.

The combination of manipulated F0 and pause (including preboundary lengthening in Chinese and English) was made to examine an interrelationship among these prosodic features.

*PARCOR-synthesis (frame: 25.6ms - 256 sample / trapezoid window / 12 bit - 10kHz sampling / LPF 4.5kHz) for this study was conducted by the Speech Processing System developed by Prof. Miyoko Sugito of Osaka Shoin Women's College

7. PERCEPTUAL EXPERIMENT

The PARCOR-synthesized stimuli were randomized and presented over a cassette tape recorder to native speakers of each language for the perceptual tests. The

subjects were asked to judge which of the two groupings, A or B, each stimulus was intended to be. The details of the subjects who participated in the experiments are as follows:

TABLE 3. Details of the Subjects

Language	Sex	Age	Total	Birth Place
Chinese	♂ ♀	20-50	75	Beijing
Japanese	♀	18-19	122	Hyogo
English	♂ ♀	25-55	34	Oregon

8. RESULTS OF PERCEPTUAL EXPERIMENT

Experimental data were analyzed by a microcomputer and a Binominal Test was conducted on the difference between the judgements of each stimulus grouping. The results of the analyses are shown in TABLE 4~9.

TABLE 4. Results of Perceptual Experiment in Chinese (with Manipulated Pause)

STIMULUS	L.B. / PAUSE	5.3.3.3 PAUSE INSERTED BEFORE 3.3.3	NUMBER OF A	NUMBER OF B	SIGNIFICANCE
18-8333 (ORIGINAL)	7.7 0 ms 4.3 2 ms 1.1 0.2 ms	0 ms	7	4	1 P < 0.001
18-8333-1	7.7 0 ms 0 ms 1.1 0.2 ms	0 ms	9	1	P < 0.001
18-8333-2	8.7 0 ms 0 ms 1.1 0.2 ms	0 ms	4	2	N.S.
18-8333-3	5.7 0 ms 0 ms 1.1 0.2 ms	0 ms	3	4	N.S.
18-8333-4	7.7 0 ms 0 ms 1.1 0.2 ms 2.0 0 ms	2.0 0 ms	3	4	N.S.
18-8333-5	7.7 0 ms 0 ms 1.1 0.2 ms 4.0 0 ms	4.0 0 ms	1	5	P < 0.001

TABLE 5. Results of Perceptual Experiment in Chinese (with Manipulated FO & Pause)

STIMULUS	L.B. / PAUSE	5.3.3.3 PAUSE INSERTED BEFORE 3.3.3 (Fo of 5.3.3.3)	NUMBER OF A	NUMBER OF B	SIGNIFICANCE
18-8333 (ORIGINAL)	7.7 0 ms 4.3 2 ms 1.1 0.2 ms	0 ms	7	4	1 P < 0.001
18-8333-6	7.7 0 ms 0 ms 1.1 0.2 ms (3.15-3.06 Hz, 2.06-1.88 Hz)	0 ms	5	2	P < 0.01
18-8333-7	7.7 0 ms 0 ms 1.1 0.2 ms 2.0 0 ms (3.15-3.06 Hz, 2.06-1.88 Hz)	2.0 0 ms	4	3	N.S.
18-8333-8	7.7 0 ms 0 ms 1.1 0.2 ms 4.0 0 ms (3.15-3.06 Hz, 2.06-1.88 Hz)	4.0 0 ms	6	6	P < 0.001

TABLE 6. Results of Perceptual Experiment in Japanese (with Manipulated Pause)

STIMULUS	L.B. / PAUSE	5.3.3.3 PAUSE INSERTED BETWEEN 5.3.3	NUMBER OF A	NUMBER OF B	SIGNIFICANCE
47-5333 (ORIGINAL)	4.2 8 ms 1.2 0 ms 8.8 8 ms	0 ms	1	2	0 P < 0.001
47-5333-1	4.2 8 ms 0 ms 8.8 8 ms	0 ms	1	1	1 P < 0.001
47-5333-2	4.2 8 ms 0 ms 8.8 8 ms 2.0 0 ms	2.0 0 ms	4	4	7 P < 0.05
47-5333-3	4.2 8 ms 0 ms 8.8 8 ms 4.0 0 ms	4.0 0 ms	3	5	7 P < 0.001

TABLE 7. Results of Perceptual Experiment in Japanese (with Manipulated FO & Pause)

STIMULUS	L.B. / PAUSE	5.3.3.3 PAUSE INSERTED BETWEEN 5.3.3 (Fo of 4.7)	NUMBER OF A	NUMBER OF B	SIGNIFICANCE
47-5333 (ORIGINAL)	4.2 8 ms 1.2 0 ms 8.8 8 ms	0 ms	1	2	0 P < 0.001
47-5333-4	4.2 8 ms 1.2 0 ms 8.8 8 ms (1.4 4 - 8.7 Hz, 1.3 5 - 1.2 6 Hz)	0 ms	1	3	9 P < 0.001
47-5333-5	4.2 8 ms 0 ms 8.8 8 ms (1.4 4 - 8.7 Hz, 9.5 - 8.5 Hz)	0 ms	8	1	14 P < 0.001
47-5333-6	4.2 8 ms 0 ms 8.8 8 ms 2.0 0 ms (1.4 4 - 8.7 Hz, 9.5 - 8.5 Hz)	2.0 0 ms	0	2	2 P < 0.001
47-5333-7	4.2 8 ms 0 ms 8.8 8 ms 4.0 0 ms (1.4 4 - 8.7 Hz, 9.5 - 8.5 Hz)	4.0 0 ms	1	2	1 P < 0.001

TABLE 8. Results of Perceptual Experiment in English (with Manipulated Pause)

STIMULUS	L.B. / PAUSE	5.3.3.3 PAUSE INSERTED BEFORE 3.3.3	NUMBER OF A	NUMBER OF B	SIGNIFICANCE
89-8333 (ORIGINAL)	8.5 4 ms 1.9 2 ms 9.9 8 ms	0 ms	3	3	0 P < 0.001
89-8333-1	8.5 4 ms 0 ms 9.9 8 ms	0 ms	3	2	1 P < 0.001
89-8333-2	8.5 4 ms 0 ms 9.9 8 ms	0 ms	3	1	2 P < 0.001
89-8333-3	4.5 4 ms 0 ms 9.9 8 ms	0 ms	19	14	N.S.
89-8333-4	8.5 4 ms 0 ms 9.9 8 ms 2.0 0 ms	2.0 0 ms	2	5	8 P < 0.05
89-8333-5	8.5 4 ms 0 ms 9.9 8 ms 4.0 0 ms	4.0 0 ms	9	2	4 N.S.

TABLE 9. Results of Perceptual Experiment in English (with Manipulated FO & Pause)

STIMULUS	L.B. / PAUSE	5.3.3.3 PAUSE INSERTED BEFORE 3.3.3 (Fo of 8.5, Fo of 5.3.3.3, Fo of 8.3.3.3)	NUMBER OF A	NUMBER OF B	SIGNIFICANCE
89-8333 (ORIGINAL)	8.5 4 ms 1.9 2 ms 9.9 8 ms	0 ms	3	3	1 P < 0.001
89-8333-6	8.5 4 ms 0 ms 9.9 8 ms (144-119-133Hz, 138-158Hz, 1212-187Hz, 0 ms 9.9 8 ms, 187-147-177Hz, 197-180Hz)	0 ms	2	1	2 N.S.
89-8333-7	8.5 4 ms 0 ms 9.9 8 ms 2.0 0 ms (1212-187Hz, 187-147-177Hz, 197-180Hz)	2.0 0 ms	2	3	1 P < 0.001
89-8333-8	8.5 4 ms 0 ms 9.9 8 ms 4.0 0 ms (1212-187Hz, 187-147-177Hz, 197-180Hz)	4.0 0 ms	0	3	3 P < 0.001
89-8333-9	8.5 4 ms 0 ms 9.9 8 ms (1212-187Hz, 187-147-177Hz, 197-180Hz)	0 ms	2	4	9 N.S.
89-8333-10	4.5 4 ms 0 ms 9.9 8 ms (1212-187Hz, 187-147-177Hz, 197-180Hz)	0 ms	1	2	N.S.

From the above TABLES and FIGURES, the following observations are made for each language.

Chinese: the boundaries in T.N.U. in Chinese show the following two characteristics; placing a pause at the boundary, lengthening the syllabic vowel of the last digit before the pause as an effect of preboundary lengthening. The perceptual experiments prove that these two prosodic features are found significant, whereas the Fo parameter does not account for the presence of the boundaries in T.N.U. for listeners of this language in comparison with Japanese.

Japanese: the boundaries in T.N.U. in Japanese are made with the same two characteristics as in Chinese. Moreover, resetting the Fo contour after the boundary is also prominent in this language. However, the lengthened syllabic vowel of the last digit before the pause is rare in Japanese. From the results of the perceptual experiments, both a pause insertion and an Fo contour resetting account for the presence of the boundaries in T.N.U. for listeners in this language.

English: the boundaries in T.N.U. in English also show the same three characteristics as in Chinese. In addition, the unique Fo rising of the syllabic vowel of the last digit before the boundary is observed in this language.

9. CONCLUSION

Each language mentioned above employs the three prosodic features in its own way respectively in expressing the boundaries in T.N.U., which seems to be based on the unique prosodic characteristics of each language.

In other words, as Chinese is a tone language, the Fo parameter cannot be manipulated, which makes its duration parameter an important prosodic feature for the boundaries in T.N.U.

In Japanese, as it is a mora-timed language, its duration parameter cannot be manipulated. Thus, an Fo parameter function as an important prosodic feature for the boundaries in T.N.U.

In English, as it is a stress-timed language, its duration parameter (both a pause insertion and the preboundary lengthening) is used as in Chinese to mark the boundaries in T.N.U. However,

as the Fo parameter in English is less exhaustively employed to make semantic differences than in Chinese, both the duration parameter and the Fo parameter are used for the boundaries in T.N.U.

However, it should also be noted that for listeners of every language we studied, all of these three prosodic features as an interactive effect could account for the presence of the boundaries in T.N.U.

These experiments also seem to prove that the prosodic characteristics for the boundaries in T.N.U. coincide with those for ordinary sentences in each language[1][2][3].

REFERENCES

- [1] AZUMA, J. and TSUKUMA, Y. (1990), "Prosodic Features Determining the Comprehension of a Syntactically Ambiguous Japanese Sentence: In the Case of the Kinki Dialect," *Syntactic Structure and Prosodic Features*, 24-33. (Report on "Prosodic Features of the Japanese Language" - Grant-in-aid for Scientific Research on Priority Areas, the Ministry of Education, Science and Culture of Japan)
- [2] LEHISTE, I. (1975), "Role of Duration in Disambiguating Syntactically Ambiguous Sentences," *JASA*, 60, 5, 1199-1202.
- [3] TSUKUMA, Y. and AZUMA, J. (1990), "Prosodic Features Determining the Comprehension of Syntactically Ambiguous Sentences in Mandarin Chinese," *Proceedings of ICSLP 90*, 505-508.

*This work was supported by the Ministry of Education, Science and Culture of Japan (Grant-in-aid for Scientific Research on Priority Areas "Prosodic Features of the Japanese Language" represented by Prof. Miyoko Sugito: Grant Nos. 01642504 and 02224204).

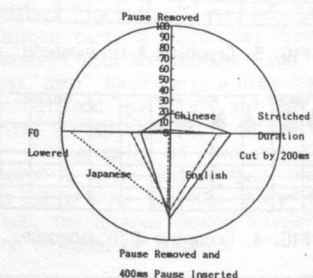


FIG. 7. Radar Chart of Perceptual Effects of Four Prosodic Features in Chinese, Japanese and English.

Isolde Wagner

Institut für Phonetik und
Sprachliche Kommunikation
Ludwig-Maximilians Universität München, FRG

ABSTRACT

Preliminary results show some influence of sentence length and focus related accent position on Fo-declination in read German declarative utterances. For the declination line linear regression lines have been computed. An additional measurement was made on the first Fo-maximum of each declining contour. The slope of the regression line decreases with increasing utterance duration, and the first Fo-maximum of long sentences is lower when the focal accent has been placed on the second Fo-maximum.

1. INTRODUCTION

During the last decade special attention has been paid again to a phenomenon of pitch behavior on sentence level, the so-called declination of the fundamental frequency contour and a simultaneous lowering of the peaks of accented syllables towards the end of declarative sentences. Depending on different types of investigations two general models explaining the phenomenon have been put forward. One refers to the underlying physiological mechanisms and thus tries to explain the phenomenon of declination by the decrease of the subglottal pressure [1], where-

as the other model - mainly found in acoustical and perceptual investigations - often assumes a "pre-planning" strategy of the speaker, which would explain the finding that at least in read sentences of different length in some languages, e.g. Danish, the Fo-onset increases and the steepness of the slope decreases with the duration of the utterance [4]. Similar relations are expected for German. Using varied utterance duration and focus related accent position, variations of the slope and of the first Fo-maximum as one starting point of Fo-contours are expected.

2. MATERIAL AND INFORMANTS

The material consisted of 36 simple sentences, divided into four blocks of utterance triples of different length (6, 10, and 14 syllables). Each block consisted of one short sentence amended twice with additional information at the end. This will be shown in the following:

- 1: *Morgen kommt Maria.* ...
 - 2: *... mit dem Auto.* ...
 - 3: *... nach Hannover.*
- Additionally different focal accent positions (1st, 5th, 9th, and 13th syllable) were posed on the utterance triples, contextually controlled by appropriate ques-

tions, e.g.:

- q: *Wann kommt Maria?*
a: *Morgen kommt Maria.*
- Four tokens of all sentences were read aloud in random order by each of three male native German speakers who were students.

3. PROCEDURE

The recorded material was acoustically analysed with respect to the Fo-variation by means of an LPC-analysis. Best-fit all-points linear regression lines have been computed, as well as the mean Fo of each intonation contour. Additional measurements were made on the first Fo-maximum and the intersection between the regression line and the y-axis. The actual duration of each sentence in ms should give further information, and the offset was expected to keep a speaker-dependent equal level.

Tab.1: FACTORS AND VARIABLES.

- F1: sentence length
(short, medium, long)
F2: accent position
(2 in short, 3 in medium, 4 in long sentences)
V1: slope of the regression line
V2: first Fo-maximum
V3: intersection of the regression line with the y-axis
V4: mean Fo of the intonation contour
V5: sentence duration

4. GENERAL RESULTS

Tab.1 shows the two factors and the five dependent variables used for the analysis. First a survey of overall means, illustrated by the figures, is given to show general trends. Then the factors have been statistically computed for each speaker separately. While the general behavior of the three speakers is quite

similar, there are some differences in detail. The design of the material required a "ONEWAY"-analysis. Since the analysis was calculated twice over the same data set, the significance level was lowered according to the "Bonferroni" procedure to $\alpha=0.025$. An a posteriori-test ("SCHEFFE") followed to find the significant differences between each of two variables within the same factor. Here the significance level was also lowered to the value $\alpha=0.025$.

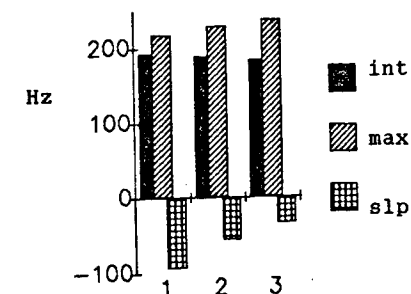


Fig.1: SENTENCE LENGTH means of intersection, first Fo-maximum and slope: 3 types of sentence length / 1 accent position on the 1st syllable / all speakers

4.1. Varied sentence length
Here the accent position was always on the first syllable. As expected, Fig.1 shows a substantial decrease of the slope with increasing utterance length. There is a main effect for all informants:

- M: $F(2,47)=61.2813$; $P<.001$;
J: $F(2,47)=27.4799$; $p<.001$;
R: $F(2,45)=54.3722$; $p<.001$.
- But there is only a weak increase of the first Fo-maximum as well as in the decrease of the intersection. The mean Fo of the intonation contours is not remarkable, and the duration in ms just illustrates the actual

duration of the linguistically varied sentences.

4.2. Varied accent position
The slopes on Fig.2 to Fig.4 also decrease as the focal accent is moved towards the end of an utterance. But in the long sentences on Fig.4 a significant difference can only be detected between the earlier and the later accent positions:

M: $F(3,62)=11.6998$; $p<.001$;
J: $F(3,63)=3.6151$; $p<.025$;
R: $F(3,61)=36.4458$; $p<.001$.

Accordingly to the decreasing slope the mean Fo of the declining contours increases significantly for each informant and ends up on a speaker-individual but equal level in all sentences.

The first Fo-maximum of Fig.3 and Fig.4 shows a peculiar, more or less speaker dependent lowering, when the accent has been placed on the 5th syllable (the second position). And it increases again as the accent is moved towards the end of an utterance.

The intersection does not show any notable variation, while the sentence duration in ms increases feebly with the accent at the end.

5. SOME INDIVIDUAL RESULTS

5.1. Speaker M

Regarding the first Fo-maximum speaker M does not show any main effect neither on the factor "sentence length" nor on the factor "accent position". Just when he produces the long sentences with focal accent on the second position the first Fo-maximum approaches a significant lowering
($F(3,62)=3.2015$; $p=.0297$).

5.2. Speaker J

Speaker J does show significant differences on the first Fo-maximum when the



Fig.2: ACENT POSITION (2)
means of intersection,
first Fo-maximum and slope :
short sentences / 2 accent
positions / all speakers

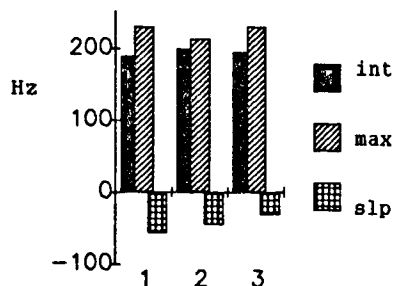


Fig.3: ACENT POSITION (3)
means of intersection,
first Fo-maximum and slope :
medium sentences / 3 accent
positions / all speakers

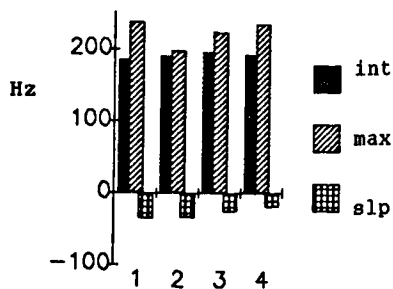


Fig.4: ACENT POSITION (4)
means of intersection,
first Fo-maximum and slope :
long sentences / 4 accent
positions / all speakers

accent position is varied. In short sentences the first Fo-maximum increases significantly when the accent is placed on the second and thus on the last position

($F(1,31)=9.4327$; $p<.025$), and in long sentences the first Fo-maximum decreases extremely significant when the accent is placed on the second position

($F(3,63)=9.2232$; $p<.001$). Medium sentence length does not show any effect.

5.3. Speaker R

For speaker R both factors "sentence length" and "accent position" show a high influence on all variables. That is not valid for speaker M and J.

Concerning the first Fo-maximum, it is significant low with the accent on the second position in medium and long sentences

($F(2,47)=6.9829$; $p<.025$;
 $F(3,61)=14.7907$; $p<.001$), and there is, contrary to speaker J, a lowering tendency in short sentences with second accent position
($F(1,30)=5.2496$; $p=.0294$).

6. DISCUSSION

So far as has been tested in this investigation, sentence length and focal accent position exercise influence on some parameters relating to the declination phenomenon. One of them is the slope of the declining contour, computed as a global all-points linear regression line. It decreases significantly with increasing utterance duration, and with varied accent position towards the end of an utterance. The first Fo-maximum only shows some significant effect with varied accent position. This means that, if the accent has a medium position the first Fo-maximum is lower than in the other cases. Concerning

varied sentence length the first Fo-maximum increases weakly but not significantly in longer utterances.

This investigation is a preliminary study which has not been totally completed yet. Hence one should be cautious about interpretation. It seems more important at this stage to point out problems remaining in this kind of analysis. One of them is the discussion about the use of an all-points linear regression line rather than a top-line or a baseline. The global regression line can be affected by a late focal accent and thus increases, while the declination contour decreases [2, 3].

Furthermore concerning this investigation, at the next stage all the focal accents must be cut off before computing linear regression lines, and then be measured separately. In this way one could obtain better information about the influence of varied accent position on the declination line.

7. REFERENCES

- [1] GELFER, C. (1987), *A simultaneous physiological and acoustic study of fundamental frequency declination*. Phil. Diss., CUNY, New York.
- [2] 't HART, J. (1986), "Declination has not been defeated - A reply to Lieberman et al.", *J. Acoust. Soc. Am.* 80, 1838-1840.
- [3] LIEBERMAN, P. et al. (1985), "Measures of the sentence intonation of read and spontaneous speech in American English." *J. Acoust. Soc. Am.* 77, 649-657.
- [4] THORSEN, N. (1980), "Intonation contours and stress group patterns in declarative sentences of varying length in ASC Danish." *ARIPUC* 14, 1-29.

THE REPRESENTATION OF INTONATION IN
MANDARIN CHINESE

Jialing Wang

Tianjin Normal University, China

ABSTRACT

This paper proposes a three-dimensional model for the representation of intonation in Mandarin Chinese. In this model, I use two different types of features, which appear on different planes. Hierarchically-related tonal features [Upper] and [Raised] are used to represent tones. Pitch range features [Expanded Range] and [Raised Range] deliver the intonational meaning. [Expanded Range] is related to Focus; [Raised Range] has to do with expressiveness, such as questions. The paper discusses the representation of the neutral tone, focus, and questions within the framework of this model.

1. INTRODUCTION

Mandarin Chinese is a tone language with four lexical tones. These distinguish lexical meaning and are specified in the lexicon, as shown in (1):

(1) Tone	Pitch Value	Ex.	Gloss
1	HH	ma ¹	mother
2	MH	ma ²	hemp
3	LLH'	ma ³	horse
4	HL	ma ⁴	scold

Since lexical tone and intonation are both characterized by pitch, their relationship has been a constant issue in the analysis of Chinese intonation. In this paper I use two different types of features in a three-dimensional model

(Halle & Vergnaud 1987) to represent tone and intonation.

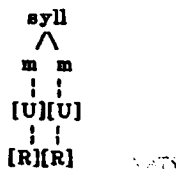
2. THE MODEL

I follow Yip(1989) in assuming two coplanar features on the tonal plane: the register feature [Upper] and the sub-register feature [Raised], which are hierarchically-related, as in (2):

(2)	+Upper	+Raised	H
		-Raised	H'
-Upper	+Raised	M	
	-Raised	L	

I differ from Yip in taking the mora as the tone-bearing unit as well as a timing unit, as shown in (3)

(3)



I follow the application of underspecification theory by Pulleyblank (1986) in assuming that the universal default values for the two tonal features are [-U] and [+R]. So these two features are not specified in the lexicon. Thus the underlying representation of

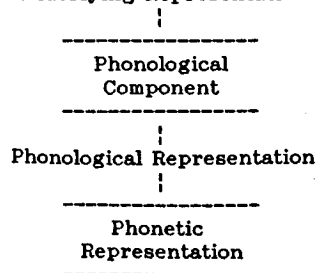
the four tones are Tone 1: [+U], [+U]; Tone 2: [-R], [+U]; Tone 3: [-R], [-R]; Tone 4: [+U], [-R].

The above four tones occur only with stressed syllables. Unstressed syllables do not bear any of the four tones, but are said to have a neutral tone. In underlying representation, the neutral tone is represented as having one mora and no tonal features, as shown in (4):



In this model I recognize a phonological component and a phonetic component on the tonal plane. The phonological component consists of a set of phonological rules which bring about a change in tonal category. These are the Tone Sandhi Rule (which change Tone 3 to Tone 2 when it occurs before another Tone 3) and the Tone Deletion Rule (which changes the tone of unstressed syllables to neutral tones). The input to the phonological tonal component is the sequence of underlying tones with stress assigned. Stress is represented on another plane (stress plane) by metrical grids. The output of the phonological component is the phonological tonal representation, which is the input to the phonetic component. This component contains a set of allophonic rules deriving the phonetic tonal representation. This is shown in (5):

(5) Underlying Representation



Phonetic Representation

The phonetic tonal representation constitutes the basic pattern upon which pitch range features (which deliver intonational meaning) interact. Pitch range features include [Expanded Range] and [Raised Range], which are on a separate plane, and linked to the relevant syllable by association lines. They expand or raise the pitch range of the tone linked with that syllable via the mora.

3. Neutral Tones

The above model can adequately resolve the issue of the representation of the neutral tone, which has been largely ignored or over-simplified in previous literature.

Traditionally, the pitch value of the neutral tone is said to be solely determined by the tone of the preceding syllable: it is H' when it is preceded by Tone 3, M when preceded by Tone 2 and L when preceded by Tone 1. But this has not taken into account the full range of positions that the neutral tone may be in. First, it does not take into account the distinction between the prepausal and non-prepausal positions of the neutral tone. Secondly, it does not take into account the fact that neutral tones can occur consecutively in a sequence. And lastly, neutral tones can sometimes appear in initial position.

Taking the above facts into consideration, we get the following data:

(6)

a. Neutral tone in prepausal position

Preceding Tone	Neutral Tones		
	1	2	3
Tone 1	L	ML	MLL
Tone 2	L	ML	MLL
Tone 3	H'	H'L	H'ML
Tone 4	L	LL	LLL

b. Neutral tone in non-prepausal position

PROSODIC ITALICS: FUNCTIONS AND PHONETIC REALIZATION

A.Panasyuk and I.Panasyuk

Leningrad State University, USSR

ABSTRACT

A phonetic study of italics in English language literary texts has shown that this graphical means of expressing emphasis is very useful in making intonationally unambiguous those utterances which may have more than one interpretation in respect of the nucleus placement and type of tone used.

Traditionally, studies of graphical means of emphasis used in a given language have not been included into the field of phonetics, but rather regarded as part of stylistics [15,18,19,23]. Since graphical means, such as italics, carry a lot of valuable information about the intonational structure of an utterance, they may be as well regarded as subject matter of intonology, a branch of phonetics dealing both with the sound form and the semantic load of speech utterances.

It is not surprising that the English language which makes enormous use of intonation in rendering various meanings should favour the use of italics much more than any other language. It was Maria Schubiger who first pointed out that italics are often found in English sentences (in literary texts) where the placement of nuclear stress is determined only by context and is not signalled either by the syntactical construction or by a modal particle, as is the case in French and German [16].

Despite their frequent use by

authors of novels and stories, italics are not approved of by many stylists. For example, in "The King's English" by H.W.Fowler and F.G.Fowler we come across a point of view that "italics are a confession of weakness" and are employed mostly by those writers "who, regarding the reader's case as desperate, assist him with punctuation, italics and the like" [6]. A more realistic opinion of italicization is found in "The ABC of Style" by R.Fleisch who states that "if you don't use italics, you're missing one of the best resources of writing; if you use too many, you spoil the effect you're after. The basic rule is to underline (for italics in print) the words that would get heavy natural stress in speaking - and not to shy away from the colloquial sentence pattern that calls for such stress" [5].

It seems to be worth mentioning that not only scholars stress the intonational significance of this graphical means of emphasis but creative writers themselves, who make practical use of italics, often comment on this subject. In "1984" by G.Orwell we find such a commentary: "... Meanwhile I shall send you a copy of the book." even O'Brien, Winston noticed, seemed to pronounce the words as though they were in italics" [14,p.146].

Another author, L.M.Montgomery, describing her heroine as a sweet-souled lass, states that " she could instil some venom into innocent italics when

occasion required" [12,p.88].

From the given above examples we can see that italics are used for emphasizing a particular word which has a special meaning in the context as well as for showing that a special kind of intonation is to be chosen when pronouncing the sentence.

Thus, italics may be said to possess two prosodic functions, namely, the function of indicating an unusual position of the sentence stress, and that of showing that the nucleus (even though in its predicted position) is to be realised with an unusual (emphatic) intonation. In most cases, however, both functions are combined and the italicized word is located in an unpredicted position and marks an unusual tone.

Very often, the use of italics is accompanied by an author's remark on the unusual position of stress, e.g. "There was the unusual mellifluous murmur from the loudspeaker about seatbelts, emergency exits, oxygen masks. He wondered why stewardesses accented such unlikely words: "On our flight this evening we will be offering..." [17,p.31].

Another example of the author's commentary on the use of a "special tone" italics is given below. "In tones of loud and hearty excitement Miss Pilchester, who had forgotten to close the door, confesses that she literally didn't know. It was all such a thrill, so absolutely unexpected. Had she been an age?" [1,p.70].

Sometimes, capital letters are employed in place of italics: "Lenore talked haut-American, a fast anglicized gabble which lit on one word now and then for emphasis. In the Brompton house she'd said to me: "It's so PEACEFUL here. D'you know, Peeder, for many years we stayed at Brown's Hotel because it's so ENGLISH" [13,p.174].

Unusual placement of an italicized word often points to a contrast between

two or more elements within the nearest verbal or situational context. Contrastive italics may be placed on a normally unstressed word (function words), or even a prefix, or any other word which is capable of carrying a contrastive meaning. Consider the following example: "I thought you said it was all very formal." "Yes. She's not usually formal. Why should she be like that? She's so direct as a rule: not exactly informal, ever, but absolutely direct" [9,p.82].

Here a complex contrast is made possible with the help of italicization.

Much more often, however, italics are used with function words, such as auxiliary verbs, pronouns, etc. In order to find out parallels to this phenomenon in other languages, we analyzed a number of translations into English and selected those sentences which contained italicized auxiliaries (in the English version).

The following example has been taken from "Lillebror och Karlsson på Taket" by Astrid Lindgren and its English translation.

"Mellanmal forstor aptiten" sa hon. "Har blir inga bullar". Och ändå hade hon bakat bullar. Det stod ett helt fat i det öppna fönstret för att svalna. [10,p.48].

"Snacks between meals ruin your appetite," she said. "There will be no buns here". She had baked buns. There was a whole dish of them on the window-sill. [11,p.38].

In the Swedish sentence italics are on the meaningful part of the predicate (participle). Additionally, there is an intensifying particle and the inverted word order which contribute to the strengthening of emphasis on "bakat". The corresponding English sentence lacks any other means but italics on the auxiliary to bring forth the contrast between the actual existence of buns and their unavailability for the boy.

Another example is from the Russian book "The Golden Calf" by I. Ilf and E. Petrov. -Вы не читали Блейлера? - спросил Кай Юлий удивленно. - Позвольте, по каким же материалам вы готовились? [20, 362].

"You haven't read Bleyler?" asked Caius Julius in surprise. "Excuse me, but with what material did you prepare yourself?" [8, p.189].

As can be seen, there are no italics in the original Russian text. The logic stress is on the word "материалам". This position of the stress is determined by the presence of the particle "же". In the English variant, there is a shift of the nucleus onto the auxiliary "did", which is an equivalent to the Russian particle.

Putting italics on pronouns is also quite frequent. We'd like to give here an illustration from "Alice in Wonderland" by Lewis Carroll, an author who is known to have used italics abundantly. The following example contains contrastive italics on a personal pronoun: "I can't help it," said Alice very meekly: "I'm growing". "You've no right to grow here," said the Dormouse. "Don't talk nonsense," said Alice boldly: "You know you're growing too". "Yes, but I grow at a reasonable pace," said the Dormouse. [2, p.144].

We've analysed the translations of this book into a number of languages, including French, Russian, and Estonian.

In the French translation, the sentence in question has no italics, but it contains a stressed pronoun "moi" preceding the unstressed one: -Oui, mais moi, je grandis a une vitesse raisonnable... [4, p.168].

The translation into Estonian gives evidence to a similar tendency, i.e., a stressed form of pronoun is used (which is normally omitted). Additionally, the final word, an adverb of manner, is in italics: "Seda küll, aga mina kasvan mõistlikult" [3, p.91].

The two Russian translations analysed

reveal two different tendencies. In the translation by V. Nabokov the pronoun is omitted altogether. In the translation by N.M. Demurova the place of italics is preserved.

-Да, но разумным образом, возразил Соня, - не раздуваюсь, как вы. [22, 186]. -Да, но я расту с приличной скоростью, - возразила Соня, - не то что некоторые... [21, с.90].

Perceptual and acoustic analysis of English sentences containing italicized words which were spoken by English speakers has shown that in most cases they had an emphatic tone, either a Rise-Fall or a High Fall. Such words were easily identified by trained Russian phoneticians as nuclei carrying one of these tones. The intonogram in Fig.1 is an illustration of a well perceived emphatic tone on the italicized word

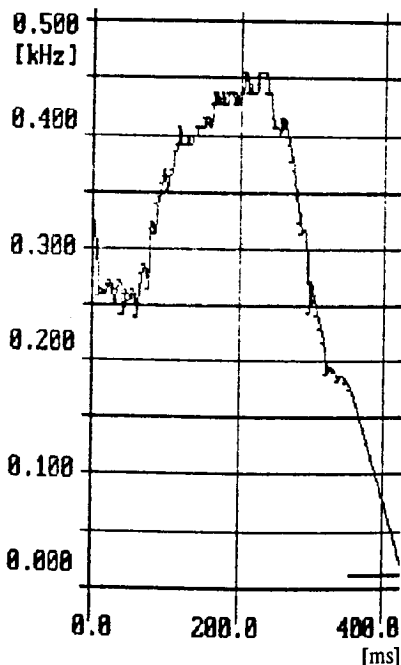


Fig.1 Intonogram of the word "there"

"there" from the sentence "What lies over there?" [7].

There are some cases, however, when the italicized word is not perceived as a nuclear one. For example, the word "only" in the sentence "It's the only thing" (see Fig.2) is perceived by all listeners as stressed while the final word "thing" is identified with the nucleus.

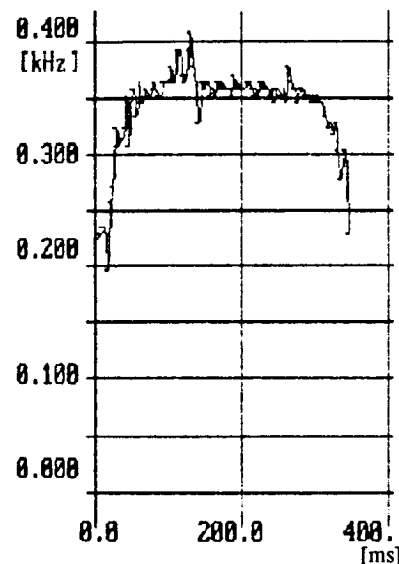


Fig.2 Intonogram of the word "only"

In conclusion, it should be said that use of italics reflects a very complicated system of accentuation which is characteristic of the English language. Italics are effective means of signalling a shift of the sentence stress which otherwise may be overlooked by the reader.

An extensive study of italicization may be very valuable in understanding the semantic functions of intonation.

REFERENCES

- [1] BATES, H.E. (1977), "Oh! To be in England".
- [2] CARROLL LEWIS (1966), "Alice's Adventures in Wonderland", Penguin Book.
- [3] CARROLL LEWIS, "Alice

Imedemaal". Tallinn.

- [4] CARROLL LEWIS "Alice Au Pays des Merveilles", Paris.
- [5] FLEISCH, R. (1977), "Look it up", A desk book of American spelling and style. London.
- [6] FOWLER, H.W., FOWLER, F.G. (1927), "The King's English", Oxford.
- [7] GRAHAME, K. "The wind in the Willows".
- [8] ILF, I., PETROV, E. "The Golden Calf".
- [9] LEHMAN, R. "The Echoing Grove".
- [10] LINDGREN, A. "Lillebror och Karlsson Pa Taket". Stockholm.
- [11] LINDGREN, A. "Midge and Karlson on the Roof", London.
- [12] MONTGOMERY, L.M. (1968), "Anne of Avonlea", Toronto.
- [13] NICHOLS, P. (1985), "Feeling You're Behind", Penguin Book.
- [14] ORWELL, G. (1989), "1984", NY.
- [15] PARTRIDGE, E. (1977), "Have a Point There". A Guide to Punctuation and its allies. London.
- [16] SCHUBIGER, M. (1965), "English Intonation and German Modal Particles". *Phonetica*, vol.12, pp.65-84.
- [17] TYLER, A. (1985), "The accidental Tourist", NY.
- [18] АРНОЛЬД, И.В. (1973), "Графические стилистические средства". *Иностранные языки в школе*, 3, с.13-20.
- [19] КОЛЕГАЕВА, И.М. (1986), "Экспрессивная графика в оригинале и переводе художественной и научной прозы." В книге: *Контрастивное исследование оригинала и перевода художественного текста*. Одесса.
- [20] ИЛЬФ, И., ПЕТРОВ, Е., "Золотой теленок".
- [21] КЭРРОЛЛ ЛЬЮИС, (1990), "Алиса в стране чудес". (пер. Демуровой И.М.). Москва, "Наука".
- [22] КЭРРОЛЛ ЛЬЮИС, "Аня в стране чудес". (пер. В.Набокова). Ленинград, "Детская литература".
- [23] ПЕТРОВСКИЙ, К.В. (1981), "Стилистическая графика англо-американской прозы XIX-XX веков". Автореферат канд.дисс., Киев.

N. D. Svetozarova

Université de Leningrad, Leningrad, URSS

ABSTRACT

It is common to describe intonation differences using sentences of the same or seemingly the same lexico-grammatical structure. However there are many characteristic intonation contours which exist only in combination with peculiar, sometimes irregular, or odd structures. A distinguishing feature of Russian intonation is its abundance of intonational stereotypes, the prosodic structure of which can be described only in conjunction with their unique lexico-grammatical structure. The study of these Russian intonational idioms reveals the importance of special prosodic means, such as different phonation types, lengthening and shortening of vowels and consonants, level tones etc.

1. INTRODUCTION

Les auteurs de presque tous les écrits sur l'intonation essaient de montrer son autonomie et ses propres moyens distinctifs. On le fait généralement en prêtant des contours intonatifs différents à une même séquence de mots. Il est facile à voir que les séquences de mots dans les exemples qu'on cite ne sont homonymes qu'au premier abord: leur structure syntaxique n'est pas la même, les acceptions des mots diffèrent aussi. Des exemples "purs" sont difficiles à trouver, mais ce fait ne prouve point que le rôle de l'intonation soit en quelque sorte diminué.

En russe c'est l'intonation qui est principalement chargée de dis-

tinguer les types de phrases et de segmenter la phrase en thème-rhème. Cela s'explique par l'emploi facultatif des moyens lexicaux et grammaticaux (particule interrogatives, mode du verbe, l'ordre des mots):

(1) Он придет? - Он придет. - Est-ce qu'il va venir? - Il va venir.

(2) Читать! - Читать?

Lire! - Dois-je lire?

(3) Поезд пришел. - Поезд пришел. Le train est arrivé. - C'est le train qui est arrivé.

Il y a aussi des faits qui laissent voir une certaine sélectivité de l'intonation envers les moyens lexicaux et grammaticaux. Une des particularités de l'intonation consiste en ce qu'en même temps elle ne dépend pas des mots et de la syntaxe et en dépend. En intonologie les faits qui témoignent de l'autonomie de l'intonation sont beaucoup plus exploités que les faits opposés. Je crois que cela est dû à l'orientation vers la fonction distinctive et non pas constitutive des unités intonatives. Une pareille orientation dans la phonétique segmentale vers la fonction distinctive du phonème a freiné les recherches sur les rapports entre le sens et la forme phonique des unités de langue. L'école phonologique de Leningrad (école de L. V. Ščerba) prévoit qu'on considère comme principale la fonction constitutive des unités phonétiques et non pas la fonction distinctive, ce qui fait voir sous une autre optique les rapports entre le son et le sens.

Le russe offre, comme peut-être toute autre langue, une grande quantité d'exemples où certaines intonations ne sont combinées qu'avec certaines classes de lexèmes ou des constructions syntaxiques. Quelques cas sont décrits dans le cadre des fonctions proprement linguistiques de l'intonation. Ainsi, dans la classification des constructions intonatives (IK) par E. Bryzgunova la construction émotionnelle-appréciative IK-7 et, en partie, IK-5 montrent une distribution limitée [1]. D'autres cas de sélectivité sont liés à d'autres fonctions de l'intonation, p. ex. à la fonction figurative. Tel est l'emploi d'un ton bas pour rendre l'idée de ce qui est grand et d'un ton haut de ce qui est petit; l'accélération du débit pour rendre l'idée d'une action dynamique etc.

2. LOCUTIONS PHRASÉOLOGIQUES INTONATIVES

Le russe parlé possède un grand nombre de formules-réactions de types différents qui sont caractérisées par une structure grammaticale incomplète, irrégulière, sont sémantiquement appauvries et quelquefois même vides de sens.

Cependant dans la communication orale ces formules acquièrent une signification bien déterminée grâce à l'intonation. Leur sens intègre et stable que l'on ne peut pas déduire des mots qui les composent, leur caractère métaphorique, expressif et émotionnel fait classer ces formules parmi les unités phraséologiques [2]. Le rôle particulier qui revient à l'intonation dans leur fonctionnement permet de les caractériser comme des locutions phraséologiques intonatives (LPhI).

Ce type de formules et, en particulier, leurs caractéristiques intonatives n'ont pas été suffisamment étudiées (voir cependant [3, 4, 5, 6]).

1. Il y a des substituts prosodiques de quelques interjections et adverbies; ces substituts peuvent être classés parmi des LPhI qui ne sont rendus dans les textes russes que d'une façon

approximative comme "yry", "ydy", "xmx" etc. Par exemple:

(4) [- ˘] yry = да (oui)

(5) [˘?] ne-a = нет (non)

Ils sont pareils aux gestes (de consentement, de refus, de perplexité etc.) par leur fonctionnement et leur caractère non-verbal, ce qui permet de les considérer comme des gestes phoniques ou bien des gestes intonatifs.

2. Des clichés intonatifs sont des complexes dans lesquels des séquences de mots, qui sont souvent sémantiquement appauvris, n'existent qu'avec une intonation spécifique. Les exemples ci-dessous n'ont pas de sens, sinon avec une intonation particulière qui leur donne un sens affectif:

(6) Ну и ну! - l'étonnement, l'admiration (Ça, par exemple!)

(7) Вот так! - la réaction à qch d'inattendu (En voilà une affaire! Quelle tuile!)

(8) Как бы не так! - le refus (Compte là-dessus!)

Des séquences de mots pareilles peuvent être potentiellement polyvalentes, mais grâce à l'intonation elles reçoivent l'unique interprétation possible dans une situation communicative donnée.

Ainsi, la formule un peu vulgaire (9) Надо же! peut selon l'intonation signifier l'étonnement, le reproche, la réprobation, l'indignation, l'admiration. [5]

3. Outre les clichés intonatifs il existe des idiomes intonatifs (au sens strict du mot). Ces constructions ont une structure syntaxique régulière et peuvent être prononcées avec une intonation neutre, non-emphatique. Mais en tant qu'idiomes, dits avec une intonation emphatique, ils reçoivent une autre signification, qui ne découle pas du sens propre des mots composants. L'intonation emphatique agit de sorte que le sens propre est remplacé par un sens figuré-appréciatif, relatif:

(10) Расскажи! - la demande, l'ordre (Parle! Raconte!)

(11) Расска-а-зывает! - la méfiance (A d'autres!)

(12) Не говори. - la détense de parler (Ne parle pas!)

(13) He говори! - la confirmation d'une opinion, d'une appréciation.

Dans ces cas-là c'est l'intonation qui différencie les homonymes: elle distingue les structures neutres régulières des locutions figées syntaxiques.

Le russe est particulièrement riche en formules expressives-négatives [3,4]. Comparez:

(14) Так я тебе и сказала.
(C'est ainsi que je t'ai dit.)

(15) Так я тебе и сказала!
(Je ne te le dirai pas!)

(16) Придет он. - Il va venir.

(17) Придет он! - Il ne viendra pas

On connaît bien des cas où l'intonation permet de donner au mot un sens diamétralement opposé:

(18) Молодец! - la louange (Bravo!)

(19) Молоде-еу! - la désapprobation, l'ironie (Tu as gagné!)

(20) Хором! - la louange (Ce qu'il est beau!)

(21) Хором! - le blâme, l'ironie (Un joli coco!)

En général, si le sens primaire et l'intonation se contredisent, c'est toujours l'intonation qui prend le dessus.

4. Il y a encore un groupe d'exemples intéressants dont le caractère spécifique est d'augmenter le sens ou, en d'autres mots, de combler une ellipse.

(22) Дел! - Il y a tant de choses à faire!

(23) Голос у нее! - Elle a une voix ravissante!

Une montée mélodique haute rend dans ces exemples l'idée d'une grande quantité en combinaison avec le génitif singulier ou pluriel du substantif et l'idée d'une haute appréciation - avec le nominatif. En cas de contact visuel entre les interlocuteurs le sens de l'intonation est souvent confirmé par des gestes (par exemple le hochement de tête) et une mimique appropriée.

Un autre cas où l'intonation comble une ellipse a lieu dans des répliques incomplètes, inachevées, qui sont typiques de la langue russe parlée:

(24) Ну ты вообще...

(25) Ну вечно ты... - En ces cas l'intonation de la continuation

remplace le prédicat manquant et ne laisse pas de doutes sur le sens de la réplique (généralement l'appréciation négative).

Malgré les distinctions considérables, il y a dans tous les cas qu'on vient d'examiner, un trait commun, notamment, le caractère stéréotypé de la séquence de mots et de l'intonation.

Les LPhI:

a) sont générées comme des unités toutes faites,

b) sont grammaticalement indivisibles,

c) ont un sens stable et intègre, qui ne peut pas être déduit des significations des éléments qui les composent,

d) étant polyvalentes sur le plan explicatif, elles sont monovalentes quant à l'aspect expressif et appellatif,

e) fonctionnent dans la langue comme des répliques-réactions,

f) sont caractérisées par une expressivité prononcée,

g) sont typiques de la langue parlée et de son imitation dans les oeuvres littéraires,

h) sont intimement liées à la mimique, à la pose, aux gestes,

i) souvent ne peuvent pas être traduites littéralement en d'autres langues.

3. CARACTÉRISTIQUES PHONÉTIQUES

Dans la phraséologie intonative on emploie des moyens phonétiques typiques de la langue emphatique:

1) allongement des voyelles - une des particularités les plus frappantes de l'intonation russe emphatique qui sert à rendre l'idée de l'étonnement, de l'enthousiasme, de la douceur, de la perplexité, de l'indignation, de la méfiance, du reproche etc.

2) allongement et renforcement emphatique des consonnes, compression des voyelles, typique de certaines émotions négatives (dépit, irritation, refus catégorique);

3) écarts de Fo augmentés à l'intérieur de la syllabe accentuée et entre les voyelles voisines;

4) utilisation des registres extrêmes: très bas et très haut;

5) tons complexes, modulés et entrecoupés par un coup de glotte;

6) ton plat, absence de mouvements mélodiques pendant des périodes bien longues;

7) divers timbres, types particuliers de phonation (voix aspirée, rauque, grinçante, tendue etc.)

3. PERCEPTION

Notre expérience linguistique témoigne d'un large emploi des LPhI dans la langue parlée. On les comprend dans une situation donnée d'une manière suffisamment univoque, surtout accompagnées de gestes et de mimiques. On peut supposer que les formules intonatives seraient bien reconnues même privées de leur contexte et du contact visuel des interlocuteurs.

Pour étudier le rôle de l'intonation dans la reconnaissance du sens des LPhI on a élaboré une méthode spéciale nommée "méthode des remarques d'auteur".

Les auditeurs doivent situer des répliques isolées, c'est-à-dire imaginer des situations où elles pourraient être employées et en caractériser l'intonation au moyen des remarques d'auteur du type: "a-t-il dit d'une voix calme, s'est-il écrié d'un ton surpris etc." A titre d'exemple on leur cite quelques fragments d'oeuvres littéraires avec des remarques d'auteur authentiques.

Le corpus expérimental est constitué par des oppositions du type:

(26) Он в город? Куда ему? - Il part pour la ville? Où donc?

(27) Он справится? - Куда-а ему! - Il va se débrouiller? - Il s'en faut de beaucoup!

(28) Хлеба я купил. Еще чего? - Du pain, j'en ai déjà acheté. Et quoi encore?

(29) Хлеба купишь? - Еще чего! - Tu achètes du pain? - Quoi encore!

Les fragments avec des éléments segmentaux identiques sont découpés du contexte. Les personnes qui suivent le cours de la théorie de l'intonation ont pris part aux expériences. Les répliques ont été présentées en paires (intonation neutre-intonation emphatique). Les auditeurs ont bien reconnu les différents types intonatifs et les ont décrit au moyen des remarques.

Les répliques des deux types sont décrites de différente manière. En caractérisant les répliques neutres les auditeurs se bornent généralement aux verbes déclaratifs communs et très fréquents (dire, demander, répondre). Quant aux LPhI, on y constate un choix beaucoup plus large des verbes et l'emploi plus fréquent des adverbes spécificateurs. Les auditeurs non seulement désignent une émotion, mais décrivent leurs impressions auditives (a-t-il dit d'une voix traînante, d'un ton tranchant). Il y a des auditeurs qui ne caractérisent pas l'intonation à l'aide des remarques, mais tâchent de rendre le sens de la réplique. Ainsi, la réplique "Сделает он!" (au sens "Il ne le fera pas") peut provoquer non seulement la réaction "a-t-il dit d'un ton sarcastique", mais aussi la réaction "Лентяй какой!" (Quel paresseux!); la réplique "Куда ему!" (au sens négatif) - la réaction "Он на такое не способен" (Il n'en est pas capable). On emploie aussi les descriptions des gestes et de la mimique. Ainsi, la réplique "А кто его знает!" (Qui sait?) a provoqué, entre autres, la réaction "a-t-il dit en haussant les épaules".

RÉFÉRENCES

- [1] BRYZGUNOVA E. A. (1977) *Zvuki i intonacija ruskoj retchi*. M.
- [2] FRAZEOLOGITCHESKIJ SLOVAR RUSKOGO JAZYKA (1967) Pod red. A. Molotkova. M.
- [3] KIPRIJANOV V. (1975) *Frazeologizmy-kommunikativy v sovremenom russkom jazyke*. Vladimir
- [4] MUKHANOV I. L. (1986) *Funkcionirovanije tchastic i intonacii v ekspressivno-otricatelnykh predlozhenijakh so slovom "kakoj"*. - "Ruštinar". Bratislava, N 8, p. 4-8
- [5] RATMAYR R. (1988) *Opredeľenje kommunikativnogo znatenija replik-reakcij s nepolnoznačnymi slovami v razgovornoj retchi*. // Problemi di Morfosintassi delle lingue slave. Univ. di Bologna
- [6] ZEMSKAJA E. (1979) *Russkaja razgovornaja retch: Lingvističeskij analiz i problemy obutčenijsa*. M.

INTERFERENCES PHONETIQUES AU COURS DE L'APPRENTISSAGE
DU FRANCAIS (GROUPE LINGUISTIQUE SLAVE)

N.Evtchik, G.Roudzit

Institut des langues étrangères, Minsk,
Union Soviétique

The paper deals with the study of phonemic and prosodic interference in Byelorussian-French "class-room" bilingualism. The study is based on the theoretical contrastive analysis of French and Byelorussian phonetic systems. The experimental research of French utterances produced by Byelorussian learners made it possible to reveal typical deviations from the norms of French pronunciation.

Il est de notoriété générale que la perception et la production des phonèmes ainsi que des structures accentuelles et intonatives d'une langue étrangère sont conditionnées par le système phonétique de la langue maternelle du sujet parlant. Les étudiants étrangers perçoivent et interprètent la prononciation de la langue qu'ils étudient à travers "le cryble" phonologique de leur propre langue (6). Cela entraîne de multiples fautes de prononciation et de nombreuses déviations de la norme admise dans la langue étudiée.

Dans le domaine de la prononciation il existe deux types d'interférences: interférences phonologiques et

interférences phonétiques. Les interférences phonologiques liées à la différenciation du sens des unités linguistiques ont une occurrence plus faible que les interférences phonétiques qui créent "cet accent" particulier étranger à la langue étudiée.

Les interférences se produisent au niveau de la prononciation des sons et des faits prosodiques. Ce sont les structures prosodiques (rythme et intonation) qui sont les plus difficiles à assimiler (4). Les fautes au niveau prosodique sont particulièrement lourdes de conséquences car elles empêchent d'acquiescer de bonnes habitudes aux étudiants étrangers.

C'est à l'aide de l'utilisation de la méthode d'analyse contrastive des systèmes phonétiques de ses deux langues, que nous sommes parvenues à établir la zone des interférences. Notre but était d'étudier les interférences phonétiques dans le parler des étudiants biélorusses qui apprennent le français. Le corpus expérimental contenait des phrases énonciatives, interrogatives (questions totales et partielles) et des phrases impératives, réalisées par deux sujets français (phrases modèles -Ph.-M.) et quatre sujets biélorusses apprenant

le français depuis trois ans à la faculté de français (variantes - Ph.-V.). Les résultats des analyses aux niveaux acoustique et perceptif ont permis de dégager un certain nombre de caractéristiques spécifiques dues aux interférences. Parmi ces traits il y a ceux qui se rapportent à tous les types communicatifs et ceux qui sont conditionnés par les particularités propres au type communicatif déterminé. Ainsi, toute phrase française se caractérise par l'isochronisme des syllabes inaccentuées relativement brèves, tandis que les syllabes accentuées sont marquées avant tout par l'augmentation de la durée. Cette durée peut être deux fois plus grande que celle des syllabes inaccentuées (2). Dans le langage des étudiants biélorusses cette tendance rythmique n'est pas toujours observée. D'une part, l'influence de la langue maternelle se révèle dans la perturbation du principe de l'isochronisme des syllabes inaccentuées. Il y a même des cas où la durée des syllabes inaccentuées dépasse celle des syllabes accentuées. Une des fautes les plus typiques observées dans la prononciation des étudiants revient au contraste trop marqué des syllabes accentuées et inaccentuées. Cela a pour résultat l'hyperprééminence des syllabes toniques par rapport aux syllabes atones. Cette faute chez les sujets biélorusses s'explique par l'absence dans le biélorusse de tension musculaire constante qui doit être maintenue en permanence sur toutes les syllabes au cours de l'émission de la phrase

française (2). Très souvent les étudiants biélorusses n'arrivent pas à répartir cet effort musculaire et nerveux en réservant un léger supplément d'énergie pour la dernière syllabe.

Puisque la succession des syllabes n'est pas prononcée avec une force sensiblement égale, nécessaire en français (5) nous voyons apparaître une déviation très répandue chez les biélorusses étudiant le français qui est la réduction des voyelles atones: p. ex. professeur, manifestation.

Les sujets biélorusses se trahissent par l'absence du savoir-faire de déplacer l'accent du mot essentiel au mot auxiliaire et inversement. Dans de nombreux cas nous avons constaté la prééminence de deux syllabes en contact, due à l'accentuation des mots non-accentogènes (adjectifs possessifs, prépositions, verbes auxiliaires).

Ph.-M.: Il vous attend dans le vestibule.

Ph.-V.: Il vous attend dans le vestibule.

Dans ses articles P. Delattre (1,3) a souligné à plusieurs reprises que l'accentuation du mot français dans la phrase varie dans les limites: accentuation-désaccentuation-inaccentuation. C'est-à-dire les accents se répartissent au minimum en trois degrés différents. Les diglottes biélorusses ne réussissent pas à prononcer ces éléments rythmiques de la phrase tenant compte de l'hierarchie des accents.

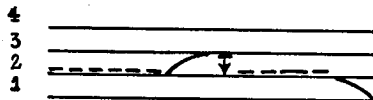
Cette interférence propre à la prononciation biélorusse ne peut entraîner confusion de sens, car l'accent tonique n'a pas de fonction distinctive en français. Mais la simplification des rapports hiérarchiques entre les syllabes accentuées amène à un

certain changement de rapports sémantiques, en particulier de rapports modaux. Ph.-M.: Pensez à lui rendre le dossier de l'effectif du personnel de l'entreprise. Ph.-V.: Pensez à lui rendre le dossier de l'effectif du personnel de l'entreprise. Tout ce que nous avons discuté jusqu'ici concerne essentiellement l'accent. Il est à noter que ces interférences accentuelles se combinent avec celles de la mélodie et englobent le niveau tonal, les écarts mélodiques, les registres. Il est connu que la phrase française énonciative commence au niveau moyen. Lorsque qu'on parle sans émotion, sans affection l'attaque de la phrase est douce (4). Chez les sujets biélorusses l'attaque de la phrase est très souvent assez forte et le niveau du début de la phrase est plus élevé. En ce qui concerne le niveau final de la phrase informative il faut dire qu'il n'atteint pas toujours la finalité complète. La phrase énonciative est alors perçue comme inachevée, en suspens. Dans la phrase interrogative (question partielle) nous constatons la réalisation du ton toujours ascendant qui est très marqué dans la partie finale. Au niveau de la perception cette phrase est comprise comme trop émotionnelle ce qui ne correspond pas toujours à la situation. Ph.-M.: Comment la maladie de Claire s'est-elle annoncée? Ph.-V.: Comment la maladie de Claire s'est-elle annoncée? Nous venons de décrire les fautes les plus fréquentes dans les parties initiale et finale de la phrase, mais il arrive souvent qu'au milieu de la phrase les sujets biélorusses baissent le ton à la fin des groupes de sens au lieu de prononcer la der-

nière syllabe sur un ton montant. Ces interférences s'expliquent partiellement par la présence dans le biélorusse de syllabes post-toniques qui doivent être réalisées sur un ton ascendant.

Ce qui est indispensable de souligner c'est que les étudiants biélorusses n'arrivent pas à réaliser les écarts mélodiques nécessaires à la jonction des unités rythmiques (4), c'est-à-dire la première syllabe de l'unité qui suit est prononcée sur le même niveau que la syllabe finale de l'unité précédente.

Ex.: Vous avez un seul changement à la station suivante:



Outre les déviations relevées au niveau prosodique il y a un certain nombre d'interférences au niveau de la prononciation des sons. Parmi les cas les plus répandus on peut mentionner les suivants:

- assourdissement des consonnes finales: cage-[kaʃ], rose -[ros];
- palatalisation des consonnes t, d devant les voyelles i, u: tu dis;
- nasalisation des voyelles orales devant les consonnes nasales: Seine - [sɛ̃];
- la non-différentiation des oppositions distinctives des voyelles telles que [oe - e]: [jə di - je di], [ø - o]: [radiø - radio], [a - o]: [blā - blō], [ɔ - o]: [pɔm - pom],

Les déviations typiques mentionnées ci-dessus sont prises en considération par les enseignants lors de la création des exercices portant sur les voyelles, les consonnes et l'intonation.

(1) DELATTRE, P. (1966), "Accent de mot et accent de groupe", *Studies in French and comparative Phonetics*, The Hague: Mouton, 69-72.

(2) DELATTRE, P. (1966), "L'accent final en français: accent d'intensité, accent de hauteur, accent de durée", *Studies in French and comparative Phonetics*, The Hague: Mouton, 65-68.

(3) DELATTRE, P. (1966), "Le mot est-il une entité phonétique en français?" *Studies in French and comparative Phonetics*, The Hague: Mouton, 73-76.

(4) FAURE, G. (1962), "Recherches sur les caractères et le rôle des éléments musicaux dans la prononciation anglaise", Paris: Didier, 370.

(5) FAURE, G. (1971), "Accent, rythme et intonation", *La grammaire du français parlée*, Paris: Hachette, 27-37.

(6) TROUBETSKOÏ, N. (1960), *Osnovi fonologii*, Moskva: Inostranaja literatura, 372.

Wim A. van Dommelen

Institut für Phonetik und digitale Sprachverarbeitung
Kiel, Germany

ABSTRACT

This paper investigates the influence of fundamental frequency (F0) contour on the perception of segment duration in isolated German words. It is shown that the established opinion concerning an increase of perceived segment duration due to a dynamic vs. a flat F0 contour must be modified. The effect appears to be dependent on word structure: While for monosyllables the lengthening effect of dynamic F0 was confirmed, a shortening effect was observed for disyllabic words.

1. INTRODUCTION

Since the experimental work of Lehiste [2] there seems to be a common opinion concerning the influence of the F0 contour in a vocalic segment upon the segment's perceived duration. A dynamic as against a flat F0 contour is generally regarded to lengthen subjective vowel duration. Other investigations [1, 3] showed that the perception of a word-final obstruent as fortis vs. lenis, too, is biased by the character of the F0 contour in the preceding vowel. Here, the increase of lenis judgements due to a moving F0 contour is explained by the subjectively longer vowel duration, biasing in turn phoneme perception towards more lenis.

In previous - hitherto unpublished - experiments I investigated the influence of a varying F0 contour on the perception of vowel quantity in German disyllabic words.

Besides the main cue of vowel duration, phoneme perception was found to be influenced by the F0 pattern of the first syllable's vowel. The direction of the effect, however, was the opposite from what was expected, a moving F0 contour causing a subjective *shortening* of the vowel. Looking for a possible explanation it was noticed that all previous experiments on this issue used either isolated vowels [2] or monosyllables [1, 3]. Two experiments were therefore set up to investigate word structure (monosyllabic vs. disyllabic) effects.

To investigate the influence of segment duration and F0 contour, the German vowel pair /a:/ - /a/ was used, since this is the only vowel quantity opposition which is mainly cued by duration. All the other long/short oppositions are associated with large differences in vowel openness. The monosyllables were represented by the word pair "Aas" (/a:s/, "carrion") - "As" (/as/, "ace"). As a maximally similar word pair containing two syllables "aBen" (/a:sən/, "[we] ate") vs. "Assen" (/asən/, "aces", dative case plural) was chosen. Variation of actual vowel duration should induce listeners to identify the vowel as being phonologically long vs. short. In addition, the vowel's F0 contour was modified (flat vs. flat-falling). According to the findings reported in the literature up to now, a dynamic F0 contour should lengthen subjective vowel duration and

thereby shift the phoneme boundary towards shorter durations. This was taken as a working hypothesis for both monosyllabic and disyllabic words.

Also, to aid interpretation, the duration of the postvocalic fricative was varied. It was expected that the listeners would interpret a shorter fricative as a momentary faster speech rate, implying a subjective lengthening of the preceding vowel. This should apply to mono- as well as disyllabic words.

2. EXPERIMENTAL PROCEDURE

Test 1

A token of the German word "Aas", spoken on a monotone by a trained male speaker, was used for stimulus generation. Prior to electronic splicing, the word was low-pass filtered at 5 kHz and digitized at a sampling rate of 10 kHz. The subsequent manipulations involved the F0 contour and the temporal structure by manipulation of the synthesis frame rate. First, following an LPC analysis two different F0 contours were created. The first one was "flat" (slightly falling from 115 - 110 Hz to avoid an unnatural vowel quality), the second one was initially flat and fell linearly during the second half of the vowel (115 - 113 - 75 Hz). Subsequently, using these two F0 contours the vowel (original duration 285 ms) was synthesized with eight different durations (varying from 110 - 250 ms in 20-ms steps). The fricative was synthesized with its original duration (373 ms) and shortened by approximately one third (250 ms). Each of these two fricatives was spliced with each of the eight tokens of the vowel duration continuum. In total, this test comprised 32 stimuli (8 vowel durations x 2 fricative durations x 2 F0 contours).

Test 2

For the second test, the same speaker from Test 1 produced the

word "aBen" on a monotone. The manipulations of the test word closely followed those from the first test. Two vowel F0 contours were generated, which were identical with those described above. The /ən/ part received a low F0 contour, which fell from 70 to 60 Hz. Its amplitude was attenuated by ca. 13 dB to be auditorily coherent with the falling F0 contour. Furthermore, the duration of the /a:/ vowel (originally 211 ms) was varied from 85 - 190 ms in eight 15-ms steps. As for the monosyllables, the fricative was synthesized with its original duration (211 ms) and a shortened one (141 ms).

For both tests, the stimuli were replicated five times, recorded on analogue tape in a randomized order, and presented to the listeners via a high-quality loudspeaker. The listeners were seated in a sound-treated room and responded by marking one of two alternatives ("Aas"/"As" and "aBen"/"Assen", resp.) on a prepared answer sheet. Twenty-four phonetically naive subjects took part in Test 1 and twenty-two in Test 2. Eight subjects participated in both tests.

3. RESULTS

The results from the first test are presented in Figure 1. It can be seen that the categorization of the vowel as phonologically short vs. long is mainly cued by its physical duration. Second, the effect of F0 contour confirms the finding of Lehiste [2]: A dynamic contour leads to an increase of "long vowel" judgements. Note that this effect holds for both fricative durations. Third, following the expectations, a shorter fricative duration causes an increase of subjective vowel duration.

The results for Test 2 partly parallel those for Test 1. As far as the influence of temporal structure on vowel quantity per-

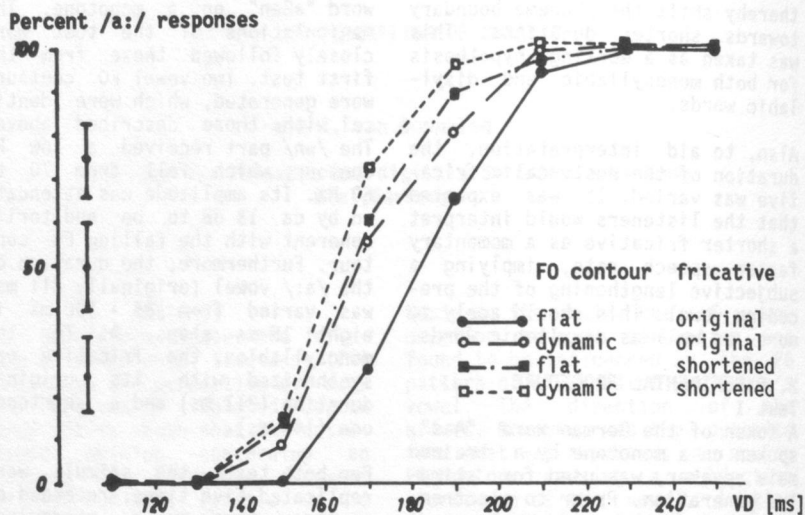


Figure 1
Percentage of /a:/ responses for the monosyllables as a function of vowel duration (VD) for flat and dynamic F0 contours in the vowel and two fricative durations. At each data point $n = 120$. Vertical bars at the left indicate 95% confidence intervals at 25, 50, and 75%.

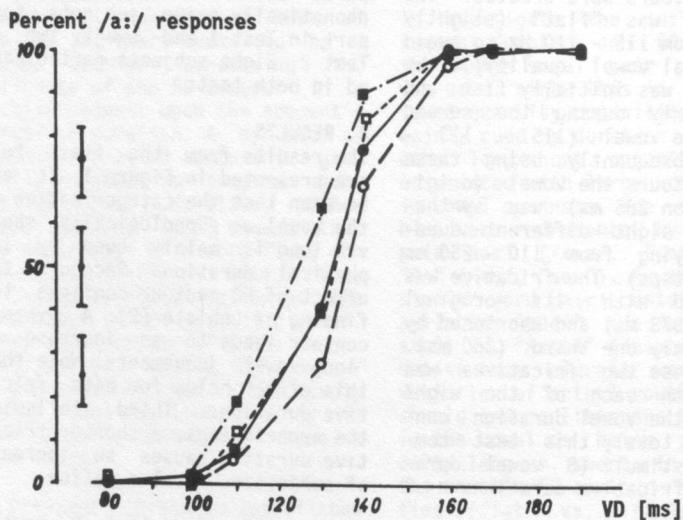


Figure 2
Same as Figure 1, but for the disyllabic words. $n = 110$.

ception is concerned, the results are fully comparable (cf. Figure 2): Whilst phoneme identification primarily depends on actual vowel duration, shortening of the postvocalic fricative biases the location of the phoneme boundary towards shorter durations. In this case, too, perception is additionally influenced by the F0 contour in the vowel. The direction of the effect is, however, exactly the opposite. Here, the presence of a dynamic vs. flat contour causes a decrease, rather than an increase of "aßen" judgements. This occurs consistently for both fricative durations.

4. DISCUSSION

With the results of the test on monosyllables the findings from the literature have been replicated: There was an increase of "long vowel" judgements due to a dynamic vs. flat F0 contour, which certainly reflects the impression of a longer vowel duration in that case. A possible explanation may be the assumption of a production-perception link for voiced segments in absolute final position. This is supported by data from Lyberg [4], who found a positive correlation between segment duration and extension of the F0 fall for a Swedish [dag] syllable in sentence-final position.

The results of the second experiment have clearly falsified the second part of the working hypothesis: The vowel in the disyllabic words was auditorily shortened by a dynamic F0 contour, and not lengthened. Since the experimental conditions (speaker, processing, and test conditions) were identical for both tests, it is highly improbable that the effects are artifactual or due to different perceptual behaviour of the listeners. This is supported by the observation that the effect of varying the fricative duration is identical in both tests.

Further investigations are required to find an adequate explanation for the different interactions of F0 and perceived segment duration in mono- vs. disyllabic words. At the moment, it might be speculated that the crucial feature of the disyllabic word in this respect is the global intonation contour. F0 information in the context surrounding the syllable in question could provide a reference for the listeners. Isolated monosyllables, in contrast, lack this reference. This speculation is supported by the results of preliminary experiments using monosyllables in carrier phrases. It seems therefore that the experiments reported before focused on an exception, rather than the rule.

5. REFERENCES

- [1] GRUENENFELDER, T.M.; PISONI, D.B. (1980), "Fundamental frequency as a cue to postvocalic consonantal voicing: some data from speech perception and production", *Perception and Psychophysics*, 28, 514-520.
- [2] LEHISTE, I. (1976), "Influence of fundamental frequency pattern on the perception of duration", *J. of Phonetics*, 4, 113-117.
- [3] LEHISTE, I.; SHOCKEY, L. (1980), "Labeling, discrimination and repetition of stimuli with level and changing fundamental frequency", *J. of Phonetics*, 8, 469-474.
- [4] LYBERG, B. (1981), "Some consequences of a model for segment duration based on F0-dependence", *J. of Phonetics*, 9, 97-103.

DURATIONAL SHORTENING AND ANAPHORIC REFERENCE

W. N. Campbell

ATR Interpreting Telephony Research Laboratories, Kyoto, Japan.

ABSTRACT

A tendency for the durational shortening over time of nominal heads and pronominal references to them was confirmed in a twenty-minute radio broadcast of a short story. Lengthening was calculated by comparing durations predicted by the timing algorithm of a computer text-to-speech program with those measured from a recording of the story after factoring out global changes in rate of speech. Considerable variation was noted in the residuals thus formed, and resetting was found to correlate with events in the narrative of the text.

1. INTRODUCTION

Fowler & Housoum [4] have shown that speakers distinguish words that are *new* to a monologue from those that are assumed *given*, by shortening subsequent occurrences of the latter. More recently, for Dutch, Eefting [3] constructed experiments to compare the effects of *information value* and *accentedness* on the duration of words but found only a weak effect for the former, compared to that of the latter, and questioned the value of *given-ness* as a predictor of duration in itself.

This paper looks at the durational correlates of anaphoric reference in a long passage of naturally occurring professionally narrated text and shows that the durations of both antecedent nominal units and subsequent pronominal references to them can be seen to reduce as the utterance progresses, but with resetting occurring after an interval of time and at major breaks in the narrative of the text.

2. MATERIALS

A twenty-minute radio broadcast of a short story was analysed for anaphoric reference, enumerating all nominal units and tagging pronominal references to each with an identifying code. These codes were linked using a text editor to a list of the syllables with measures of the lengthening undergone by each. In this way every occurrence of a nominal unit or any of its referring pronouns can be associated with a value representing the degree of reduction in the duration of its syllables.

2.1. Measuring length

The passage was digitised and measured for duration at the syllable level. The timing component of a computer text-to-speech system was optimised for the passage and used to predict the durations from input that had been manually coded to describe each syllable in terms known to control a large portion of the durational variance. The output from the program was compared with the original durations, and a set of residuals produced ($\text{residual} (\%) = (\text{predicted duration}/\text{observed duration} \times 100) - 100$) which can be taken to represent the amount of over- or under-prediction, which in turn can offer clues to the identity of significant factors not being sufficiently taken into account by the prediction algorithm [2].

The strongest of these factors is presumed to be variation in speech rate, which is clearly evident on perceptual evaluation and can be visualised in a graphical plot of the residuals as a slowly

changing, low frequency offset. The *supsmu* smoothing function of the Splus statistical package [1] was used to model this offset and the smoothed representation then subtracted from the residuals to factor out, in a simple way, that part of the variance that can be assumed due to changes in speaking rate. The new residuals thus obtained indicate whether the speaker was rendering each syllable faster or more slowly than the local norm. If the program predicts a duration greater than that observed, then the residual will be positive, reflecting a presumed reduction in the duration of the spoken syllable, and vice versa.

Such data are by no means perfect, and results for individual samples are subject to considerable uncertainty due to measurement errors etc., but if trends can be observed from large numbers of observations then appropriate rules to describe the trends can be formulated and incorporated into the model to improve the quality of future predictions. Some loss of certainty is inevitable when working with large numbers of samples, and the method lacks the controls that experiments with laboratory-produced sentences may have, but this is felt to be a small price to pay for insights into the more delicate timing processes of naturally-occurring speech data.

3. RESULTS

Seven nominal units were repeated more than ten times in the passage:

nominal:	freq:	references:
Gerry	18	he 174, him 44, you 12, I 4
the tunnel	14	it 4, there 1
the rock	14	it 5
the water	14	- -
his mother	12	she 29, her 17
his head	12	it 1
the boys	11	they 16, them 11

No unit was monosyllabic, and some references contained adjectives or were

homonyms, as in *a well of blue sea, the stinging salt water, the blue well of water, etc.* for *the water*. In such cases the mean value of lengthening for all syllables in the group was taken to represent the lengthening of the unit.

The duration prediction algorithm accounted for 86% of the variance, with a correlation of $r = 0.93$. After deduction of the smoothed fit, the standard deviation of the residuals was reduced from 20ms to 7ms, and the range reduced from 138ms to 58ms. Values of lengthening were in the range of $\pm 20\%$.

In very few cases was there a simple linear reduction in duration over time. The overall trend was towards a reduction in syllable duration, but there was considerable oscillation about the reducing mean. With the exception of the hero's name, *Gerry*, for which the slope was -0.12 , robust regression lines fitted to the samples showed a positive slope, the steepness of which ($0 = \text{flat} = \text{no change}$) indicates the rate at which later syllables are reduced in duration.

If we look, for example, at references to *the rock* (Figure 1), a slope of 1.29 fits in the overall case, but there is a major change of direction in the narrative after ref. 4, and a new sequence can be considered to begin from ref. 5. The resulting slopes for the two groups thus formed would become 7.8 and 3.6 respectively.

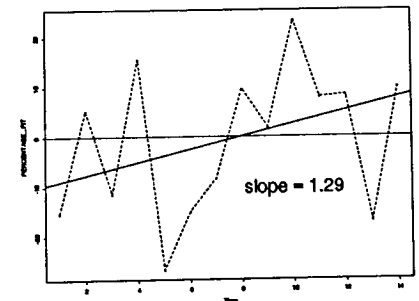


Figure 1: Residuals for *rock*.

For occurrences of *the water* (Figure 2), the overall slope was positive at 0.035,

but not significant. Again there is a major break in the theme of the text after ref. 7, and a line fitted to the first 7 occurrences shows a steeper slope of 6.14, from the percentage values -16, -12, -1, 7, 5, 18, and 20. The next three occur closely in time at 6m 40s, 6m 50s, and 7m 40s and if a reset is assumed to have occurred between these and ref. 11, occurring much later at 10m 25s, corrected slopes for the next two groups thus distinguished become 10.83 and 1.41.

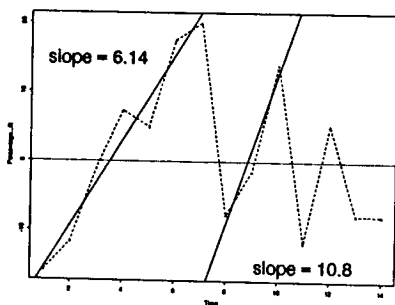


Figure 2: Residuals for *water*.

Time alone is not a sufficient cue to resetting; references to *the boys* show a similar pattern, and resetting after the first four would change an overall slope of -1.9 into a slope of 9.2 for the first part. However, although refs. 4 & 5 are very close together (at 5m 55s and 6m 10s), during these 15 seconds of narrative, Gerry goes swimming underwater, exploring, comes back to the shore, and then sees the boys once again in the same place as he first saw them. Not only is the location reset, but also apparently the syllable timing.

Pronominal reference too, shows similar reduction in duration with time. A robust regression line fitted through the 12 references to 'Gerry' as *you* (Figure 3) had a slope of 1.76 to fit the percentage values (-12, 22, 10, 7, 34, 24, 10, 22, 6, 24, 31, 24) measured for each occurrence.

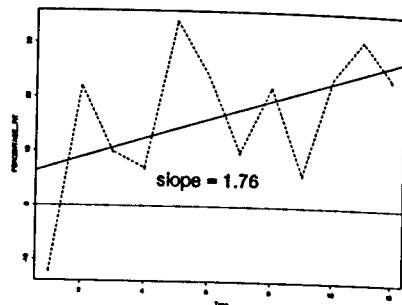


Figure 3: Residuals for *you*.

If the first six tokens, which occur together in the first 1m 50s of the story are considered separately from the next, which does not come until 11m 29s, steepness of the two groups becomes 6.2 and 3.2 respectively.

References to Gerry as *him* fit a slope of 0.6 (Figure 4). It seems that the first few group separately, and the last, too, may be an exception.

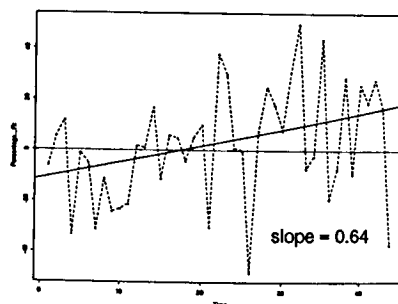


Figure 4: Residuals for *him*.

References to his mother as *her* (Figure 5) show less variation about the regression line, which has a slope of 2.79. The first three points here, too, appear exceptional; no explanation has yet been found for why these tokens should be spoken faster than the subsequent ones, but it is possible that the smoothing algorithm used to factor out global changes in speaking rate coped less well with the edge samples. However, such local details are beyond the necessarily

general scope of this paper and would require more sophisticated statistical and linguistic analyses.

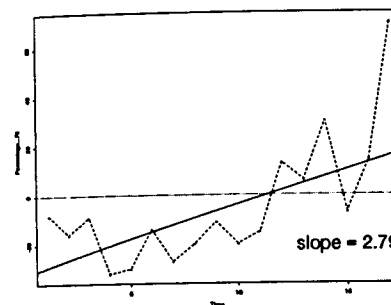


Figure 5: Residuals for *her*.

5. DISCUSSION

It would be unwise to draw strong conclusions from such a small study, but a clear trend has been shown, with regression lines drawn through selected groups of the data having a positive slope of around 5, indicating a 5% difference in the fit of each subsequent occurrence after factoring out the most obvious effects on duration by a predictive model. This notional line describes the best fit to the distribution of the points, but there is considerable scatter around the line and it would be difficult to predict the degree of fit for any one token from it. This difficulty is compounded by the difficulty in knowing where to allocate a reset. Clear separation in time appears to be one factor, and in retrospect, with the aid of the difference measures, marked changes in the flow of the narrative can be found that correlate well with resets in the cycle, but work has yet to be done to determine whether such resetting points could be determined *a priori*, from knowledge of the text alone.

On the other hand, the data do show that such work would be of use to the prediction of timing in speech, and confirm that there is a tendency to shorten the duration of items in a narrative when those items can be assumed to be

known, or shared information. The samples were taken from a long passage of naturally occurring professionally narrated text and the measuring of their length made use of a differencing between observed and predicted durations using the timing algorithm developed for a computer text-to-speech system. This one-step-removed form of measurement introduces a degree of uncertainty of its own, but allows a finer view of the effects on speech timing by factoring out the grosser, more predictable components.

Acknowledgements

I am particularly grateful to Steve Fligelstone of Lancaster University for providing the Anaphora tagging, and to ATR for enabling the continuation of this research.

REFERENCES

- [1] BECKER, R. A., CHAMBERS, J. M. & WILKS, R. A. (1988) "The New S Language: A Programming Environment for Data Analysis and Graphics", AT&T Bell Laboratories, Wadsworth & Brooks/Cole Advanced books & Software, Pacific Grove California.
- [2] CAMPBELL, W. N. (1990) "Measuring Speech-Rate in the Spoken English Corpus", pp 61 - 81 in *Theory and Practice in Corpus Linguistics*, Eds J. AArts & W. Meijs, Rodopi, Amsterdam.
- [3] EEFING, W. (1991) "The effect of 'information value' and 'accentuation' on the duration of Dutch words, syllables, and segments". *JASA* # 89, pp 412 - 424.
- [4] FOWLER, C. A. & HOUSOUM, J. (1987) "Talkers' signalling of 'new' and 'old' words in speech and listeners' perception and use of the distinction". *J. Mem. Lang.* #26, pp 489 - 504.

THEORIES OF PROSODIC STRUCTURE: EVIDENCE FROM SYLLABLE DURATION

†D. R. Ladd & ‡W. N. Campbell

†Department of Linguistics and Centre for Speech Technology Research,
Edinburgh University, Edinburgh, Scotland.

‡ATR Interpreting Telephony Research Laboratories, Kyoto, Japan,
and CSTR, Edinburgh University, Edinburgh, Scotland.

ABSTRACT

A recent theoretical proposal to enrich the traditional fixed hierarchy of prosodic domain types (foot, phrase, etc.) by allowing the possibility of "compound phrases", has been tested with a model of syllable duration for English. By marking the input text to identify both subordinate and superordinate major and minor tone-group boundaries, a finer specification of the durations of phrase-final syllables can be achieved. The new description explains significantly more of the error in the predictions for these syllables in the duration model.

1. INTRODUCTION

Rules for segment and syllable duration remain one of the least satisfactory aspects of most speech synthesis-by-rule systems. Empirical studies [3] have established many of the factors that affect duration, including both segmental differences (manner and place of articulation, vowel height, etc.) and prosodic factors such as degree of stress and position in phrase. However, current models still fall well short of accurately reproducing the timing of natural speech.

There is reason to believe that part of the difficulty in modelling duration stems from theoretical shortcomings in the identification of the prosodic factors involved. A number of current issues in phonological theory concern the nature of prosodic structure and the relationship among different prosodic features. Obviously, if the definition of e.g.

'phrase' is open to debate, then this will affect the way 'phrase boundaries' are marked in any given corpus or text sample, which in turn will affect any empirical findings about the influence of phrase boundaries on syllable and segment duration.

The study reported here is an attempt to assess whether such theoretical issues are of any practical significance for empirical models, and more specifically to evaluate Ladd's theoretical claim (Ladd [4][5]) that there is no principled limit to the depth of prosodic structure. We do this by comparing the durational effects of phrase boundaries *within* and *between* what Ladd calls 'compound' phrases, i.e. phrases that are themselves composed of two or more phrases. We report two kinds of results: first, whether there are significant differences on this comparison, and second, whether inclusion of the distinction makes possible a significant improvement in the amount of variance accounted for. If the answer to both questions is yes, it will illustrate the potential relevance of these issues in phonological theory for practical applications in phonetics and speech technology.

2. MATERIALS & PROCEDURES

2.1 Modelling syllable duration

Our starting point is the model of syllable duration reported by Campbell [1], and its application to a sample text. 3959 syllable durations were measured from a twenty-minute passage of speech

recorded (with permission) from a BBC Radio broadcast of a short-story. The passage was prosodically annotated by two British-trained phoneticians to indicate stress, accent-type, and both major (maj-tg) and minor (min-tg) tone-group boundary locations¹. Each syllable was then tagged with a number of identifiers (e.g. stressed syllable in one-syllable foot, final syllable in maj-tg, etc.) and a neural network was trained to predict the durations from the annotated input.

In this study we concentrate on the model's predictions of the effects of position in the phrase. Five categories of syllable are defined with regard to two types of prosodic phrase: initial in maj-tg (1), initial in min-tg (2), medial (3), final in min-tg (4), and final in maj-tg (5).

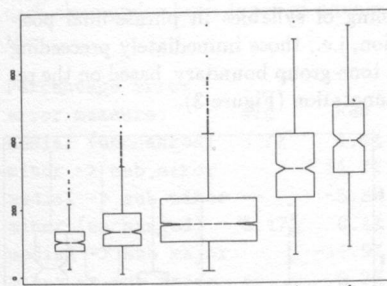


Figure 1: Durations of syllables factored by position in phrase.

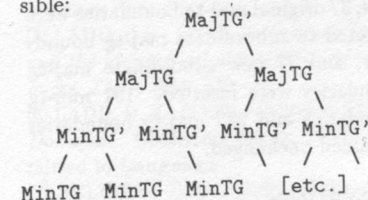
The boxes in Figure 1 are drawn with horizontal lines indicating the 25th, 50th and 75th percentiles of the durations of these syllables in milliseconds; the notches indicate significance at the 5% level in the difference of the distributions if they show no overlap. Analysis of variance of the durations factored in this way yields $F_{4,3954} = 529.4$ ($p < 0.001$), showing that the effect of position is highly significant.

¹Maj-tg corresponds roughly to Pierrehumbert and Beckman's 'intonational phrase' and min-tg to their 'intermediate phrase'.

However, the model's predictions, as suggested in the introduction, are only as good as the transcription on which they are based. The transcription is a traditional 'British school' analysis in which utterances are composed of major tone groups, and major tone groups are composed of minor tone groups. This categorisation of prosodic phrases conforms to the 'Strict Layer Hypothesis' (Selkirk [8]), according to which the prosodic structure of any utterance consists of a hierarchical arrangement of a fixed number of prosodic domain types. The Strict Layer Hypothesis is what is challenged in Ladd's work: specifically, Ladd has argued for the existence of 'superdomains' or 'compound prosodic domains', in which two or more adjacent domains of a given type are gathered together in a larger prosodic constituent *which is itself of that type*. Evidence for this proposal includes studies of acoustic cues to 'boundary strength' (e.g. Cooper and Paccia-Cooper [2], Ladd [6]), in which considerable depth of structure is reflected in segmental duration and F0 properties in the vicinity of boundaries.

2.2. Refining the model with an enriched prosodic structure

With Ladd's proposal in mind, one of us (DRL) retranscribed the phrase boundaries in the corpus to allow for both compound maj-tgs and compound min-tgs. That is, we assumed that at least the following depth of structure is possible:



We thus have four hierarchically arranged types of tone group boundary rather than, as in the original traditional transcription, two.

It is important to note that there is no regular mapping from the traditional transcription onto the new one: the new one is based on a richer categorisation of the data. For example, many boundaries that were marked as min-tg boundaries in the original transcription became subordinate maj-tg boundaries in the new transcription, but so also did many of the original maj-tg boundaries. On the other hand, other boundaries marked as min-tg in the original transcription were 'demoted' rather than 'promoted', becoming subordinate min-tg boundaries in the new transcription. In addition, there were many places at which no boundary was marked in the original transcription but where a subordinate min-tg boundary was marked in the retranscription.

Space does not permit any discussion of the kinds of phonetic cues that motivated the choice of boundary type in the retranscription; to some extent, as with the original transcription, choices were made on partly intuitive or impressionistic grounds. However, the point is not to argue in detail for one impressionistic transcription over another, but rather to show that, given a transcription that permits richer distinctions of boundary type, we can make more accurate predictions of syllable duration.

In the re-annotation, 174 new min-tg boundaries were inserted, 33 min-tg boundaries were demoted to subordinate, 190 min-tg boundaries were promoted to subordinate maj-tg boundaries, 37 original maj-tg boundaries were demoted to subordinate maj-tg boundaries, and 27 new subordinate maj-tg boundaries were inserted. 191 min-tg boundaries and 233 maj-tg boundaries remained unchanged.

3. RESULTS

Examination of the durations of the new tone-group-initial syllables (Figure 2) showed no significant difference be-

tween the sub-minor (4), minor (3), and sub-major (2) classes, although all three were significantly shorter than those in medial position (5), and major-initial syllables (1) were significantly shorter than any other group ($F_{4,3954} = 40.8$).

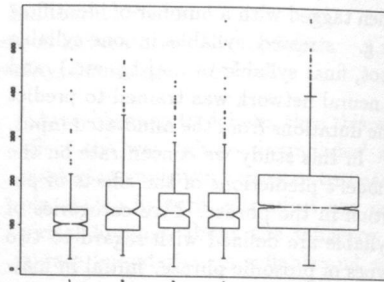


Figure 2: Durations of initial syllables.

Better separation is found in the lengthening of syllables in phrase-final position, i.e., those immediately preceding a tone-group boundary, based on the re-annotation (Figure 3).

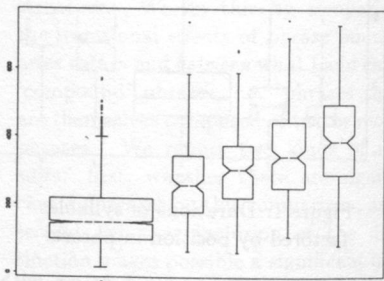


Figure 3: Durations of final syllables.

Clear and significant differences that correlate well with boundary strength can be found between those in subordinate min-tgs (1), superordinate min-tgs (2), subordinate maj-tgs (3), and superordinate maj-tgs (4). All are significantly longer than those in medial (0) position ($F_{4,3954} = 610.5$) and it can be concluded that the adoption of the finer classification provides better discrimination of the syllable durations.

3.1. Improving the prediction

In order to quantify the improvement that can be expected from incorporation of tone-group subordination in the synthesis model we can examine the residuals from a prediction. The best current prediction accounts for 86% of the variance in the durations, and by computing $\text{predicted duration}/\text{observed duration} \times 100$, we have a percentage measure of the degree of fit for each syllable. Many factors contribute to the prediction error, and much of it may be randomly distributed. If a significant portion can be associated with any one factor, however, then retraining of the model with improved factorisation should account for that part of it.

The following table shows the percentage error that can be attributed to each class under both types of annotation.

Percentage error measure:	old	new
medial (unchanged)	3.7%	4.4%
minor → sub minor	--	14.7%
medial → sub minor	--	-5.3%
minor (unchanged)	5.17%	6.1%
medial → sub major	--	-14.9%
minor → sub major	--	2.7%
major (unchanged)	-3.4%	-2.4%
major → sub major	--	-9.5%

For example, there was a 3.7% misprediction distributed among the syllables that had originally been classified as medial; those that are now classed as subordinate-major-tone-group-final syllables form a small subgroup of that set which account for 14.9% of the error. By focussing the mispredictions in this way and retraining the network with data tagged according to the compound model, an improvement of up to 14% can be expected in the durations of that subgroup of syllables, which should significantly reduce the error in the group of medial syllables as a whole.

4. CONCLUSION

In this study we have compared two approaches to modelling position-in-phrase effects on syllable duration. The model previously employed defined such effects in terms of two levels of phrase, maj-tg and min-tg. This was replaced by a model distinguishing four levels of phrase (subordinate and superordinate groupings of phrases at both maj-tg and min-tg levels), to test the independently developed theoretical notion that there is actually no principled limit to the depth of prosodic structure. The second model gave a significantly better account of the distribution of syllable durations. This suggests that the notion of indefinite prosodic depth has merit and may be of practical empirical relevance.

Acknowledgements

We would like to thank CSTR in Edinburgh and ATR in Japan for their continued support for this research.

REFERENCES

- [1] CAMPBELL, W. N., (1991) *Analog i/o Nets for Syllable Timing*, in *Speech Communication #9*, Elsevier Science Publishers B. V. (North Holland).
- [2] COOPER, W. & PACCIA-COOPER, J., (1980) *Syntax and speech* Harvard Univ. Press, Cambridge MA.
- [3] KLATT, D. H., (1976) *Linguistic uses of segment duration in English* JASA #59 pp 1208 - 1221.
- [4] LADD, D. R., (1986) "Intonational phrasing: The case for recursive prosodic structure" *Phonology Yearbook 3*, 311 - 340.
- [5] LADD, D. R., (1988) "Declination 'reset' and the hierarchical organization of utterances." JASA #84: 530 - 544.
- [6] LADD, D. R., (1990) "Compound Prosodic Domains". Occasional Paper, EULD; submitted to *Language*.
- [7] SELKIRK, E. O., (1984) *Phonology and Syntax: The relation between sound and structure*. Cambridge, Mass.: MIT Press.

ON VOWEL QUANTITY AND POST-VOCALIC CONSONANT DURATION IN DUTCH

Allard Jongman* and Joan A. Sereno*

Max Planck Institute for Psycholinguistics, The Netherlands

ABSTRACT

In Dutch, CV:Cən words contain a long vowel in syllable-final position while CVCən words contain a short vowel followed by an ambisyllabic consonant. Our measurements indicate that the duration of the intervocalic consonant is not affected by the quantity of the preceding vowel or its differential status as a tautosyllabic or ambisyllabic consonant. Instead, the duration of the second syllable is inversely affected by the duration of the vowel in the first syllable. These results are discussed in terms of the differential moraic representation of words containing long and short vowels.

1. INTRODUCTION

Durational properties of the speech signal have been well-studied for a variety of languages, including English, Swedish, Estonian, and Dutch. Factors known to influence segment and word durations range from phonetic and phonological factors up to syntactic and semantic factors. In this paper, we will concentrate on some phonetic and phonological factors influencing segment durations in Dutch. In particular, we will focus on the durational properties of minimal word pairs containing long and short vowels. Dutch has a phonemic vowel length contrast, as illustrated by the nouns 'taak' ([ta:k], 'task') versus 'tak' ([tak], 'branch'). In Dutch, long vowels can occur in both open and closed syllables whereas short vowels occur only in

closed syllables. When these nouns are pluralized by adding the suffix '-en', 'taken' ([ta:kən]) is said to consist of a first open syllable [ta:], containing the long vowel [a:] and a second syllable [kən] with a tautosyllabic [k]. On the other hand, 'takken' ([takən]) consists of the closed syllable [tak], containing the short vowel [a] and is closed by a so-called ambisyllabic [k]. In metrical phonology, these words would be represented as shown in Figure 1.

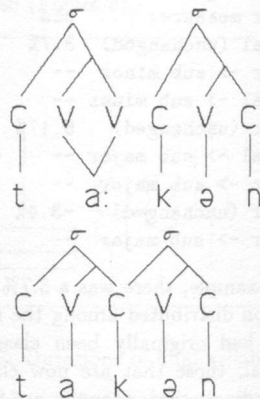


FIGURE 1: Metrical representation of 'taken' (top) and 'takken' (bottom).

The long [a:] in 'taken' is represented by two vowel slots on the CV tier. For 'takken', short [a] is represented by one vowel slot while the ambisyllabicity of medial [k] is reflected by the fact that it is attached to both the first and second syllable.

Given this difference in phonological representation between tautosyllabic and ambisyllabic medial consonants, one could ask whether this phonological contrast would surface as a phonetic difference in terms of consonant duration. In fact, our interest was triggered by a figure in the standard textbook on Dutch phonetics ([3]), reproduced here as Figure 2.

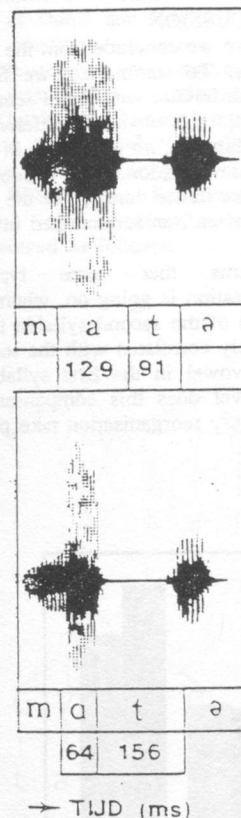


FIGURE 2: Waveforms of 'mate' (top) and 'matte' (bottom) and the relevant segment durations. (Figure taken from [3], p. 126, with permission from the authors).

This figure shows the waveform of 'mate' [ma:tə] in the top panel and that

of 'matte' ([matə]) in the bottom panel. The ambisyllabic [t] in 'matte' is much longer in duration than the tautosyllabic [t] in 'mate'. In fact, this figure suggests that the total duration of V plus C is constant for both the short and long vowel word, indicating a compensation whereby the consonant is lengthened by the same amount that the vowel is shortened. Nootboom's dissertation [2] also reported differences in Dutch nonwords of the type 'papápap' versus 'pa:pá:pa:p'. Nootboom found that the ambisyllabic consonant following the stressed short vowel was significantly longer (approximately 10 ms) than the tautosyllabic consonant following the stressed long vowel.

Given these intriguing findings, we felt that a closer look at these effects was warranted using additional minimal word pairs.

2. METHODS

Thirty-two test words (16 minimal pairs) were selected. These word pairs contained four long-short vowel pairs which have minimal spectral differences ([a:]-[a], [o:]-[o], [e:]-[I], and [ø:]-[œ]). These words were embedded in a carrier phrase in randomized order. Five speakers (three males, two females) were then recorded on a DAT-recorder. The test words were digitized at 10 kHz, and segment durations (initial consonant 'C1'; stressed vowel; medial consonant 'C2'; and second-syllable '-en') were then measured from a graphics display terminal, using standard visual and auditory criteria. All segment durations represent average values across all speakers and test words.

3. RESULTS

No significant differences in the duration of the medial consonant were found between long and short vowel word pairs. Contrary to earlier findings, there was no difference in duration between the ambisyllabic consonant following a short vowel (99 ms) and the tautosyllabic consonant following a

long vowel (101 ms).

Since there was no difference between medial consonants, one might expect that the difference in total word duration between words like 'taken' and 'takken' would simply amount to the difference in duration between long [a:] (176 ms) and short [a] (82 ms). Interestingly, however, this turned out not to be the case: The mean difference in vowel duration is 94 ms, while the difference in word duration is only 68 ms. Thus, there is a discrepancy of some 26 ms which has to be accounted for. The question then is: where did this 26 ms go?

There was a small but reliable difference between long and short vowel words in the duration of the initial consonant. Initial consonants preceding long vowels were somewhat longer (81 ms) than those preceding short vowels (77 ms). However, this only increases the difference in total word duration between 'taken' and 'takken', so that we now have to account for a 30 ms difference.

The only remaining possibility was

■ CVCCVC ■ CV:CVC

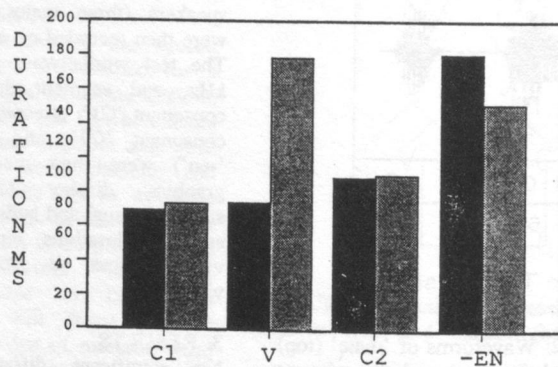


FIGURE 3: Segment durations in ms. Segment durations of words with short vowels and ambisyllabic consonants are represented by black bars while segment durations of words with long vowels and tautosyllabic consonants are represented by grey bars.

the second-syllable portion '-en'. The duration of '-en' turned out to significantly differ, depending on whether the first syllable contained a long or a short vowel. The duration of '-en' (179 ms) was longer when preceded by a first syllable containing a short vowel as compared to the duration of '-en' when preceded by a long vowel (147 ms).

4. DISCUSSION

What can we conclude from the present findings? To summarize, we found a small difference in initial consonant duration, an obvious difference in vowel duration, no difference in medial consonant duration, and a substantial difference in the duration of the second syllable '-en', as summarized in Figure 3.

It seems that some type of compensation is going on, where the duration of the second syllable is negatively correlated with the length of the vowel in the first syllable. At what level does this compensation or articulatory reorganisation take place?

One way to gain some insight into these data may be to look at the moraic representation of these words. The mora is essentially a phonological unit involved in the determination of syllable weight, such that light syllables are represented by one mora and heavy syllables by two moras. This distinction in terms of syllable weight plays an important role in the assignment of word stress in languages such as Dutch and English. Although what counts as a heavy or a light syllable varies across languages, it has been argued that in order to account for Dutch word stress, one has to assume that closed VC syllables are heavier than open long vowel syllables [1]. Within this framework, then, 'taken' and 'takken' would be represented as follows:



FIGURE 4: Moraic representation of 'taken' (top) and 'takken' (bottom). M indicates 'mora', F indicates 'foot'.

The long vowel in 'taken' is represented by one mora. The second syllable containing schwa is also represented by one mora. Together these two moras combine into one

Foot. For 'takken', the first syllable is represented by two moras, and the second syllable by one mora. The two moras of the first syllable combine into one Foot, and the second syllable forms a Foot of its own. (While there is some controversy about the moraic representation of schwa, it is important to note that under any analysis of schwa, 'taken' and 'takken' would still differ in terms of moraic structure). At this level of representation, then, the difference between words like 'taken' and 'takken' becomes obvious: 'taken' consists of one foot, while 'takken' consists of two feet. The puzzling effect of '-en' duration can now be described as follows: the stressless Foot of 'takken' is longer in duration than the weak branch of the Foot in 'taken'. In other words, the longer duration of '-en' when following a first syllable with a short vowel may be due to the fact that in this case '-en' forms an independent Foot.

At present, very little is known about the effect of metrical foot structure on phonetic segment duration, since it is very difficult to find minimal word pairs in terms of foot structure while keeping the phonetic context the same. The present results, where the same syllable is longer in duration when preceded by a heavy syllable relative to a light syllable, at least seem to suggest that foot structure may systematically affect segment durations.

REFERENCES

- [1] Lahiri, A., and Koreman, J. (1988). Syllable weight and quantity in Dutch. *Proceedings of WCCFL*, 7, 217-228.
- [2] Nootboom, S.G. (1972). Production and perception of vowel duration. Doctoral dissertation, University of Utrecht, The Netherlands.
- [3] Nootboom, S.G., and Cohen, A. (1984). *Spoken en verstaan*. Assen: Van Gorcum.

* Both authors are currently at: Department of Linguistics, UCLA, 405 Hilgard Avenue, Los Angeles, CA 90024, USA.

RHYTHMIC CATEGORIES: A CRITICAL EVALUATION ON THE BASIS OF GREEK DATA

Amalia Arvaniti

Department of Linguistics, Cambridge University, U.K.

ABSTRACT

This paper endeavours to show, on the basis of Greek data, that accent organises rhythm in all languages that have accent, irrespective of its exact acoustic manifestation and the rhythmic category a language is said to belong to. A formalisation of this idea can be achieved by means of a hierarchical abstract representation of rhythm. Such a representation can account both for the rhythm of Greek, in which stresses are sparse, and for the alternating rhythm of English by showing that they are both based on the same principle, grouping created by accent. Thus rhythm is shown to have an acoustic basis (i.e. accent) rather than a purely perceptual one.

1. INTRODUCTION

The validity of the stress/syllable-timing distinction has often been questioned, due to the well-known lack of acoustic evidence in its favour (for a review see [5]). Several efforts have been made to rescue the notion of two rhythmic categories from oblivion, usually by appealing to the perceptual rather than acoustic basis of the distinction [5].

One of the latest efforts is that of Dauer [4] who proposed a system of "quantifying" rhythm on the basis of phonological and phonetic criteria, such as syllable weight and the acoustic correlates of accent. A language's "score" in this system is meant to show its rhythmic tendency towards either stress- or syllable-timing. Dauer's system is not accurate however: the "score" of Greek

indicates that in this language accent (manifested as stress) is barely perceptible; hence Greek has a tendency for syllable-timing. Yet Greek stress has a high functional load. Moreover, phonetic research on Greek has shown first, that stressed syllables in running speech are easily identified both by native speakers and non-native phoneticians [3]; second, that the acoustic and perceptual correlates of Greek stress are F₀, duration and amplitude [2].

I believe that the problem with Dauer's system is that it attributes the same weight to factors which are relevant to the description of speech rhythm and others which, in my opinion, are not, such as the acoustic implementation of accent. Dauer's [4] insistence on phonetic correlates seems to be related to her opinion that the abstract phonological representation of rhythm, which is based on accent grouping, "tends to make all languages look [rhythmically] alike" (p. 447). I would like to suggest, instead, that this is an advantage of the phonological representation because it captures the fact that the main contributor to rhythm is always accent, irrespective of its acoustic correlates. Furthermore, I would like to propose that the differences between languages lie in the precise way in which accent achieves rhythmic grouping. My hypothesis is that the difference between Greek and English is that in Greek stresses are less frequent than in English, and that this is due to the lack of rhythmic stress in Greek. The latter part of my hypothesis disagrees with the proposals of Malikouti-

Drachman & Drachman [6] and Nespor & Vogel [7] who suggest that rhythmic stresses are used in Greek to create alternating rhythm. An experiment was conducted in order to test the above hypothesis.

2. METHOD

The experiment's material consisted of two sets of test words, although due to space limitations only the data from one set are presented here; there were no differences between the two sets.

The tetrasyllabic test words of each set were phonemically identical (see Table 1), but while (a) had antepenultimate primary stress, (b) had penultimate stress and (c) final stress. The differences in the carrier phrases resulted in different possibilities for rhythmic stress on each test word's first two syllables: (a) cannot carry rhythmic stress; (b) may have rhythmic stress on the initial syllable, while (c) may have rhythmic stress on the antepenult. Within a set, the initial syllable of (b), /*xa*/, is compared to the unstressed initial syllables of (a) and (c). Also, the antepenult of (c), /*mo*/, is compared to the antepenult of (a), /*mo*/, and to the unstressed antepenult of (b), /*mo*/.

Table 1: The recording material.

- (a) /*eleje xa'moyela ka'la*/
She/he used to say smiles (n.) well.
- (b) /*i'pe .xamo'yela kaθa'ra*/
She/he said smile (imper.) clearly.
- (c) /*θa 'po xa.moye'la kanoni'ka*/
I will say she/he smiles properly.

The test sentences were read 6 times each by 4 native speakers of Greek, from a randomised list typed in Greek. The speakers were in their twenties, spoke standard Greek and were naive as to the purposes of the experiment.

The material was low-pass filtered at 8 kHz and digitised at 16 kHz. F₀, amplitude integral (AI), average amplitude (RMS) and duration measurements were obtained. F₀ was measured using a signal-processing package which performed F₀ measurements every 10 ms over a 32 ms Hamming window. AI was calculated automatically between specified points of the waveform

which included the syllable nucleus. The original AI measurements, which were in arbitrary units given by the signal-processing package, were normalised; the values presented in Figures 2 and 3 are ratios of syllable to word AI expressed as percentages (for details see [1]). Two-way ANOVAs (stress type x speaker) were performed on the AI data.

RMS was measured and normalised in the same way as AI. Duration was measured from spectrograms. Although duration and RMS were analysed statistically, they will not be discussed here as their effect is reflected in AI, in which durational and average amplitude information are combined.

3. RESULTS

There is no evidence that syllables said to carry rhythmic stress are associated with F₀ perturbations. Figure 1 shows that the F₀ contour is determined solely by the position of primary stresses; F₀ starts rising on a stressed syllable, reaching its peak on the beginning of the following unstressed one; at this point F₀ starts falling until the next stressed syllable is reached.

AI yields very slim evidence for rhythmic stress. Figure 2 shows that the unstressed /*xa*/s and /*xa*/ have the same AI (F(2,40)=0.23, n.s.), while Figure 3 shows that /*mo*/ has lower AI than /*mo*/ (F(1,20)=189.23). In the comparison of /*mo*/ with /*mo*/ there is interaction between speakers and type of stress; while in VK's and SC's speech /*mo*/ has the same AI as /*mo*/, in AA's and DT's speech /*mo*/ has higher AI than /*mo*/ (VK: F(1,20)=0.99 n.s.; SC: F(1,20)=0.04 n.s.; DT: F(1,20)=20.89 p<0.000; AA: F(1,20)=5.22 p<0.03). DT's and AA's data are the only ones to show evidence for rhythmic stress.

In short, only in 2 out of 8 possible instances, does rhythmic stress materialise as high AI. These two instances are due to high RMS rather than duration.

4. DISCUSSION

The empirical evidence is very slim: rhythmic stress appears in a few cases in the speech of some speakers only, while its single acoustic correlate

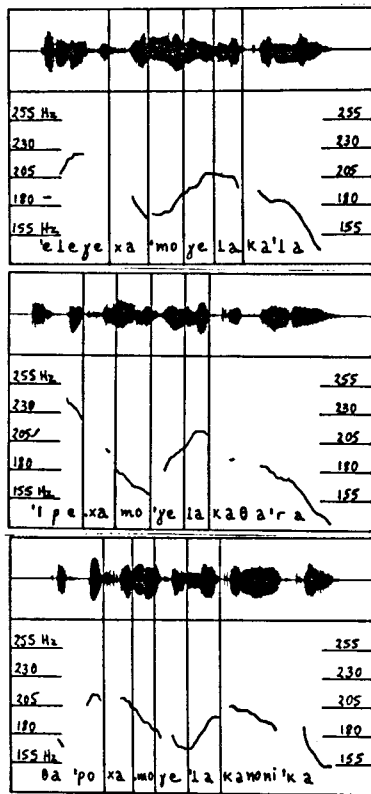


Figure 1: Typical F0 contours (smoothed) of /xa'moyela/, /xamo'yela/ and /xa.moye'la/; speaker DT.

is amplitude which Botinis [2] has shown to be the least reliable perceptual cue to stress. Moreover, the native speakers of Greek are not aware of the presence of rhythmic stress. Thus, there is no justification for representing phonologically stresses other than the primary stress of each word.

The fact that there is only one stress per word in Greek, combined with the high number of syllables per word, leads to the conclusion that stresses in Greek are indeed sparse. Does this, however, justify classifying Greek as having a tendency for syllable-timing? I believe that such a classification is irrelevant, and also

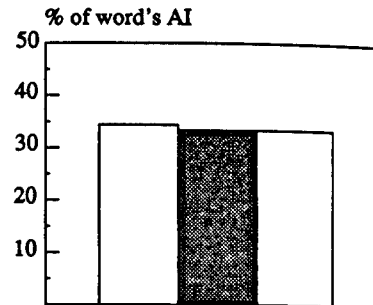


Figure 2: Mean values of amplitude integral of /xa/ for all subjects. Clear bars represent AI of the unstressed /xa/s of /xa'moyela/ (left) and /xa.moye'la/ (right), dark shaded bar represents AI of /xa/ of /xamo'yela/.

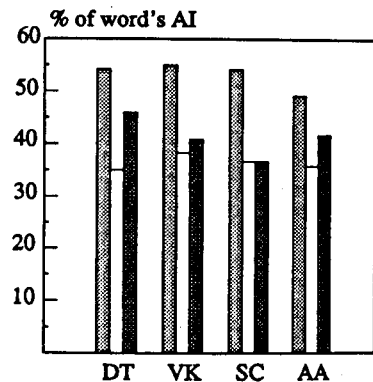


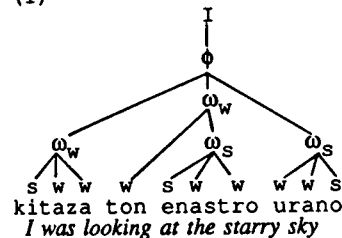
Figure 3: Mean values of amplitude integral of /mo/ for each subject separately. Light shaded bars represent AI of /mo/ of /xa'moyela/, clear bars AI of /mo/ of /xamo'yela/, and dark shaded bars AI of /mo/ of /xa.moye'la/.

incorrect because it ignores the importance of stress in Greek rhythm.

The rhythmic features of Greek can be more adequately represented by an abstract formalisation which takes the form of n-ary branching trees. As shown in (1), these trees follow the Strict Layer Hypothesis and comprise only 4 prosodic levels: the syllable (σ), the phonological word (ω) (a compound domain which includes

clitics), the phonological phrase (ϕ), and the intonational phrase (I). Thus, Greek has a "flatter" rhythmic structure than English, i.e. fewer degrees of stress. In particular, Greek has no foot structure. The lack of foot structure explains why the Rhythm Rule does not operate in Greek; as English examples show, this rule does not operate when there is no strong position within a word to which a stress may move; e.g. *mároon júmper. As Greek has only one strong syllable per word, the stress cannot move from this syllable.

(1)



This structure captures the fact that in Greek stresses, without being as frequent as they are in English, are still the prominences around which rhythm is organised. In addition, this structure shows that Greek rhythm is not based on binary patterning, since long sequences of unstressed syllables are not eliminated by means of rhythmic stress.

The fact that the rhythm of Greek is not binary has important consequences for the phonology of rhythm, as it has often been assumed that rhythmic patterns are universally binary. I believe that this superficial disagreement between theory and empirical data stems from the confusion between the phenomena linguists attempt to describe (and if possible explain) and the representations used for this purpose. While the phenomena, such as the rhythmic organisation of speech, are universal, formalisms need not be. As the present data suggest, binary branching, which successfully represents the alternating rhythm of English, is not adequate for the representation of Greek rhythm as

well. If, however, the constraints on the formalism are relaxed so that binary branching is not the only option, both languages with binary rhythmic patterns and languages with n-ary ones will be adequately represented by abstract structures of the form presented in (1).

Relying on an abstract representation does not mean that the search for the acoustic manifestation of rhythm is not a legitimate target; indeed there are compelling reasons for such as step, like the implementation of speech synthesis and automatic speech recognition models. However, using the acoustic features of rhythm as a means of describing languages as stress- or syllable-timed (or as having a tendency for either) does not serve any purpose apart from classifying languages according to the impressions of non-native speakers.

REFERENCES

- [1] ARVANITI, A. (in press), "Secondary stress: evidence from Modern Greek", in Ladd, R. & G.J. Docherty (eds), *Papers in Laboratory Phonology II*, C.U.P.
- [2] BOTINIS, A. (1989), "Stress and Prosodic Structure in Greek: A Phonological, Acoustic, Physiological and Perceptual Study", Lund University Press.
- [3] DAUER, R.M. (1980), "Stress and rhythm in Modern Greek", unpublished Ph.D. dissertation, University of Edinburgh.
- [4] DAUER, R.M. (1987), "Phonetic and Phonological components of language rhythm", *Proceedings of 11th I.C.Ph.S.*, 5, 447-449.
- [5] LEHISTE, I. (1977), "Isochrony reconsidered", *Journal of Phonetics*, 5, 253-263.
- [6] MALIKOUTI-DRACHMAN, A. & G. DRACHMAN (1980), "Slogan chanting and speech rhythm in Greek", in Dressler W., O. Pfeiffer, J. Rennison (eds), *Phonologica 1980*, Innsbruck, 283-292.
- [7] NESPOR, M. & I. VOGEL (1989), "On Clashes and Lapses", *Phonology*, 6, 69-116.

ON THE ACQUISITION OF SEGMENTAL DURATION IN NORMAL AND ARTICULATION DISORDERED 4-YEAR-OLDS

J.A. Bauman-Waengler and H.- H. Waengler

University of South Alabama, Mobile, USA

ABSTRACT

Aspects of sound duration were longitudinally investigated in the speech of two groups of preschoolers, children with normal speech/language development and children with articulation disorders. Results included increased longitudinal dissimilarities between the groups.

1. INTRODUCTION

The importance of suprasegmentals within the speech/language development of children has recently been emphasized [3], [5]. This prevailing tendency revives the old question of segmental duration which has been extensively investigated for well over 100 years [6], [4].

Speech sounds not only vary in intrinsic duration, they also influence another durationally in regular ways. The possibility to establish frequently reconfirmed durational rules bears witness to this fact.

Comparable findings in speakers of many different languages suggest common physiological properties behind these relational manifestations. Vowels seem to be generally longer before voiced than before voiceless stops, for example. Durational peculiarities between languages, on the other hand, point to simultaneous language dependent learning effects.

More recently several investigators have examined consonant and vowel

duration in the speech of children with specific articulatory deficits [8], [1], [7]. These studies have shown a greater durational variability in the children with articulatory difficulties. However, certain durational contrasts such as voicing appear to be maintained by children with disordered articulation as well.

When and how do children acquire the durational standards of their native language, whether they are based on physiological maturation of articulatory timing events or on learning processes? Are there differences in the acquisition of these standards between children with and without certain speech/language deficiencies? In an attempt to procure initial results, this study was designed to investigate sound duration longitudinally in the speech of two groups of preschoolers, children with normal speech/language development and children with disordered articulation.

2. METHOD

2.1. Subjects

Twelve children (7 boys, 5 girls) between the ages of 4;0 and 4;11 served as subjects. All subjects had passed a hearing screening test and had no reported history of neurological, physical, psychological or behavioral impairment.

Standardized tests, the Test of Early Language Development (TELD) and

the Weiss Comprehensive Articulation Test (WCAT), established the two groups of subjects (N = 6/group): The NORM Group, children who achieved age appropriate scores on both, the language and the articulation tests, and the ARTICULATION DISORDERED group, children who performed within one standard deviation of the norm on the language test but at least four points below age on the articulation test.

2.2. Data Collection/Analysis

The material consisted of 42 CVC contracts with initial /p/ plus either /i/, /k/ or /u/, followed by all voiced and voiceless stops and fricatives of American English.

These stimuli were presented orally and repeated three times in three month (10-13 weeks) intervals. At each test time all stimulus items were modelled by the same person. Misarticulations of a stimulus item in any part resulted in repetition of the logatom by the investigator. After 3 incorrect responses, the child's final production represented his/her response to that particular item. The responses of the subjects were audiorecorded and later analyzed from oscillographic displays utilizing the Micro Speech Lab computer program, Kay Elemetric Corp., 1985. To ensure comparability of the measurements, certain segmentation principles were established and strictly adhered to.

3. RESULTS

For the statistical analyses, the data were organized in two ways: absolute and relative duration. The absolute duration consisted of the real time measurements of each acoustic sound representation: initial [p], medial vowel, and final consonant. The relative duration is a calculation devised to minimize rate differences between speakers and consists of the sound

length/average sound duration of the entire word [2].

Voicing Contrast: The first question addressed was: Do the two groups of subjects adhere to the durational standard that vowels preceding voiceless consonants are shorter than those preceding voiced consonants? Absolute and relative durations were analyzed by means of a two tailed t-test. An alpha level of .05 was established for significance.

For all three testing times both the Norm and the Articulation Disordered group produced significantly longer durations when the vowels preceded voiced consonants. This held true for the three vowel qualities, regardless of whether the vowel preceded stops or fricatives, and for both absolute and relative duration. Therefore, although the group of children with an articulation disorder often misarticulated the final consonant, the durational contrast of the preceding vowel was observed. This is especially interesting in light of the fact that clearly over half of the misarticulations represented devoicing of the final consonant.

Group Specific Durational Differences over Time: The second question addressed was: Do the Norm versus Articulation Disordered subjects demonstrate group specific variations in sound duration for each of the three testing times? To answer this question, a multivariate analysis of variance was used to test for interactions among group by absolute and relative duration of the 1) initial [p] (IP), 2) medial vowel (MV) and 3) final consonant (FC) for Time 1, 2, and 3 (T1, T2, T3).

As can be noted in Tables 1 and 2 there was always a statistical significance between the Norm and Articulation Disordered groups for all absolute and relative durations at T3. From these results there appears to be a change over time with the two groups

of subjects becoming more dissimilar during this testing period.

Table 1: MANOVA Results for the Absolute Durations of Initial-p (IP), Medial Vowel (MV), and Final Consonant (FC) X Group for Time 1, 2, and 3 (T1, T2, T3)

	F-Ratio	df(1,11)	p
IP-T1	7.29		.007
IP-T2	17.22		.0001
IP-T3	5.20		.02
MV-T1	2.42		.122
MV-T2	12.67		.0005
MV-T3	5.07		.027
FC-T1	2.51		.116
FC-T2	3.60		.06
FC-T3	8.93		.004

Table 2: MANOVA Results for the Relative Durations of Initial-p (IPR), Medial Vowel (MVR), and Final Consonant (FCR) X Group for Time 1, 2, and 3 (T1, T2, T3)

	F-Ratio	df(1,11)	p
IPR-T1	11.40		.001
IPR-T2	5.35		.02
IPR-T3	11.64		.001
MVR-T1	2.26		.135
MVR-T2	.16		.69
MVR-T3	6.48		.01
FCR-T1	10.00		.002
FCR-T2	3.51		.06
FCR-T3	13.26		.0005

Individual Durational Variations for Each Group: The third question addressed was how each individual child manipulated segmental duration throughout this testing period. It was also important to determine whether or not the children in both groups varied from known norms. To establish this theoretical norm, the means of the Norm group for both absolute and relative durational measures were utilized.

An overall mean for each context condition (the three vowel qualities, voiced and voiceless stops and fricatives) was calculated from the Norm group. A 95% confidence interval was set around these means. Each child's performance was then compared to this confidence interval. A percentage was determined based on how many means of the individual child were found within this 95% confidence interval. As is to be expected, the Norm group had a much higher percentage of means within the 95% confidence interval. However, individual variations could be seen, for example: Subject 3 = 37%, Subject 5 = 46%, Subject 1 = 89%. Thus, not all members of the Norm group reacted durationally in the same manner. For the group of children with articulation disorders percentage scores were lower as well as less variable (range = 39-69%). While three of the children fell clearly outside the Norm means (Subject 8 = 42%, Subject 9 = 39%, Subject 12 = 45%), the overall range of percentages was clearly reduced when compared to the Norm group.

4. DISCUSSION

This investigation supports previous findings [1], [7], [8]: even children with articulation disorders adjust the vowel duration preceding voiced consonants although they produced in fact devoiced final consonants. However, in this respect group variations were found. In addition, more incidences of significant differences could be noted between the two groups at time 3. One feasible hypothesis is that the subjects of the Norm group did develop fairly fast during the testing time interval while the Articulation Disordered group showed little or no durational maturation. According to our findings, then, it is not that the children within the Articulation Disordered group were developing durationally in a dif-

ferent direction, but rather that their pace of change was drastically different. Finally, a fair degree of individual variation was noted in both groups of subjects, although the range of variability found was wider within the Norm group.

5. REFERENCES

- [1] CATTS, H. and JENSEN, P. (1983). "Speech timing of phonologically disordered children: Voicing contrast of initial and final stop consonants", *JSHR*, 26, 501-510.
- [2] ESSEN, O. v. (1979). "Allgemeine und Angewandte Phonetik", 5th ed., Akademie Verlag, Berlin.
- [3] KRAUSE, S.E. (1982). "Developmental use of vowel duration as a cue to postvocalic stop consonant voicing", *JSHR*, 25, 388-393.
- [4] MEYER, E.A. (1903). "Englische Lautdauer", *Skripter utgifna af K. Humanistiske Betenskaps-Samfundet i Uppsala*, VIII, 3, Uppsala.
- [5] OLLER, D.K. and SMITH, B.L. (1977). "The effect of final-syllable position on vowel duration in infant babbling", *JASA*, 62, 994-997.
- [6] SIEVERS, E. (1881). "Grundzuge der Phonetik", Leipzig.
- [7] SMIT, A. and BERNTHAL, J. (1983). "Voicing contrasts and their phonological implications in the speech of articulation-disordered children", *JSHR*, 26, 486-500.
- [8] WEISMER, G. and ELBERT, M. (1982). "Temporal characteristics of 'functionally' misarticulated /s/ in 4- to 6- year old children", *JSHR*, 25, 275-287.

This research was supported by a University of South Alabama Research Council grant.

LES CONTRAINTES TEMPORELLES DES TYPES CONSONANTIQUES
SUR LE TIMING MANDIBULAIRE DE LA QUANTITÉ
en arabe tunisien

BOUSSAFFA F. JOMAA M. SOCK R.

Institut de la Communication Parlée
URA CNRS n° 368 Grenoble France

ABSTRACT

The general aim of this study is to examine the effects of consonants on the timing of the jaw in producing length contrasts (of the type [CeCCa] vs. [CeeCa]) in Tunisian Arabic. More specifically initial consonants [ʒ] and [z] test the possible influence of the protrusion gesture at the lips on the up-down movement of the jaw. Our results show that the influence of consonantal protrusion tends to enhance length contrasts and that – among measured cycles – the “vocalic” velocity one should offer the best temporal domain to run comparisons of language variation in the field of Arabic dialectology.

1. INTRODUCTION

Notre recherche a pour but de comprendre l'organisation des gestes qui gouvernent la production de la quantité phonologique et celle des types consonantiques en arabe. Étant donné que l'organisation temporelle ou *timing* implique, en général, des contrôles temporels fins à des finalités sémiotiques, nous essayerons de mettre en évidence les stratégies motrices, propres à nos tâches phonologiques complexes.

Les études antérieures portant sur l'organisation temporelle de la parole et plus spécifiquement sur le *timing* de la quantité phonologique en arabe, nous montrent qu'il est plus efficace d'examiner les stratégies articulatoires-acoustiques, en dégagant les structurations globales de patrons de phases, au lieu de se livrer exclusivement à une recherche d'invariance (cf. [5], [1], [6]).

Notre étude analyse les caractéristiques temporelles de la mandibule pour des gestes à buts linguistiques définis, cet articulateur pouvant être considéré comme la porteur rythmique des articulateurs portés, langue et les lèvres. Nous nous focaliserons sur deux cycles mandibulaires – initiation des gestes et vitesse “vocalique” – comme champs d'observation des deux tâches linguistiques. L'augmentation de la vitesse d'élocution, nous permettra de tester les degrés de résistivité de l'opposition.

2. MÉTHODE

2.1. Corpus

Dans le cadre de notre étude [2] nous avons élaboré un corpus qui, non seulement appartient à une série de mots linguistiquement pertinents, mais qui permet aussi d'examiner les différents degrés de couplage de la mandibule avec la langue et les lèvres. Ce couplage langue et lèvres présente des coordinations interarticulateurs différentes selon la tâche linguistique à accomplir. Notre corpus est conçu pour examiner les effets des couplages entre la lame de la langue [ʒ, z] ou encore sa masse [e, a] et la mandibule. Les consonnes initiales [ʒ, z] nous permettent de tester l'influence possible du geste de protrusion des lèvres sur le déplacement mandibulaire vertical. Ainsi les mesures recueillies sont révélatrices, des couplages linguo- et/ou labio-mandibulaires. Pour un test sur la cohérence structurale des phasages entre les niveaux articuloire et acoustique, cf. [4].

Les items choisis forment une paire minimale : [ʒezza] (variante [zezza]) “il a tondu” vs. [ʒeeza] (var. [zeeza]) “il a récompensé”, ou bien “ça suffit”.

Le locuteur choisi (pour une comparaison avec d'autres locuteurs, cf.

[3]) possédait les variantes dialectales [ʒ] et [z].

2.2. Acquisition

Tous les items ont été introduits dans des phrases porteuses affirmatives. Chaque phrase a été répétée 12 fois, en ordre aléatoire, en chambre sourde, par un locuteur tunisien, à deux vitesses d'élocution. La première série a été réalisée à la vitesse normale d'élocution du locuteur. La deuxième a été obtenue avec un débit rapide.

Les signaux de déplacement vertical de la mandibule ont été recueillis à l'aide d'un kinésiographe mandibulaire (K5AR) et échantillonnés à 160 Hz pour édition en synchronie avec le signal acoustique numérisé à 8 kHz.

2.3. Mesures

À l'aide des événements qui peuvent être repérés sur les signaux articuloires des items choisis (cf. [3]), nous avons pu retenir deux cycles, initiation des gestes et vitesse “vocalique” qui se sont révélés être des domaines privilégiés pour la programmation des syllabes VC.

– le cycle d'initiation des gestes, peut être repéré par la reproduction de l'événement VCO (Vocalic Cycle Onset), soit le premier signe d'abaissement de la mandibule pour produire le cycle vocalique. Dans ce cycle, l'arrivée de l'événement CCO (Consonantal Cycle Onset) – soit les premiers signes de l'élévation de la mandibule pour le cycle consonantique – nous donne la phase “vocalique” (VCO-CCO).

– le cycle de vitesse “vocalique” est défini comme la reproduction de l'événement MVV (Maximum Vocalic Velocity), soit le pic de vitesse dans l'abaissement de la mandibule pour produire la voyelle. Dans ce cycle l'arrivée de l'événement MCV (Maximum Consonantal Velocity) – soit le pic de vitesse de l'élévation de la mandibule pour produire la consonne – nous donne la phase “vocalique” (MVV-MCV).

3. RÉSULTATS

3.1. Opposition de quantité dans le contexte initial [ʒ]

[ʒezza] vs. [ʒeeza]

Dans le cycle d'initiation des gestes VCO (fig. 1), la séparation des deux tâches linguistiques est possible en débit normal. Nous observons une différence de phase entre les deux tâches linguistiques. En outre, la catégorie VVC est plus longue en cycle que VCC.

L'opposition repose donc sur une différence de phase et dans une moindre mesure de cycle. En débit rapide l'opposition a tendance à être neutralisée

Dans le cycle de vitesse maximale MVV (fig. 2), la séparation des deux tâches se réalise encore mieux par la phase, que dans le cycle VCO. Par contre il n'y a plus de différence significative sur le cycle : nous assistons à un phénomène d'isochronie (VVC = VCC). Comme dans le cycle précédent cette opposition est neutralisée avec l'augmentation de la vitesse d'élocution.

3.2. Opposition de quantité dans le contexte initial [z]

[zezza] vs. [zeeza].

Dans le cycle d'initiation des gestes VCO (fig. 3), l'opposition entre les classes VCC et VVC ne repose pas aussi nettement sur la phase que dans le cas des items avec [ʒ]. Par contre, l'opposition se fait encore plus nettement sur le cycle. Comme dans les cas précédents elle se neutralise avec l'augmentation de la vitesse d'élocution.

En ce qui concerne le cycle de vitesse “vocalique” MVV (fig. 4), l'opposition entre les classes phonétiques VCC et VVC repose, en débit normal, autant sur le cycle que sur la phase. En augmentant le débit, nous assistons, là aussi, à une neutralisation des différences temporelles entre les tâches linguistiques.

4. CONCLUSION

En conclusion, nous pouvons dire que ces résultats montrent qu'en débit normal, l'opposition de quantité se réalise dans les deux cycles articuloires retenus, que ce soit par la phase et/ou par le cycle. Mais elle se neutralise toujours en débit rapide.

Le cycle de vitesse “vocalique” MVV reste le plus discriminant pour mettre en évidence les oppositions de quantité phonologique. Rappelons que ceci a été confirmé par ailleurs (cf. [4]) pour d'autres dialectes de l'arabe et pour le français.

En ce qui concerne l'influence des types consonantiques sur l'opposition de quantité, nos résultats montrent que les types choisis ont un effet caractéristique sur les cycles. Un effet qui n'est pourtant pas général pour toutes les tâches phonologiques : l'influence de la protrusion consonantique a tendance à augmenter la durée des cycles d'initiation (VCO) ou de vitesse (MVV), mais essentiellement pour la catégorie avec voyelles brèves, VCC. Notons que cette différence de durée entre les items avec [ʒ-] et [z-] n'existe plus en débit rapide.

Une possible explication de l'influence intrinsèque de chaque type consonantique sur le patron temporel, serait que le recrutement multi-articulateurs de la mandibule (langue et lèvres), demandant plus de précision de contrôle, pourrait être à l'origine de gestes articulatoires plus lents et, par conséquent, de cycles beaucoup plus longs pour les classes de brèves, VCC.

Ainsi une approche relative s'avère rentable pour comprendre la programmation des gestes de quantité en fonction des types consonantiques. Nous nous proposons donc, dans le cadre de la dialectologie arabe, de nous baser sur le cycle de vitesse "vocalique" (MVV) – comme domaine temporel – pour établir la typologie des stratégies de réalisation de l'opposition de quantité. Dans ce sens, une hypothèse pourrait être que les dialectes qui présentent des "réflexes" [3] (comportant une composante labiale) pourraient renforcer les contrastes.

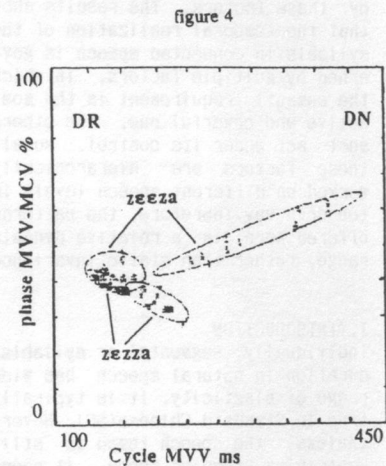
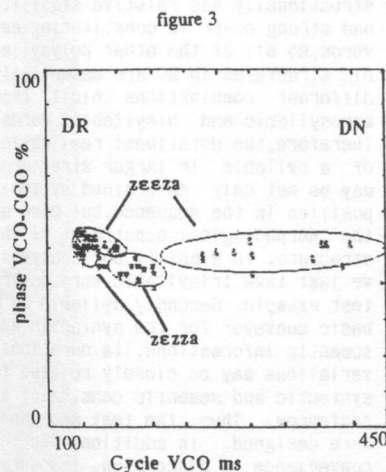
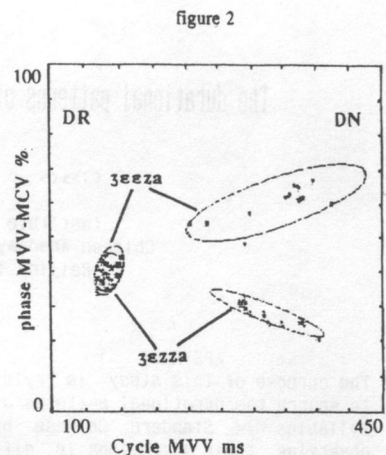
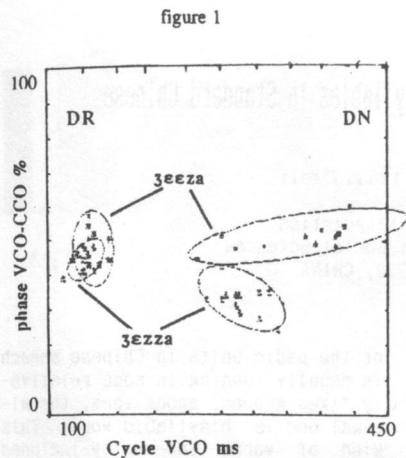
REMERCIEMENTS

Nos remerciements vont à Christian ABRY, qui a suivi ce travail de près et à Christine DELATTRE pour son aide dans la collecte des données, ainsi qu'à Gang FENG pour son aide en traitement des signaux.

5. RÉFÉRENCES

- [1] AL DOSSARI, A. (1989), *Le phasage des gestes mandibulaires vocaliques et consonantiques en arabe koweïtien*, Mémoire de D.E.A., Université Stendhal, Grenoble.
- [2] BOUSSAFFA, F. (1990), *Organisation temporelle et types consonantiques en arabe tunisien*, T.E.R de Maîtrise, Université Stendhal, Grenoble.
- [3] DELATTRE, C. JOMAA, M. WORLEY, C. ABRY, C. (1989), "The Phasing of the Jaw in Consonant and Vowel Lengthening : Arabic and French Patterns", *Europ. Conf. on Speech Comm. and Techn.* 2, 416-419.
- [4] DELATTRE, C. JOMAA, M. AL DOSSARI, A. WORLEY, C. SOCK, R. (1990), "Comparaison articulatoire-acoustique des structures temporelles en arabe et en français. Ou peut-on séparer les classes dans les VC?", *18èmes J.E.P du G.C.P de la S.F.A.*, Montréal, 113-118.
- [5] JOMAA, M. ABRY, C. (1988), "La résistivité de la quantité vocalique aux variations de la vitesse d'élocution. Le cas de l'arabe tunisien", *17èmes J.E.P du G.C.P de la S.F.A.*, 231-236.

[6] RHARDISSE, N. (1989), *La résistivité des oppositions de quantité vocalique et consonantique de l'arabe marocain aux variations de la vitesse d'élocution*, T.E.R de Maîtrise, Université Stendhal, Grenoble.



Figures 1 à 4 (arabe tunisien)

Patrons de phasage en débit normal (DN) et rapide (DR) pour les oppositions de quantité [3eeza] vs. [3ezza] (en haut, fig. 1 et 2) et [zeeza] vs. [zezza] (en bas, fig. 3 et 4) dans deux cycles de la mandibule : le cycle d'initiation des gestes vocaliques (VCO, à gauche, fig. 1 et 3) et le cycle de vitesse "vocalique" (MVV, à droite, fig. 2 et 4). Pour la définition de ces cycles et des phases correspondantes, cf. texte.

The durational patterns of syllables in Standard Chinese

Cao, Jianfen

Institute of Linguistics
Chinese Academy of Social Sciences
Beijing 100732, CHINA

ABSTRACT

The purpose of This study is trying to search the durational patterns of syllables in Standard Chinese by observing their variations in different contexts and examining what the main control factors are and how the syllable's duration is affected by these factors. The results show that the temporal realization of the syllable in connected speech is governed by multiple factors, in which the semantic requirement is the most active and powerful one, the others must act under its control, so all these factors are hierarchically worked on different speech levels in top-down way. Therefore, the patterns offered here is a relative dynamic range, rather than static invariance.

1. INTRODUCTION

Individually, segments' or syllables' duration in natural speech has wide range of elasticity. It is typically true in Standard Chinese(SC). Nevertheless, the speech tempo is still kept in a regular range, it seems that there may exist some relational invariance in temporal distribution, by which the speech tempo is controlled.

Considering of the importance of the syllable in SC, the present investigation is concentrate on the control of syllable duration. Against this target, two sets of materials were tested in this study. The first set consists of 14 pairs of trisyllabic words, and the second set contains 6 pairs of sentences, all of them were designed according to the background as follows. Firstly, syllable, as one

of the basic units in Chinese speech, is usually running in some relatively fixed groups, among them, the minimal one is bisyllabic word. This kind of words inherently includes two types of stress and the corresponding temporal patterns[2]. Moreover, the bisyllabic words in SC constructionally has relative stability and strong power in constituting new words, so all of the other polysyllabic structures in SC are essentially different combinations built from monosyllabic and bisyllabic words. Therefore, the durational realization of a syllable in larger structures may be not only determined by their position in the sequence, but also by the morphological constraint of the structure. To simplify our analysis, we just take trisyllabic word as the test example. Secondly, syllable as a basic conveyer for the syntactic and semantic informations, its durational variations may be closely related to syntactic and semantic constraint of sentences. Thus, the test sentences were designed. In addition, for the convenience of discussion, the durational measurement is also extended to the acoustic record of monosyllabic and bisyllabic words which were made in our previous investigations [7,2].

2. EXPERIMENTAL RESULTS AND ANALYSES

2.1. Average duration

Generally, it is very hard to assign an average value for syllables' duration, since in natural speech, it is contextual-dependent. However, our test results do show a tendency that the number of syllables uttered per second is regularly between 5-6 for

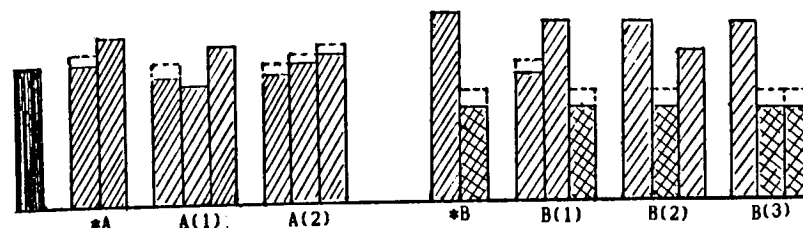


Figure 1. Durational patterns of syllables summarized from various isolated polysyllabic words:
*A. NM type bisyll. word; A(1) 2+1 NM type trisyll. word; A(2) 1+2 NM type trisyll. word.
*B. NT type bisyll. word; B(1) 1+2 NT type trisyll. word; B(2)&(3) 2+1 type trisyll. words.

Table 1. Average duration (ms) of different variations for syllables /dan/, /jie/, and /shi/ in normal stress utterance

position	in moderate speech			in faster speech		
	/dan/	/jie/	/shi/	/dan/	/jie/	/shi/
final	332.5	357.5	390	245	290	260
initial	267.5	268.8		222.5	263.3	
inter-vocalic						
* (a)		193	175		185	135
* (b)	187		148	161.3		122.5

* (a) 2nd syllables in trisyllabic words;
* (b) 3rd syllables in quadsyllabic words.

moderate speech and 6-7 for faster speech. This tendency seems not to be an accidental event, some previous studies have provided evidence from either production [3] or perception [4,6]. Based on these findings coming from different languages, we would suppose that the marginal seven or less of syllable number uttered per second might be one of the relational invariance in speech tempo existed throughout the languages in the world. Thus, it may be served as a relative scale for the comparison in this aspect.

2.2. Inherent difference

Some inherent difference of syllable duration in SC is observed from the isolated syllables. The main control factors for this difference, as it is universally acted in many languages, includes articulatory manner of the initial consonants and the opening of component vowels. In addition to these in SC, this kind of difference

is also closely related to the tonal distinction and stress type. Generally, a syllable with the third tone is usually longer than that with other tones ceteris paribus; and a syllable with normal(NM) type stress is clearly longer than that with neutral(NT) type's. Some previous studies have reported that the durational ratio of NT to NM type syllables is about 1:2 [5] to 3:5 [1], and it is contextually dependent, when neutral type bisyllabic words are uttered in connected speech, the durational difference between NT and NM type syllables becomes even more sharper [2].

2.3. Contextual variability

2.3.1. Variations in isolated words
Figure 1. gives specific variations of syllable in isolated polysyllabic words. The dark bar on the far left in the figure represent a relative scale for the comparison among the variations, its value is taken from

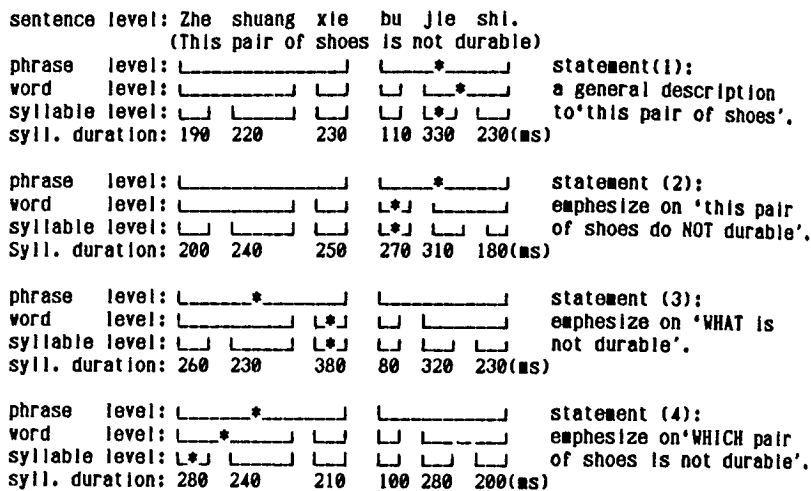


Figure 2. The schematic digram for the relationship between syllable duration and semantic requirement of the sentences.

the average duration of the word-initial syllables; The bars next to the dark one represent those patterns summarized from bisyllabic words tested in our previous investigation (2); and all of the rest are the patterns occurred in different type of trisyllabic words. Here the bars with oblique lines are the NM type tokens, and those with cross lines are the NT type tokens. In each case, the shadowed part of the bar shows a scope that a syllable's duration is exactly /likely reached, and the part within dotted lines illustrates the extent that possibly reached.

From figure 1. some effects could be clearly observed. (a) Position effect: Syllable duration in SC does follow position-effect rule, which is universally acted in many languages, such as word-initial shortening, word-final lengthening, and the intervocalic ones are often shorter in average. This effect also can be seen from Table 1; (b) Constructional constraint: As one of the basic units in SC, the durational patterns of bisyllabic words are often appeared in global. Look at the cases of B(1), B(2) and B(3) in Figure 1, it's obvious

that the long-short duration model for NT type's bisyllabic structure is kept in any position within the trisyllabic words, so the second syllable in B(1) and the first syllables in B(2) and B(3) have not been shortened as expected, but obviously lengthened instead, though all of them occur at the intervocalic position or word-initial position. This phenomena illustrate that the power of position-effect is limited, and it must be played under the control of morphological restriction.

2.3.2. Variations in connected speech

(a) Syntactic effect

The influence comes from syntactic aspect has been observed in our previous study (2). It may be helpful to present an example here: when the test word in a sentence is followed by an independent syntactic element, the average duration of last syllable in these words is around 292 ms; while it is compressed to 254 ms when it is followed by an auxiliary word, since this kind of words is usually not an independent element and has to be attached to its previous words, consequently, results in a limitation to the last syllable of the test words.

(b) Units' length and speech rate
If we look at back to the situation shown in Table 1, and make a simple comparison between the first row and the last two rows respectively or that between last two rows. It can be seen that syllable duration does decrease with the increase of syllable number in certain unit. At the same time, the relationship on inverse proportion between syllable duration and speech rate is also found from the cases in 'moderate' column and 'faster' column in the Table. However, that is not a simple and linear relation, because speech tempo is also adjusted by presence / absence and longer / shorter of the possible pauses occurred at phrase boundary, and syllable's elongation or compression performed in either case is only a tendency in gross.

(c) Semantic influence

Speech sound is the surface output represented the underlying semantic input, so the surface realization of a syllable duration in real speech is reasonably related to semantic content. Our observation indicates that this role is usually carried out through the difference in accent stress. Figure 2 gives one of the examples, in which, four different statements are made from the same syllable sequence but represent different semantic contents, and results in different distribution of syllable duration.

3. CONCLUSION

Based on the limited observation taken in this study, the durational patterns of syllables in Standard Chinese could be summarized as follows:

(a) Syllables as monosyllabic words uttered moderately in isolation is around 500 ms. Some intrinsic difference is existed among them, and it is conditioned by some inherent factors related to syllable itself. However, all of these tokens will be shortened when they occur in connected speech, the average ratio between this two cases is around 100:70.

(b) The durational realizations for individual syllable in real Chinese speech have a wide dynamic range, the lower threshold is around 100 ms, and

the higher one's is about 600ms. The contextual variations are shown in Table 1 of the text.

(c) The durational patterns summarized above in SC is governed by multiple factors acted in the different levels. The main factors, besides those inherent ones related to syllable itself, are coming from speech rate, syllable's position in the utterance, units' length in which the syllable appeared, morphological constraint of speech unit, difference in syntactic contexts, and the semantic requirement of certain sentences. Moreover, it seems that the last factor is the most active and powerful one, all of the others must be acted under its control. Therefore, these factors are hierarchically worked on a top-down way, the elongation or compression for a syllable is performed non-linearly through the mechanisms of compensation and adjustment, rather than in average or complete proportion.

4. REFERENCES

- [1] Cao, J. (1986), "Analysis of neutral tone syllables in Standard Chinese", *Applied Acoustics*, Vol.5, No.4:1-6.
- [2] Cao, J. (1989), "Temporal distribution of the bisyllabic words in Standard Chinese: An evidence for relational invariance and variability from natural speech", *RPR-IL(CASS)*/1989, 38-56.
- [3] Kohno M. & Tsu, S. (1989), "Rhythmic phenomena in a child's babbling and one-word sentence", *The Bulletin No.191, 6-13, The Phonetic Society of Japan*.
- [4] Kohno M. & Tomoko Tanioka, (1990), "The nature of timing control in language", *The Proceedings of ICSLP 90, Kobe, Japan*.
- [5] Lin, M. & Yan, J. (1980), "Acoustic characteristics of neutral tone in Beijing Mandarin", *Fangyan No.3: 166-178*.
- [6] Meiler G. (1956), "The marginal seven, plus or minus two?", *Psychological Review*, 63-2, 81-97.
- [7] Wu, Z. (1986), *The Spectrographic Album of Monosyllables of Standard Chinese*, Chinese Social Sciences Press, Beijing, 1986.

THE RHYTHM OF *TANKA*, SHORT JAPANESE POEMS: READ IN PROSE STYLE AND CONTEST STYLE

Yayoi Homma

Osaka Gakuin University, Osaka, Japan

ABSTRACT

In *Tanka* read as prose, the duration of each line including pauses was fairly regular. In the contest style, however, the duration of each line was not regular, but the durations of the first group of 5-7-5 lines and the following group of 7-7 lines were fairly regular, with a great amount of prepausal lengthening before each pause. The mora in Japanese as an isochronous unit of timing seems to be abstract, but it is a basic rhythmic unit. This basic unit coexists with larger rhythmic units, such as the poetic line read in the prose style or the group of lines read in the contest style of *Tanka*.

1. INTRODUCTION

This paper attempts to observe the rhythm of *Tanka*, short Japanese poems, produced in two different styles: prose and contest styles. According to Lehiste (1990)[3], "the prosodic system of a language is crystallized in the metric structure of its traditional poetry." *Tanka*, basically composed of 31 moras in 5-7-5-7-7 lines, is the most traditional Japanese poetry handed down since the Seventh Century.

In Homma (1985)[2], I investigated why this traditional type of verse sounds rhythmic, measuring the duration of each segment, mora, pause, line, and whole poem read as prose. I found that although the duration of each segment and mora had a greater range of difference and the number of moras of the lines was different,

the average durational differences of lines and whole poems were small.

The purpose of this paper is to confirm the results of my previous experiment and to compare the two different styles producing the same *Tanka* poems.

2. EXPERIMENT OF THE PROSE STYLE READING

2.1. Methods

I selected fourteen poems from the *One Hundred Poems from One Hundred Poets* (13th Century). Three of them have the regular 31-mora form, and the others have one hypermeter line in one of the five lines: one extra mora is added to one of the lines. The poems read as prose by five native speakers of Japanese were recorded on tape in the phonetic laboratory of the Ohio State University.

2.2. Measurements

Wide-band spectrograms of 70 poems (14 poems x 5 speakers) were made with a Kay-Sonograph (5500). In the present experiment, I measured the duration of lines, pauses, and whole poems, because I thought that the units in which moras would be manifested in Japanese must be larger than a line in poetry (Homma, 1985[2]). I got the average duration of moras by dividing the duration of the lines by the number of moras. The duration of initial stop consonants after pause was impossible to measure. All the values were rounded to the nearest 5 milliseconds.

Table I. Average duration (\bar{x} , ms) and standard deviation (SD) of moras, pauses, lines, and whole poems of *Tanka* read in the prose style

lines	(1)5-mora	(2)7-mora	(3)5-mora	(4)7-mora	(5)7-mora	whole poem
mora \bar{x}	155	136	149	137	138	---
SD	8.4	7.2	7.3	3.6	7.5	---
pause \bar{x}	223	41	437	59	---	---
SD	55.5	29.0	46.8	26.6	---	---
line \bar{x}	1017	1018	1200	1035	994	5264
SD	74.7	87.2	64.4	54.0	64.4	149.4

2.3. Results

Table I shows the average duration (\bar{x} , ms) and standard deviation (SD) of moras, pauses, lines, and whole poems of 70 *Tanka* read in the prose style.

From Table I, the following points were observed.

(1) The duration of each line including

style was a unit of temporal programming in Japanese (Lehiste, 1990[3]).

3. EXPERIMENT OF THE CONTEST STYLE READING

3.1. Methods

The same fourteen poems from the experiment of the prose style reading were

Table II. Average duration (ms) of moras of regular and irregular lines as read in the prose style

lines	(1)	(2)	(3)	(4)	(5)
regular	156	136	150	137	139
irregular	146	135	141	133	133
average	155	136	149	137	138

pause was fairly regular except for the third line, at the end of which the speakers took a breath.

(2) The adjustment for equidistant lines was achieved in two ways: first by changing the duration of pause, and secondly by changing the duration of moras by means of the speech rate.

Longer lines, especially irregularly longer lines, were read a little faster; thus the average duration of moras became shorter.

Table II presents the average duration of moras of regular and irregular lines with hypermeter.

These results supported Homma, 1985[2]. The poetic line of *Tanka* read in the prose

studied. This time I used a tape which was made by a publisher for the people who would like to participate in the time-honored contest played with these hundred verses on cards. The tape was recorded by one of the authorized speakers who read the poems in the contest style. The speaker made no pause between the lines, but took a long 9-second pause between 5-7-5 and 7-7 lines in accordance with the rules of the contest.

3.2. Measurements

Measurements were taken in the same way as in the first experiment.

3.3. Results

Table III shows the average duration (\bar{x} ,

Table III. Average duration (\bar{x} , ms) and standard deviation (SD) of moras, pauses, lines, and whole poems as read in the contest style

lines	(1)5-mora	(2)7-mora	(3)5-mora	(4)7-mora	(5)7-mora	whole poem
mora \bar{x}	208	258	434	388	454	---
SD	19.9	19.2	92.1	38.6	42.3	---
pause \bar{x}	0	0	0	65	---	---
SD	0	0	0	166.7	---	---
line \bar{x}	1073	1853	2221	2820	3265	11232
SD	141.6	88.2	435.7	278.8	211.6	788.4

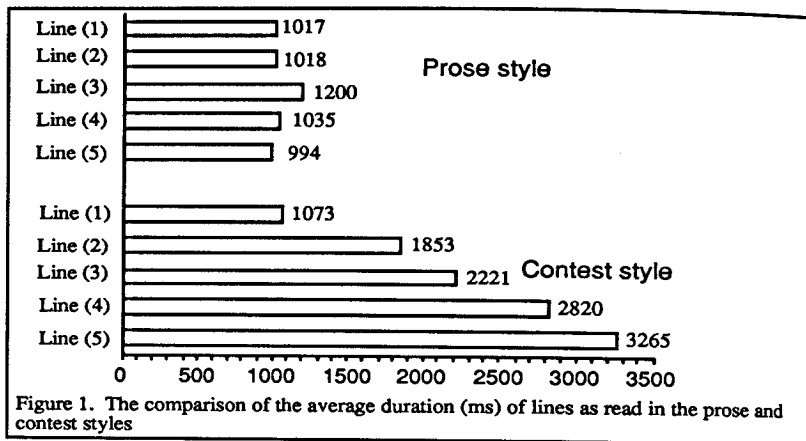


Figure 1. The comparison of the average duration (ms) of lines as read in the prose and contest styles

ms) and standard deviation (SD) of moras, pauses, lines, and whole poems of the fourteen *Tanka* read in the contest style; 9-second pauses after the third line were excluded.

Figure 1 compares the average duration of lines as read in the prose and contest styles.

Table IV. Average duration (ms) of moras of regular and irregular lines as read in the contest style

lines	(1)	(2)	(3)	(4)	(5)
regular	207	265	446	394	470
irregular	216	231	362	348	396
average	208	258	434	388	454

From Table III and Figure 1, the following points were observed.

(1) The speech tempo was much slower in the contest style. The average duration of whole poems excluding 9-second pauses was about twice as long as that of the prose style.

(2) The values of SD were smaller in the prose style. A great regularity existed in production of the prose style.

(3) The average duration of lines in the contest style was not equal; the duration of lines gradually increased. This means that in the contest style the unit of temporal programming was not a line.

(4) There were no pauses after the first, second and fourth lines, but instead a long

9-second pause was put between the groups of 5-7-5 and 7-7 lines. Extremely long prepausal lengthening was observed before pauses. At the end of the lines without pause, however, some amount of preboundary lengthening was observed.

(5) The average duration of moras of 5-mora lines was not longer than of 7-mora lines as in the prose style, although the

average mora duration of regular lines was longer than that of irregular lines. Table IV shows the average duration of moras of regular and irregular lines as read in the contest style.

The two kinds of efforts to keep equidistant lines in the prose style, the adjustment of the duration of pauses and moras, were not effectively made in the contest style.

(6) *Tanka* lines read in the contest style were divided into two groups: 5-7-5 and 7-7, before and after the long pauses. The average duration of the first and the second groups was 5,147 ms and 6,085 ms, respectively. Table V presents the average duration of each group in percentages. The whole duration of the poems excluding the long pause was measured as 100%.

Table V. Average duration of each group in percentages

5-7-5 group	7-7 group
46%	54%

Although the ratio did not show perfect isochrony, there was a strong tendency for the speaker to read the two groups isochronously with prepausal lengthening at the end of each group. It seems to me that the group of lines larger than the poetic line emerged as a unit of temporal programming in the contest style reading of *Tanka*.

4. CONCLUSION

The basic rhythmic unit in Japanese is a mora. [To attest the existence of "a foot consisting of two morae" suggested by Poser (1990)[5] is beyond the scope of this paper.] The mora coexists with larger units such as a word in prose (Homma, 1981[1]; Port et al., 1987[4]), and a poetic line in the prose style (Homma, 1985[2]; Lehiste, 1990[3]). Moreover, a much larger rhythmic unit was found in the contest style reading of *Tanka*. To what extent the temporal programming works in Japanese is yet to be resolved at the moment, but as far as this experiment is concerned, the rhythmic unit was larger than the poetic line in the contest style reading of *Tanka*.

REFERENCES

- [1] HOMMA, Y. (1981), "Durational relationship between Japanese stops and vowels", *Journal of Phonetics* 9, 273-281.
- [2] HOMMA, Y. (1985), "The rhythm of *Tanka*, short Japanese poems", *Acoustic Phonetics in English and Japanese*, Kyoto: Yamaguchi Publishing House, 182-193.
- [3] LEHISTE, I. (1990), "Phonetic investigation of metrical structure in orally produced poetry", *Journal of Phonetics* 18, 123-133.
- [4] PORT, R. F., J. DALBY, and M. O'DELL (1987), "Evidence for mora timing

in Japanese", *Journal of the Acoustical Society of America* 81(5), 1574-1585.

[5] POSER, W. J. (1990), "Evidence for foot structure in Japanese", *Language* 66(1), 78-105.

LATERAL CONSONANT PRODUCTION IN BILINGUAL SPEAKERS LEARNING A THIRD LANGUAGE

J. Llisterra and G. Martínez-Daudén

Departament de Filologia Espanyola - Laboratori de Fonètica
Universitat Autònoma de Barcelona, Spain.

ABSTRACT

The degree of velarization of the lateral consonant [l] is studied in three Catalan-Castilian bilingual subjects reading texts in Catalan, Castilian and French. It is suggested that language dominance for each subject may be related to the use of an alveolar or a velarized variety of this consonant.

1. INTRODUCTION

Although the problem of phonetic transfer has been frequently treated in the second language acquisition literature, transfer phenomena arising at the phonetic level when a bilingual speaker acquires a third language have been hardly approached. Previous work by Llisterra and Poch [4] [5] describes the vowel system of French and English as produced by monolingual Castilian and by bilingual Catalan-Castilian speakers, showing a different pattern of distribution in the F1-F2 plane for each group of subjects. However, the analysis of the long-term averaged spectrum of Catalan-Castilian bilinguals carried out by Bruyninckx et alii [1] reveals the need to introduce the notion of language dominance when bilingual subjects are studied. This paper aims at presenting some preliminary data on the pattern of phonetic transfer found in third language learning by bilingual speakers according to dominance in one of the two languages in which they are proficient.

In order to carry out this investigation, we have chosen to analyze the degree of velarization of the lateral consonant [l] in the French spoken by Catalan-Castilian

bilinguals. In an early work comparing lateral consonants in Castilian and in Catalan, Navarro [7] established the velarized character of Catalan [l] in front of the alveolar point of articulation characteristic of this consonant in Castilian. Further experimental studies have confirmed Navarro's remark by showing the existence in Catalan of the lingual retraction common to the so-called "dark" varieties of [l]. This is one of the more prominent features of the Catalan accent when native speakers of this language speak Castilian.

Since French is a language with a "clear" (i.e. alveolar) lateral consonant, the production of this sound should not be difficult for a monolingual Castilian subject having this variety in his native language. However, one may suppose that some transfer from Catalan may occur in the case of bilingual speakers, and that this will result in the use of a velarized variety of [l] when speaking French. On the other hand, it is also possible to hypothesize that this transfer will be stronger for a bilingual with linguistic dominance towards Catalan than in a speaker with Castilian dominance.

2. PROCEDURE

2.1. Subjects

A large group of students at the Universitat Autònoma de Barcelona were asked to answer a questionnaire aimed at determining their use of Catalan and Castilian and requesting information about their knowledge of French. The questionnaire addresses four skills:

listening, speaking, reading and writing, both in Catalan and Castilian. The subjects had to answer with a percentage of use of each language in a wide range of situations, together with questions about their geographical origin and about the language spoken with each member of the family.

In this paper we will present the data obtained for 3 male bilingual speakers: FC, a bilingual subject who always uses Castilian at home and speaks Catalan around 20% of the time when he is at the University; JJ, also using Castilian as the only language at home, but with a fairly balanced use of Catalan and Castilian (50% each) in other environments; and JT, who uses only Catalan at home and speaks Castilian around 40% of the time when he is at the University and no more than 30% of the time in other situations.

As far as their knowledge of French is concerned, it may be considered quite homogeneous: they have all been studying French at primary and secondary school, with no further contact with the language after this period.

2.2. Corpus

Three texts in Catalan, Castilian and French were selected from contemporary fiction. They contained between 770-550 words and a large number of tokens of the lateral consonant [l]. For the purpose of the present study, only sequences of vowel + [l] + [e] have been retained; this offers a suitable environment for inter-language comparisons, since [Vle] sequences are found in the three languages. A total number of 296 lateral consonants have been measured (104 for Catalan, 93 for Castilian and 99 for French).

2.3. Recording and acoustic analysis

The speakers read the texts at their normal rate in one single recording session. It took place in a sound isolated room in semi-anechoic conditions at the Phonetics Laboratory at the Universitat Autònoma de Barcelona. A Sennheiser MD 441N directional cardioid microphone and a Revox A77 tape recorder were used.

The signal was digitized at 10 KHz sampling rate, and a 13 pole LPC model was used in the acoustic analysis; a pre-emphasis of 6dB was also applied. This was done using the routines implemented in the MacSpeech Lab II software package by GW Instruments running on an Apple Macintosh SE/30.

The boundaries of the lateral consonant were determined using changes in the waveform as the main criteria; perceptual checking was also applied listening to the segmented consonant when necessary.

Lingual retraction can be related to the frequency of the second formant; a low F2 corresponds to a high degree of tongue retraction and to a raising of the posterior part of the tongue towards the velum. Then, F2 frequency was measured at the center of the lateral consonant in order to assess whether our speakers were using a "clear" or a "dark" variety.

3. RESULTS

Results for each language are shown in tables 1, 2 and 3 below.

Table 1: F2 mean values for Catalan [l] for each bilingual subject

Subject	Catalan F2 (Hz)		
	mean	s.d.	n.
FC	1338	96.2	34
JJ	1269	68.6	35
JT	1198	84	35

As far as Catalan is concerned, second formant frequency values for [l] are significantly different for the three speakers (t-test $p = 0.000$ between FC and JT and $p = 0.001$ between FC and JJ and JT).

Table 2: F2 mean values for Castilian [l] for each bilingual subject

Subject	Castil.		
	F2 (Hz)		
	mean	s.d.	n.
FC	1473	105	33
JJ	1466	82.1	27
JT	1301	111.5	33

However, for Castilian values, significant differences (t - test, $p = 0.000$) are only found between JT and the two other subjects.

Table 3: F2 mean values for French [l] for each bilingual subject

Subject	French		
	F2 (Hz)		
	mean	s.d.	n
FC	1607	80.1	31
JJ	1575	84	31
JT	1319	93.8	37

French values for the lateral consonant exhibit highly significant differences between JT and the other two subjects (t test, $p = 0.000$); nevertheless, differences between FC and JJ for French are more significant than those found when comparing their values for Castilian (t obs = 0.3, d.f. = 58, $p = 0.803$ for Castilian and t obs = 1.5, d.f. = 60, $p = 0.140$ for French).

On the other hand, intra-speaker comparisons between the three languages show very significant differences in all cases but one: differences between French and Castilian values for speaker JT are not significant (t obs = 1.1, d.f. = 23, $p = 0.288$).

4. DISCUSSION

In order to assess the linguistic profile of our subjects, the values obtained from the analysis of their productions of the lateral consonant may be compared to reference values for monolingual speakers.

Quilis et alii. [9] obtained an average value of 1534 Hz for F2 in intervocalic context for monolingual Castilian speakers; neither of our subjects approaches this value, and significant

differences between this mean value and all the values obtained are found (p between 0.000 and 0.002 for the three subjects). This fact may be explained by the influence of Catalan in the Castilian production of our bilingual subjects, as shown in previous research by Martínez-Daudén [6].

Catalan [l] is given a mean value of 1039 Hz by Panyella [8]. Observed Student's t values are greater when comparing results found for speakers FC and JJ (t obs = 18.1 and 18.3, d.f. = 33 and 35, $p = 0.000$) than those obtained when comparing this population mean with F2 values for JT (t obs = 11.1, d.f. 34, $p = 0.000$). It is clear then that values found for JT are more similar to those reported for dominant Catalan speakers than those observed for the two other subjects.

For native French, Chafcouloff [2] quotes values between 1461 Hz and 1849 Hz, depending on the vowel context; average value is of 1656 Hz; the only subject approaching this value is FC, but the difference between the population mean for native French and the sample mean produced by FC is still significant (t-test $p = 0.002$). This shows that none of our speakers are producing a French native lateral consonant, at least as far as F2 is concerned. This result is in agreement with the analysis of a different group of bilingual speakers previously carried out by the authors [3], although differences appear more clearly in the present study.

However, there is a strong difference between FC and JJ on the one hand and JT on the other. As will be seen later, this has to do with their linguistic dominance.

The values observed so far are coherent with the results of the initial test dividing the speakers in terms of their linguistic dominance: FC exhibits an F2 frequency for the lateral consonant in Castilian which is intermediate between the native Castilian values and the ones found for dominant Catalan speakers when speaking Castilian; on the other hand, they are significantly different from his values for Catalan, which are themselves higher than the F2 values found for Catalan dominant speakers; he may be then considered a bilingual with dominance towards

Castilian. JJ shows a very similar pattern for his Castilian productions, but his values for Catalan are intermediate between those found in Catalan dominant speakers when using Castilian and the values encountered in the Castilian dominant subject. He may then be classified as a more balanced bilingual. Finally, JT shows Castilian values which are different from his Catalan ones, but which are also significantly different from the values found for native Castilian speakers; this corresponds to the pattern usually found in bilingual speakers showing a strong Catalan dominance.

The next question one may ask is how linguistic dominance is related to transfer phenomena in the French texts produced by our subjects. A general trend in the differences between subjects can be observed: a high F2 in French seems to be correlated with a high degree of Castilian dominance. Anyway, the fact that the Castilian values for the speaker FC - a Castilian dominant subject - and also for JJ - a balanced bilingual - do not attain the F2 frequency of a monolingual Castilian has not to be disregarded. A comparison with the performance of monolingual Castilian subjects speaking French would be needed to clarify whether knowledge of Catalan in exerting some influence in the spoken French of these individuals.

A strong case for transfer from Castilian could be made observing the lack of significant differences between the mean for French and the mean for Castilian in speaker JT ($p = 0.288$); however, one has to take into account the fact that this speaker may be trying to approach a target value for F2 higher than the one found in Catalan - his dominant language -, attaining equal unsuccessful results in Castilian and in French; this second hypothesis will favour the idea of a transfer from Catalan.

5. CONCLUSION

It has been shown that the degree of velarization of the lateral consonant in Catalan-Castilian bilingual speakers using French is influenced by their linguistic dominance. The study of phonetic transfer in third language learning reveals then that bilingualism is a notion that embodies

different degrees of language competence for different speakers. However, more research is still needed in order to formulate an adequate model of transfer in third language learning by bilinguals.

6. REFERENCES

- [1] BRUYNINCKX, M. et alii. (1991), " Effects of language change on voice quality. An experimental study of Catalan-Castilian bilinguals ", *12th International Congress of Phonetic Sciences*, 19-24 August 1991, Aix -en- Provence, France.
- [2] CHAFCOULOFF, M. (1980), " Les caractéristiques acoustiques de /j,y,w,l,r/ en français ", *Travaux de l'Institut de Phonétique d'Aix*, 7, 5-56.
- [3] LLISTERRI, J.- MARTINEZ DAUDEN, G. (1990), " Phonetic interference in bilingual speakers learning a third language: the production of lateral consonants ", *Proceedings of AILA-90*, Thessaloniki, Greece, vol 2, 452.
- [4] LLISTERRI, J.- POCH, D. (1987), " Les réalisations sonores dans l'apprentissage d'une troisième langue par bilingues ", *Revue de Phonétique Appliquée*, 82-83-84, 209-219.
- [5] LLISTERRI, J.- POCH, D. (1987), " Phonetic interference in bilingual's learning of a third language ", *Proceedings XIth ICPhS. The Eleventh International Congress of Phonetic Sciences*. Estonia: Academy of Sciences of the Estonian SSR, Vol.5, 134-137.
- [6] MARTINEZ-DAUDEN, G. (1989), " caracterización acústica de la consonante lateral alveolar en un corpus de habla continua realizado por hablantes bilingües ", *Revista Española de Lingüística*, 19/1, 160.
- [7] NAVARRO TOMAS, T. (1917), " Sobre la articulación de la /l/ castellana " in BARNILS, P. (Ed), *Estudis Fonètics*, Barcelona, Institut d'Estudis Catalans, 265-275.
- [8] PANYELLA, R. (1985), *Determinació experimental del context de velarització de [l] del català a partir de la freqüència dels tres formants acústics*, Unpublished ms., Laboratori de Fonètica, Universitat Autònoma de Barcelona.
- [9] QUILIS, A. et alii (1979), " Características acústicas de las consonantes laterales españolas ", *Lingüística Española Actual*, 1/2, 233-343.

Pour que l'enseignement des langues étrangères
donne les moyens de les comprendre à l'étranger

Regina LLORCA

Université du Queensland, Australie

ABSTRACT. This paper shows that methods of learning a language do not give the ability to deal with it in a country where it is spoken because they only develop phonological memory. Suggestions are proposed for exercises to complement methods which are based on the process of acquisition but apply it to the class context and do not give general skills.

1. L'ÉCOUTE CONDITIONNÉE

Pour l'enseignement traditionnel des langues étrangères, la perception de la parole est d'abord la reconnaissance des éléments appartenant à la langue et le message est considéré indépendamment du *sujet parlant* :

- l'élocution privilégiée est celle qui efface les caractéristiques individuelles au profit du message transmis : elle est claire, neutre et sans accent ;

- les exercices favorisés sont les exercices de reproduction : l'étudiant doit substituer sa voix à celle d'une personne entendue pour reconstruire le *texte* dit par celle-ci, intégralement ou dans son contenu global.

Les conséquences sont les suivantes :

- le type d'élocution privilégié en classe *conditionne* l'écoute de l'étudiant. Lorsqu'il se rend à l'étranger, l'ex-apprenant compare les réalisations qu'il entend au modèle de référence, c'est-à-dire à sa propre prononciation : il ne perçoit et ne comprend strictement rien dès qu'elles s'en éloignent sensiblement (le locuteur a un accent, crie,

marmonne, parle très vite, etc.).

- dans les exercices pratiqués en classe, l'étudiant développe un certain mode de mémorisation de la parole : il ne mémorise pas les sons qu'il a entendus mais les mots *qu'il a reconnus à travers les sons entendus* et pour lesquels il donne mentalement une nouvelle réalisation, marquée du mode de prononciation qu'il a appris, comme il le fait oralement dans les exercices de reproduction. La mémoire auditive est réduite à la capacité de retenir l'information recueillie par l'écoute *phonologique* mais ne s'appuie pas sur la *mémoire phonétique* ou *musicale* qui enregistre les sons tels qu'ils sont produits par le sujet parlant dans une séquence donnée. Quand une séquence de parole a été enregistrée par la mémoire musicale, on peut la *réentendre* mentalement (comme on peut réentendre par exemple une séquence célèbre de cinéma avec la voix et l'accent de l'acteur qui l'a dite).

En fait, l'apprentissage traditionnel traite la parole comme on traite l'écrit. En présentant une référence de prononciation, il donne à l'étudiant les moyens de *représenter* les sons qu'il entend. L'étudiant peut alors mémoriser cette représentation au lieu des sons entendus et donc réduire la parole au message exprimé, sans la voix du sujet parlant. Ainsi, le souvenir qui reste d'une production orale est comparable à celui qui reste d'un texte lu.

2. L'ÉCOUTE INNOCENTE

Examinons maintenant le mode d'écoute et

de mémorisation qui est celui de "l'étranger innocent" qui acquiert la langue dans un pays qui la parle et qui n'en a aucune connaissance préalable.

Étant incapable de comparer les réalisations diverses des natifs à un modèle de prononciation, l'étranger innocent ne peut que s'efforcer de mémoriser les *sons* qu'il entend ; la seule stratégie qu'il peut adopter est d'accumuler en mémoire des faits sonores marqués d'une voix et d'une façon de parler individuelle. Cette mémorisation s'effectue d'abord pour des séquences entendues des dizaines de fois et qui deviennent de ce fait *musicalemment intelligibles* : elles cessent d'être perçues comme des amalgames confus de sons inarticulés pour être perçues comme des suites intonées de syllabes.

Dans chaque séquence, l'étranger innocent peut alors comparer les sons *entre eux* et les différencier dans le cadre de la suite de sons produits par *le* sujet parlant dans *cette* production.

En comparant des séquences de parole produites par des locuteurs différents, l'étranger innocent peut déterminer les traits sonores qui sont communs à des groupes géographiques ou sociaux. En comparant des séquences produites par un même locuteur dans des situations différentes, il peut déterminer les transformations que les sons subissent : lorsque le locuteur s'adresse à des personnes de statut différent ou lorsqu'il change d'attitude.

Après détermination des traits communs aux sons produits dans un pays ou une région et des règles qui régissent leurs transformations, toute séquence de parole devient intelligible.

Dans une première étape, l'étranger innocent acquiert les éléments qu'il entend le plus souvent ; à partir du stade d'intelligibilité générale, tout élément est susceptible d'être acquis puisqu'il apparaît sous une forme sonore *perceptible et mémorisable*.

En résumé, l'écoute innocente consiste à percevoir chaque séquence de parole comme un fait sonore nouveau et unique, même si

elle contient des mots déjà connus, parce qu'elle est produite par un être humain qui ne ressemble à aucun autre et dont l'humeur change d'un moment à l'autre. Cette attitude d'écoute permet la compréhension de tout individu parlant dans une communauté car elle conduit à établir les rapports structurels, non pas entre les unités abstraites de la langue (phonèmes ou prosodèmes) mais entre *les sons concrets réalisés par ceux qui le parlent* et ceci grâce à la mémoire musicale qui établit des comparaisons entre les éléments successifs d'une même chaîne sonore et entre des façons de parler différentes.

Ainsi, l'auditeur innocent acquiert la langue en même temps qu'il élabore les systèmes de variantes de réalisation : variantes expressives, stylistiques, sociales, géographiques. C'est d'ailleurs ce qui lui permet d'acquérir les éléments de la langue avec leur *valeur culturelle* : on sait qu'un mot appartient à un registre de langue familier ou soutenu parce qu'on l'a mémorisé avec une façon de parler relâchée ou soignée ; on sait qu'un mot appartient au jargon politique parce qu'on l'a mémorisé avec la voix des leaders politiques ; en France, on sait que "peuchère" est un mot du Midi parce qu'on l'a mémorisé avec l'accent du Midi, et aucun autre.

Pour l'auditeur innocent, la parole n'est pas d'abord emploi de la langue, elle est en premier lieu le produit d'une activité vocale et par là, l'expression d'une *identité* (physique, sociale, géographique). La séquence de parole est alors une suite de sons particuliers à un individu avant d'être une suite d'éléments phonologiques communs au groupe culturel et la perception de la parole est un acte de *connaissance* des sons particuliers avant d'être un acte de *reconnaissance* des éléments communs. Cet ordre se justifie ainsi : l'intelligibilité, qui est le résultat de l'acte de connaissance des sons particuliers, permet leur mémorisation musicale et c'est par comparaison entre les sons mémorisés que l'on découvre leur valeur fonctionnelle puisque cette valeur

n'existe que par opposition à celle des autres. L'auditeur innocent s'intéresse donc aux différences *phonétiquement* ou *musicalement* pertinentes avant de les réduire aux différences phonologiques : par exemple, la différence entre [R] et [r] en français ou entre les réalisations du phonème /p/ par un Alsacien et un Parisien n'est pas phonologiquement pertinente mais elle est phonétiquement pertinente puisqu'un Français perçoit les deux sons comme différents. Les traits phonétiques sont pertinents non pas pour la signification mais pour l'identification du sujet parlant. Or, on ne peut pas comprendre le message si l'on est incapable de comprendre la structure du système de sons dans lequel il est transmis. C'est ce qui arrive à l'étranger conditionné qui cherche immédiatement les mots dans les paroles qu'il entend, sans s'intéresser au sujet parlant, souhaitant même que celui-ci n'existe pas car il "déforme" les sons. En l'empêchant de comprendre, ledit sujet parlant ne fait que lui retourner l'insulte.

L'auditeur innocent s'efforce de percevoir les sons réels au point de les mémoriser musicalement avant de les interpréter linguistiquement. Les situations où l'on peut deviner la signification des formes sonores étant rares, l'étranger innocent détermine la valeur de la plupart des expressions qu'il acquiert en comparant les situations mémorisées où elles ont été entendues. Lorsqu'il commence à comprendre, il mémorise à court terme des séquences de sons plus ou moins longues avant de les décoder en mots (décodage qui se fait en même temps que la séquence suivante est enregistrée dans le cas de la parole suivie). A mesure que sa connaissance de la langue s'élargit, l'auditeur développe des capacités d'anticipation et reconnaît les mots l'un après l'autre mais il continuera d'appliquer sa première stratégie pour comprendre la parole à débit très rapide et les séquences totalement imprévisibles qui existent dans la plupart des conversations. Cette démarche s'oppose aux pédagogies qui recommandent de deviner les éléments manquants d'après

un certain nombre de "mots-clés" (sans qu'on sache comment ces mots-clés sont identifiés et compris) et qui développent chez l'étudiant l'habitude d'entendre en fonction d'attentes, du fait qu'elles favorisent les discours prévisibles et construits sur les acquis antérieurs.

Enfin, l'étranger innocent construit sa prononciation sur celle des natifs qu'il réentend parler en mémoire alors que l'étranger conditionné entretient constamment l'articulation qu'il a adoptée tout au début de l'apprentissage.

Bien sûr, les avantages de l'écoute innocente ne sont donnés qu'à ceux qui ont suffisamment de contacts avec la population du pays.

Cependant, il apparaît que la possibilité de traiter une langue à l'étranger ne requiert pas seulement des *conditions* favorables ; elle requiert aussi une *aptitude* à mémoriser la parole avec la voix du locuteur et la façon exacte dont ont été prononcés les sons dans une situation donnée. Il s'agit de mémoriser la parole par la mémoire *sensorielle* des sons, comme on mémorise les bruits, avant ou en même temps que par la mémoire *intellectuelle* du sens et des mots reconnus.

3. APPLICATIONS

Les nouvelles pédagogies des langues créent des conditions d'*acquisition*. En particulier, elles retardent la phase où les étudiants parlent eux-mêmes [1], [2], ce qui est un facteur déterminant pour qu'ils mémorisent ce qu'ils entendent et non une représentation marquée de leur voix. Cependant, il ne s'agit de conditions d'acquisition et de compréhension que *dans la classe*. Pour préparer les étudiants aux conditions de l'étranger, il faut aussi leur donner les moyens de traitement de la parole en général et dans le cadre de l'enseignement à des adultes, cet objectif signifie le plus souvent *redécouvrir* l'écoute innocente et *rééduquer* la mémoire musicale. En effet, les stratégies décrites plus haut sont reconstituées à partir de l'observation d'enfants à l'étranger : les enfants utilisent la mémoire sensorielle de la

parole dans une langue étrangère parce qu'ils l'entraînent constamment dans leur propre langue ; on diminue cet entraînement à mesure que l'on développe la maîtrise de l'écriture, en particulier lorsqu'on accède au stade de la prise de notes qui développe un type de mémorisation où les propos entendus sont réduits à leur contenu lexical. C'est ce modèle que reproduit le processus d'*apprentissage* d'une langue : il consiste à prendre des *notes mentales*, au moyen des représentations phoniques que l'on apprend à articuler au départ.

Nous proposons ici un ensemble d'exercices destinés à développer la mémoire musicale d'une langue étrangère et à favoriser un mode d'acquisition où la découverte des invariants phonologiques et de la signification se fait par comparaisons entre les données de cette mémoire. Ces exercices exploitent les résultats d'une recherche sur les facteurs favorisant la mémoire sensorielle du son [4] et sont appliqués au Français Langue Etrangère ; certains ont été expérimentés à l'Université du Queensland, grâce à la collaboration des enseignants et étudiants du Département de Français et sur l'initiative de Dr. Jacques Montredon, responsable du programme de première année.

L'approche propose d'abord une activité appelée "Théâtre Rythmique". Le professeur interprète des discours marqués d'effets expressifs et phonostylistiques tout en faisant effectuer par les étudiants une suite de mouvements construite en combinant les trois critères suivants : les mouvements ont une valeur illustrative du contenu du discours, visualisent ses qualités prosodiques et s'enchaînent de façon à ce que chacun amène le suivant. Il s'agit de proposer une chaîne kinétique qui "déclenche" le souvenir des sons entendus sur chacune de ses portions, d'autant plus que son exécution par les étudiants est le prétexte à ce qu'ils entendent la chaîne sonore plusieurs dizaines de fois. Le Théâtre Rythmique s'apparente avec la méthode Total Physical Response [1] par sa Fonction,

avec certaines approches musicalistes [5] et avec les langages de signes par ses formes. Dans une phase ultérieure, les étudiants interprètent les discours eux-mêmes (comme dans la démonstration vidéo présentée, où le Théâtre Rythmique est appliqué à l'interprétation de poèmes modernes choisis par Jacques Montredon).

Parallèlement, on travaille sur le cinéma et les feuilletons télévisés, qui présentent la plus grande variété de réalisations. Les exercices proposés conduisent les étudiants à analyser la structure phonique de séquences diverses, accédant ainsi à leur intelligibilité et à leur mémorisation musicale. Ici, il s'agit de développer une technique générale de traitement des variations individuelles et régionales.

Dans les deux types d'exercices, la phase de mémorisation est suivie d'une phase de comparaisons entre les données mémorisées, où les étudiants sont amenés à découvrir les mots communs apparaissant sous diverses réalisations sonores. Dans le Théâtre Rythmique, cette découverte se fait par l'intermédiaire des éléments gestuels récurrents, dans les documents filmés, elle se fait par les éléments situationnels communs aux scènes comparées et que le professeur doit choisir en conséquence.

RÉFÉRENCES

- [1] ASHER, J. (1972), "Children first language as a model for second language learning", *Modern Language Journal*, 56, 133-139.
- [2] KRASHEN, S. et al. (1984), "A theoretical basis for teaching the receptive skills" *Foreign Language Annals*, 17, n° 4, 261-271.
- [3] LLORCA, R. (1991), "Le rôle de la mémoire musicale dans la perception d'une langue étrangère", *A paraître*.
- [4] LLORCA, R. (1991), "Facteurs de développement de la mémoire musicale d'une langue étrangère", *A paraître*.
- [5] MACARTHUR, S. et TROJER, L. (1985), "Learning language through music", *Revue de Phonétique Appliquée*, 73-74-75, 211-222.

A COMPARISON OF ENGLISH AND GREEK
ALVEOLAR FRICATIVES

L. Panagopoulos

Aristotle University of Thessaloniki
Greece

ABSTRACT

This paper investigates a production problem faced by Greek speakers of English involving the English alveolar fricatives. It seeks acoustic evidence supporting a physiological hypothesis. The hypothesis is confirmed.

1. INTRODUCTION

In Panagopoulos (1985) the problem was examined by reference to physiology and was attributed to a wider 'swing' of the tongue for the Greek alveolars. Greek speakers, like all users of a foreign language, go through training for the acquisition of English by trying to extract the properties of the new phonological system through the knowledge of their own native system. The Greek inventory of consonants includes an alveolar pair, [s, z], but lacks a palato-alveolar opposition. This lack of opposition allows for a wider 'swing' of the tongue in Greek, which presumably includes the articulatory area allotted to the English palato-alveolar fricatives. As a result, the pronunciation of an English [s] by a Greek

speaker, often sounds unacceptably retracted, verging on a [ʃ].

The aim here is to provide acoustic evidence for or against the physiological hypothesis made in Panagopoulos 1985.

2. METHOD

Two native speakers of English and Greek produced spectrograms on a Kay Sonagraph. Frequency, amplitude and duration measurements were made on appropriate parameter configurations, often with the help of a calibrated transparent overlay, and the results were normalized and processed statistically. The alveolars were embedded in phrases in intervocalic positions with a stressed vowel preceding them and an unstressed vowel following them.

3. RESULTS AND DISCUSSION

3.1. Spectral peaks and amplitude

The spectral distribution of energy for the alveolars, bands of fricative noise as well as voicing for the voiced segments are as follows:

E n g l i s h :

[s]: 3.8 - 8kHz

[z]: 3.6 - 8kHz

[ʃ]: 2.1 - 7kHz

Greek:

[s]: 3.0 - 8kHz

[z]: 3.0 - 8kHz

The lower bottom range for the Greek [s] became even lower, 2.6kHz, when the alveolar was followed by a voiceless velar plosive. This difference is attributed to less amplitude of noise rather than to the transitions (Harris 1958) which are less prominent features than the centre frequencies. In creating synthetic stimuli the transitions were omitted in May (1976). The same source reported that noise frequencies for (English) [s], varied with vocal tract size, male vs female, so that the lower energy limit was raised for a female speaker, from 3.5kHz and 1.6 for [s] and [ʃ] respectively (Hughes and Halle 1956) to 5.1kHz and 2.6 for the same fricatives. Such sex-specific differences can be safely assumed to be universal though and any amount of shifting applying to English should apply to Greek under the same circumstances. Amplitude, as expected, was higher for the voiceless set than for the voiced set, in both languages by about 8dB. The spectral energy of the Greek voiceless alveolar fricative lies between the English alveolar and palato-alveolar. The English values are close to those by Strevens (1960) and Behren & Blumstein (1988). As far as the Greek alveolar articulation goes

it is definitely retracted by comparison to the English [s].

3.2. Durational measurements

The distribution of duration for the alveolars in the two languages was:

English:

	[s]	[z]	[ʃ]
mean:	170 ms	54 ms	172 ms
s.d.	3.5	4	3.6
variance:	12	16	10
range:	7	8	8

Greek:

	[s]	[z]
mean:	73 ms	61 ms
s.d.:	8.5	9.8
variance:	16	17

On the basis of these data, the duration of the English [s] was almost three times as long as that of the Greek [s]. Together with reduced amplitude (see above), the reduced duration of the Greek fortis articulation was much less energetic than the corresponding English articulation and reduced energy supports retraction. In addition, the higher statistic figures for standard deviation, variance and range associated with the Greek alveolars underline their instability, which in our opinion is due to the lack of a palato-alveolar contrast in Greek. An interesting indication is also the fact that the voiced fricatives in the two languages, apart from the higher instability for the Greek [z], are more similar than the voiceless fricatives. This instability of the Greek [z] is reflected in the remarkably higher percentage of devoicing for this

sound than for the English [z] (Panagopoulos 1975), a fact that can be attributed to higher intraoral pressure.

4. CONCLUSION

The acoustic data support the physiological hypothesis mentioned in the beginning of this presentation.

Although there are differences in the acoustic analyses of the sets of the alveolar fricatives in English and Greek, the common articulatory parameters they share place them in the same phonological class, which incidentally consists of the same number of oppositions in the two languages (with the addition of the single voiceless glottal fricative in English) because the lack of a palato-alveolar pair in Greek is balanced by the inclusion of a velar pair of fricatives.

The measurements of the spectrograms revealed stable patterns of spectral activity for the English fricatives. This stability is probably due to the restricted freedom of movement of the tongue which is a consequence of language-specific phonological organization.

5. REFERENCES

- [1] BEHRENS, S. J. & S. E. BLUMSTEIN, (1988), "Acoustic Characteristics of English voiceless fricatives: a descriptive analysis", Journal of Phonetics, 16, 295-298.
[2] HARRIS, K. S. (1958), "Cues

for the discrimination of American English fricatives in spoken syllables", Language and Speech, 1, 1-7.

[3] HUGHES, G. W. & M. HALLE, (1956), "Spectral properties of fricative consonants", Journal of the Acoustical Society of America, 28, 303-310.

[4] MAY, J. (1976), "Vocal Tract Normalization for /s/ and /ʃ/", Haskins Laboratories Status Report on Speech Research, SR-48, 67-73.

[5] PANAGOPOULOS, E. (1975), "The Place of the Phonetic Component in Linguistic Theory", Faculty of Philosophy Yearbook, 14, Aristotle University of Thessaloniki.

[6] PANAGOPOULOS, E. (1985), "Tongue Constraints Affecting the Acquisition of L1/L2", Proceedings of the International Conference on Foreign Language Learning and Interpersonal Tolerance and Understanding, Thessaloniki.

[7] STREVENS, P. (1960), "Spectra of fricative noise in human speech", Language and Speech, 3, 32-49.

LA VALEUR ET LES FONCTIONS DE LA PAUSE DANS LE PARLER ORAL MONOLOGIQUE SPONTANÉ DANS LA LANGUE ÉTRANGÈRE ET DANS LA LANGUE MATERNELLE

M.AVDONINA

Université Linguistique de Moscou, URSS

ABSTRACT

The present study concerns pause's status as semiotic sign and experimental analysis of its functions. We chose for material oral spontaneous monological speech in native and foreign languages. The comparison lead to principal conclusions: 1) pause's functions in foreign language have qualitative difference with the native language ones; 2) quantitative individual differences correlated in our experiment with progresses in speech in foreign language

1. INTRODUCTION

Les approches à l'étude de la pause sont nombreux et varient selon les buts de recherche. Dans le cadre des études que nous poursuivons depuis plusieurs années et qui portent sur les particularités du parler en langue étrangère il s'est manifesté un problème, assez peu étudié, de la caractéristique comparative de la valeur et des fonctions de la pause dans le parler oral monologique spontané dans la langue étrangère et dans la langue maternelle.

Nous partons du fait que la pause présente d'une part un élément de l'acte de la parole et, d'autre part, un signe sémiotique de nature particulière (signe zéro) qui se prête difficilement à l'analyse comme à l'interprétation. Nous la considérons comme un des universaux de la langue, surtout dans sa fonction du démembrement de la parole. L'interprétation de la pause peut dépendre des facteurs de nature différente, contextuels et contextuels. C'est une des unités qui, dépourvues de valeur nominative, assurent la communication et qui, étant facultatives, sont marquées par l'individualité du sujet parlant. Nous proposons de faire en-

trer la pause dans une catégorie pragmatique particulière, ensemble avec d'autres caractéristiques individuelles de la parole, telles que répétitions mécaniques, corrections de sens, corrections de forme, mots et propositions intercalés. Étant des signes sémiotiques de nature complexe, ces composantes de l'acte de la parole participent non seulement au démembrement syntagmatique du texte, mais à la communication proprement dite, influant à l'expressivité et au caractère performatif du texte.

Cette étude a pour but d'établir les corrélations entre: 1) la valeur et l'emploi de la pause; 2) certaines qualités de la personnalité du sujet parlant (autoévaluation, réflexion) et 3) ses progrès en langue étrangère.

2. PROCEDURE.

2.1. ENREGISTREMENT DES TEXTES.

Nous avons proposé à 40 étudiants en français de l'Université Linguistique de Moscou (étape avancée: 3^e année d'étude) de développer un des trois sujets à leur choix en limitant le temps à 15 min. Après la fin de l'enregistrement les étudiants écoutaient la cassette et répondaient aux questions (préalablement élaborées à la base des conversations avec des professeurs et des étudiants) qui portaient sur leurs impressions du texte et leur réflexion pendant la production du texte. Les étudiants répondaient en langue maternelle sur la même cassette.

2.2. ANALYSE QUANTITATIVE DES TEXTES.

Les textes en français étaient soumis à l'analyse psycholinguistique élaborée par nous qui comprenait une quarantaine de pa-

ramètres ayant pour but de relever les particularités du parler en langue étrangère. Les textes en russe (les comptes-rendus) étaient analysés après, chaque fois que la comparaison s'imposait. Cette analyse comprend seulement les pauses qu'on pouvait mesurer (plus de 3 secondes).

2.3. EVALUATION DES PROGRES EN FRANCAIS.

Cinq professeurs ont écouté les enregistrements et appréciaient le parler des étudiants (anonymement) sur 7 caractéristiques chacun selon l'échelle de trois niveaux: "bon" (100 points), "moyen" (50 points), "mauvais" (0). La moyenne arithmétique de ces 35 notes a acquiert la valeur d'évaluation des experts. A partir de cette moyenne nous avons subdivisé les étudiants en quatre groupes: 1) 0-24 points: les "mauvais", 6 pers.; 2) 25-49 points: les "mauvais-moyens", 7 pers.; 3) 50-74 points: les "bon-moyens", 17 pers.; 4) 75-100 points: les "bons", 10 pers.

2.4. ANALYSE QUALITATIVE DES FONCTIONS DE LA PAUSE.

Cette fois nous analysons toutes les pauses, y compris les pauses audibles, mais trop courtes pour être mesurées. En nous basant sur les réponses des comptes-rendus nous tâchons de relever les fonctions des pauses en les liant aux particularités de la réflexion et de l'autoévaluation de chaque groupe d'étudiants.

2.5. Toutes les informations ont été soumises au traitement par ordinateur pour confirmer la fiabilité des corrélations obtenues.

3. RESULTATS.

3.1. LES FONCTIONS DE LA PAUSE DANS LA LANGUE ETRANGERE ET DANS LA LANGUE MATERNELLE.

L'analyse des 40 protocoles du parler oral monologique spontané en russe et en français permet de constater l'absence dans le parler en français - langue étrangère d'éléments qu'on trouve sans peine dans le parler en russe de ces mêmes étudiants: composantes allusives, humour, implications, procédés rhétoriques, valeurs autonomes de l'intonation, etc. décrits dans nos articles - ce qui amène à l'absence des écarts entre les aspects sémantique et pragmatique du texte, à la pénurie de moyens d'expressivité, à la réduction de la valeur performative du texte français.

L'interprétation de la pause peut dépendre des facteurs de nature différente, y compris ceux extralinguistiques: les connais-

sance de fond de l'auditeur, la situation de communication, l'attitude du destinataire à l'égard du sujet parlant et de la charge informative du texte.

La pause d'un monologue spontané dans une communication normale dans la langue maternelle peut servir 1) pour faire comprendre ce qui est dit (l'interlocuteur a le temps pour coder l'information dans son code individuel de la parole intérieure); 2) pour fasciner, pour préparer l'auditeur à la perception d'une information importante qui va suivre; 3) pour provoquer la réaction de l'auditeur (approbation, rire, etc.); 4) pour adhésion; 5) parcellisation; 6) pour répondre à sa propre question implicite.

Nous avons trouvé des exemples de tous ces types de pauses dans les comptes-rendus autoévaluatifs de nos étudiants (faits en russe).

Entre les pauses d'hésitation nous avons pu distinguer deux types: 1) certains étudiants ne croient pas possible de se faire des louanges (le parler spontané en russe contenait un compte-rendu, l'impression de l'étudiant après avoir écouté le texte enregistré de son allocution en français); 2) d'autres ne peuvent pas exprimer une impression négative sur eux-mêmes et sur leur parler en français. Les premières ont à l'origine des tabous sociaux et moraux, les deuxièmes sont dues aux particularités de l'autoévaluation et de la réflexion de l'individu.

Dans les textes en français toutes les pauses portent le caractère d'hésitation. En employant la méthode du compte-rendu des sujets parlants et l'analyse du texte, nous avons pu y distinguer certains sous-types: 1) pause pronostique, prospective: a) due aux difficultés formelles, b) due à celles du contenu; 2) pause retrospective: a) de correction, surtout grammaticale; b) de recherche des bifurcations du contenu dans un fragment distant; c) de réaction émotionnelle, affective sur sa propre activité langagière.

La pause comme un des éléments pragmatiques est directement liée à la transformation de la parole intérieure à la parole extérieure, en assurant et en exprimant explicitement le dialogue du sujet parlant avec lui-même (qu'est-ce que je vais dire et comment? qu'est-ce que j'ai dit et comment? est-ce que je parle bien? m'écoute-t-on attentivement? que pense de moi mon interlocuteur?, etc.). Nous avons supposé que

les particularités de la réflexion de chaque groupe d'étudiants peuvent changer son rôle dans la structure du texte.

3.2. LE RÔLE DE LA PAUSE DANS LA STRUCTURE DU TEXTE EN LANGUE ÉTRANGÈRE.

Chaque énoncé en français a été traité avec une procédure spéciale de comptage des éléments du texte. Le tableau 1 reproduit une partie de ces résultats, qui, d'une part, donnent l'idée générale de la longueur du texte, de la durée de la production et de la vitesse de la parole et, d'autre part, démontrent les différences quantitatives d'emploi de la pause et d'autres composantes privées de valeur nominative entre les quatre groupes d'étudiants (voir 2.3.)

Le tableau 1 fait voir que l'évaluation subjective des experts est confirmée par les différences réelles des caractéristiques des textes: la longueur, la durée et la vitesse augmentent du groupe 1 au groupe 4. Les deux groupes des faibles emploient moins de mots et phrases intercalés (le groupe 1 les réduit presque à zéro). On peut supposer que leurs fonctions sont réalisées dans ce cas par les répétitions mécaniques et les

pauses qui sont sensiblement plus nombreuses. En plus, remarquons l'absence presque totale des corrections de sens dans le groupe 1.

D'après leurs comptes-rendus les étudiants du groupe 1 au moment de production de la parole pensaient qu'à leur échec, les pauses nombreuses étaient donc compensées par des réflexions sur leur personnalité, le produit de leur activité n'y était pas reflété. Les étudiants du groupe 2 concevaient leur échec eux-aussi, mais, parallèlement, se cherchaient des excuses dans les conditions de l'expérience, ils faisaient non seulement des pauses nombreuses, mais ils étaient sûrs de bien parler (très peu de répétitions mécaniques et de corrections de forme); 3) le groupe 3 se caractérise par une approche formelle à l'exécution de ce travail, ce qui amenait aux difficultés de développer l'idée (un très grand nombre de répétitions mécaniques); 4) enfin, les bons étudiants du groupe 4 pensaient surtout à la qualité de l'exécution, à la qualité de leur produit (texte), cela amenait à un bon contrôle et règlement de leur production: très peu de pauses, mais le nombre sensiblement plus grand des mots intercalés et de correc-

Tableau 1

CARACTÉRISTIQUES PSYCHOLINGUISTIQUES DES TEXTES EN FRANÇAIS (LANGUE ÉTRANGÈRE)

Caractéristiques des textes	Groupes d'étudiants			
	Groupe 1 les "mauvais"	Groupe 2 les "mauvais- moyens"	Groupe 3 les bons- moyens"	Groupe 4 les "bons" 25%
1. Longueur du texte (nombre de mots)	294	493	864	1 014
2. Durée de la production (min)	4,1	5,7	9,6	7,6
3. Vitesse (mots/min)	71	89	116	133
4. Pauses (sec.)	13,0	14,3	3,2	1,7
5. Répétitions mécaniques (% de longueur)	4,4	1,9	4,5	2,4
6. Corrections de sens (% de longueur)	0,2	1,4	1,3	1,6
7. Corrections de forme (% de longueur)	1,2	1,1	1,3	0,7
8. Mots intercalés (% de longueur)	3,6	3,2	4,0	4,1

tions de sens.

A notre avis cela permet aux bons étudiants, d'une part, de concevoir les idées et de choisir la forme sans s'arrêter et en gardant toujours le rythme de la parole et, d'autre part, de vérifier la construction formulée où ils ont senti une difficulté.

L'analyse du texte prouve que la valeur et les fonctions de la pause dans la langue étrangère sont apauvries par l'absence de

l'intention de l'emploi de la pause dans un but autonome, tandis que les mêmes sujets parlants l'emploient dans la langue maternelle avec toutes sortes d'effets pragmatiques. En même temps chaque groupe d'étudiants en fonction de leurs progrès en langue étrangère a des particularités qualitatives et quantitatives de l'emploi de la pause dues au caractère de leur réflexion et de leur autoévaluation.

LEARNING ENGLISH SEGMENTS
WITH TWO LANGUAGES

Mohamed Benrabah

Institut des Langues Etrangères
Université d'Oran, Algérie

ABSTRACT

To discover the extent to which the bilingual's both native languages influence the sound system of English we have made a phonetic analysis of the errors made by Algerians at the segmental level. It appears that for Algerians, the native language with the most complex sub-system will have a major influence on the corresponding sub-system of the target language.

1. INTRODUCTION

The phonological system of an average adult speaker is so firmly rooted that any attempt to alter it may encounter resistance. In the production of the sounds of another language, this resistance is expressed as phonic interference which mainly consists of the transfer of some phonological habits of the first language(s) into the one being learnt.

Is the expression of this phonic interference the same for the monolingual as well as the bilingual? Evidence exists to support the influence of the native language of the monolingual on the target one [4], while studies concerned with what happens in the case of the bilingual do not abound.

The present paper is an attempt to shed some light on the influence of the bilingual's two languages on the sound system of the target language. In dealing with the errors made by bilinguals at the level of segmentals, we will try and show the extent to which each native language influences these segments.

2. METHOD

Twenty four balanced bilinguals in Algerian Arabic and French were recorded speaking English spontaneously. These advanced speakers had just graduated or were about to. Conversational-like speech has the advantage of revealing the influence of the native language(s) because of the strains that result from the conditions due to conversation [2]. The material recorded for each speaker was edited and administered later as a dictation task to ten native British listeners. Thus, a total of 240 listeners took part in the various listening sessions during which they wrote down as accurately as possible what they had heard. These British informants were selected on the grounds of their performance with an R.P speaker using a similar

but shortened version of the material used with the Algerians.

3. ANALYSIS OF THE RESULTS

In the final analysis of the listeners' responses, we considered only instances where miscomprehension occurred between each Algerian speaker and the ten native listeners.

The actual utterances which led to miscommunication were analysed phonetically and the causes classified along a small number of error-categories.

Segmental substitution proved the most important in the distortion of the speech of Algerians. Vowels accounted for the majority of the mispronunciations. The following pure vowels (in rank order) were the most commonly mispronounced segments: / \wedge ɒ ə e/

Note that all these vocalic sounds are lax and this tends to support the claim that R.P lax vowels are the most difficult sounds for the non-native to make [1].

When we consider the various misarticulations of the above vowels, we notice that Algerians tend to replace them with a certain regularity.

For the vowel / \wedge /, speakers produce either of the following:

- quality in the general region of secondary cardinal vowel No.11; e.g. "studies", "sumumer", "suddun".

- quality in the general region of primary cardinal vowel No.6; e.g. "young", "oother", "monoth".

From the above data we can state that whenever 'u' occurs in the spelling

speakers would usually produce [œ]. Orthographic 'o', on the other hand, incites subjects to produce [ɔ]. These mispronunciations are the result of a negative transfer from French as well as spelling.

R.P /ɒ/ is regularly realised as [əʊ] and occurs when speakers try to give an English 'flavour' to certain lexical items which show in their spelling certain similarities with French. Words of this kind contain an /ɒ/-type of sound spelt as 'au' or 'o' (as in e.g. "Maurice", "mosque" and "cost"). These inaccurate qualities may stem from an initial awareness of the presence of diphthongs in English and their absence in French. The realisation of words like "gone" and "knowledge" with a diphthong represents a typical error due to overgeneralisation; that is, since both "go" and "know" contain the diphthong in question, the same vocalic glide is maintained in "gone" and "knowledge". The different incorrect realisations of the schwa vowel by Algerians do not show any general tendency. However, the majority of the words involved were structural items pronounced in their strong form based on spelling and without any vowel weakening.

R.P /e/ tends to be realised as a vowel in the general region of cardinal No.2 in items such as "better", "embassy" and "definite". This could result from a confusion with the French vowel of "été" ('summer') which is negatively transferred into English.

The mispronunciation of some R.P diphthongs is also a common feature of the spoken English of Algerians. Out of eight English diphthongs, /eə əʊ eɪ/ and to a lesser extent /ɪə/ were the most difficult.

/eə/ is commonly realised as a monophthong with a quality in the general region of cardinal vowel No.2.

Among the usual inaccurate articulations of /əʊ/ we can mention a monophthong which fluctuates between cardinal vowels No.6 and No.7, especially when the sound is spelt 'o'.

As to the mispronunciation of /eɪ/ monophthongization also tends to be the rule. A typical realisation varies between cardinal vowels No.2 and No.3.

The substitution of [i:] for /ɪə/ in words such as "here" and "really" could be related to the presence of 'e' in the spelling.

In the production of R.P consonants, six segments proved most problematic with the plosives /t/ and /d/ being among the most difficult. These are given a dental articulation as in French and Arabic, although an affricated alveolar [tʃ] exists in a number of Algerian Arabic varieties. The misarticulation of R.P /t/ and /d/ also lacks aspiration which is neither present in Arabic nor in French.

One of the striking features of the spoken English of Algerians is the substitution of /t/ and /d/ by the corresponding emphatic plosives. These typical realisations occur in the context where /d/,

and particularly /t/, are followed by an open and/or back vowel as in e.g. "time" and "talk".

Three R.P fricatives proved particularly difficult for Algerians. Subjects usually replace dental /θ/ by [t] or [f] but never by [s] as in the case of French native speakers. As for the lenis dental /ð/, it is almost always realised as dental [d] and never as [z].

The mispronunciation of /h/ fluctuates between the lenis glottal fricative [ɦ] and the fortis laryngeal [h] with the former being the most widespread.

Finally, the approximant /r/ is typically realised as an alveolar tap in the spoken English of Algerians.

4. DISCUSSION

The above analysis of the various sound substitutions allows us to make a number of observations. First, in the case of mispronounced vocalic segments (pure and diphthongal) the negative transfer seems to result from the French language. The produced vowel either exists in French or is derived from the influence of spelling based on the Algerians' knowledge of the sound/letter correspondence in this same language.

Second, the inaccurate articulation of certain R.P consonants seems to find an answer in the effect that Arabic consonants have on the Algerian speaker. The non-use of [s] and [z] for dental fricatives, the use of emphatics and the tap are evidence to support this point.

It thus appears that

interference does not occur haphazardly but seems to express itself in a specific way. That is to say, most of the vocalic errors and consonantal errors were attributed to the influence of French and Algerian Arabic respectively. Compared with Arabic, French has a more complex vowel system, whereas Arabic has a much more complex consonant system. It appears, then, that for the Algerian bilingual, the native language with the most complex sub-system will have a major influence on the corresponding sub-system of the target language. Hence, because the French vowel system is more complex than that of Algerian Arabic, speakers are more likely to be influenced by French in their mispronunciation of English vowels. On the contrary, the Arabic consonant system being more complex, it is more likely to affect the articulation of certain English consonants.

5. CONCLUSION

In conclusion we will exercise a word of caution. Our attempt to explain certain phonetic errors made by Algerians does not take into account all the processes involved in second language learning. In the present paper we dealt with only one such process which is language transfer from the two native languages. But we do realise that language transfer on its own is not enough to explain the various deviations from the norm made by the non-native in the process of learning a

foreign language [3].

6. REFERENCES

- [1] GIMSON, A.C. (1980), *An Introduction to the Pronunciation of English*, (3rd ed.), London: Edward Arnold.
- [2] RITCHIE, W.C. (1968), "On the explanation of phonic interference", *Language Learning*, 18, 183-197.
- [3] SELINKER, L. (1972), "Interlanguage", *IRAL*, 10, 219-231.
- [4] TARONE, E. (1978) "The phonology of interlanguage", in Richards, J. (ed.), *Understanding Second and Foreign Language Teaching*, Mass.: Newbury House.

PROBLEMATIQUE DE L'ENSEIGNEMENT DE LA PROSODIE DU RUSSE AUX FRANCOPHONES

M. BILLIERES

Laboratoire Jacques-Lordat, Université de Toulouse II, France

ABSTRACT

The errors of pronunciation made by French speakers in Russian are caused by an inadequate temporal perception of the phonetic particularities of the language, by the phonetic correction method employed and by the choices made by the learner.

Les erreurs de prononciation des Français en russe sont dues entre autres à une mauvaise perception temporelle de sa matière phonique. Le temps n'est pas une donnée physique : c'est une notion, un concept, une donnée abstraite. Le développement de la notion de temps est lui-même une forme d'adaptation à la réalité.

Un rythme langagier dépend de la périodicité subordonnée à la temporalité. La périodicité n'est elle même possible que si elle est structurée. Une langue étrangère, c'est d'abord un rythme étranger. Il est constitué par des contrastes perçus en termes de durée, de hauteur et d'intensité sur les syllabes successives d'une suite sonore. En français /F/ comme en russe /R/ il existe une dizaine de rythmes "de base" où les phénomènes accentuels ont un rôle clé dans la perception de la périodicité en tant que prééminences rythmiques constituant autant de points de repère possibles.

Dans les 2 langues où ils reviennent à intervalles réguliers, toutes les 4 à 5 syllabes en moyenne, il existe :

- un accent de groupe : en /F/ il a la durée pour paramètre premier, en /R/ l'intensité est le trait physique dominant ;

- un accent secondaire : en /F/ c'est un relief mélodique et/ou d'intensité en /R/, il dépend de la durée, toujours inférieure à celle de la voyelle accentuée ou de la prétonique mais toujours supérieure à celle des autres voyelles atones. La voyelle porteuse de cet accent n'est jamais perçue comme étant réduite.

L'intonation se déroule sur la trame rythmico-temporelle. En /F/ les unités intonatives sont caractérisées par l'égalité de durée des syllabes non accentuées et l'allongement notable (x 2 en gros) de la voyelle affectée par l'accent de groupe. Il en va tout autrement avec les "constructions intonatives" (IK) du russe. Ainsi, IK-4 provoque l'allongement de la voyelle accentuée et de la dernière syllabe atone du syntagme. Dans certaines réalisations modales, IK-5 et IK-7, IK-2 et IK-5 se distinguent par des différences de durée et de débit. En syntagme non terminal, IK-2 a un tempo plus rapide que IK-1 dans des phrases affirmatives.

Au niveau des voyelles et des consonnes, la durée est fonction de la longueur intrinsèque des sons, de leur combinaison dans la syllabe, de la nature de cette dernière, du débit, etc. En /R/ la quantité vocalique dépend en outre de la dureté/mouillure des consonnes avoisinantes et de l'influence de l'accent lexical.

L'accent lexical, neutralisé en /F par l'accent de groupe, affecté au contraire en /R/ chaque mot significatif à l'intérieur du syntagme. Son trait dominant est la durée, soumise au phénomène de la réduction vocalique : habituellement, la durée de la voyelle accentuée est supérieure à celle de toutes les autres voyelles du mot, la longueur des voyelles préaccentuées étant elle-même plus importante que celle des voyelles postaccentuées.

En /R/, le rythme est constitué par l'inégalité quantitative des syllabes internes d'un syntagme, contrairement au /F/. La réduction vocalique constitue un obstacle perceptif majeur car certaines voyelles ont des durées très brèves - entre 40 et 50 ms- auxquelles l'oreille française n'est pas accoutumée. Un handicap supplémentaire est dû à l'accent lexical que le Français tend à ramener vers la fin du mot. Cette destruction rythmique déclenche un dérèglement en chaîne de la structure phonique au niveau de l'intonation et à celui de la réalisation des phonèmes. Une source supplémentaire de difficultés est liée à la perception globale de la hauteur, de l'intensité et de la durée qui caractérisent tous les types d'accent. Les rapports entre ces 3 paramètres posent encore de nombreuses énigmes. Le Français a tendance à accentuer les mots du /R/ en privilégiant la hauteur et l'intensité, ces 2 corrélats étant intimement liés. En /R/, l'intensité et la durée caractérisent l'accent de groupe, et la durée l'accent lexical. Ce dernier toutefois a été pendant très longtemps qualifié d'accent d'intensité. Or, en règle générale, une augmentation d'intensité peut être perçue comme une augmentation de durée. De même, une élévation de la hauteur peut être ressentie comme un renforcement de l'intensité. Quoiqu'il en soit, la durée est toujours le paramètre le plus fragile et le plus difficilement décelable. En fait l'oreille est un instrument de mesure acoustique bien imparfait.

La méthodologie soviétique de correction phonétique se fonde sur la conscientisation. La comparaison des "bases articulatoires" du /R/ et du /F/ s'effectue par le biais de cours de phonétique articulatoire privilégiant le visuel -schémas, miroir devant la bouche, lecture- et le tactilo-kinesthésique grâce à de nombreux exercices de "gymnastique articulatoire". La prononciation étant directement reliée au contrôle neuro-musculaire, un bon entraînement sensori-moteur joint à la connaissance précise de l'articulation d'un son donné constituent un préalable au développement de la "base perceptive".

Cette approche intellectualisante, axée sur le sens et l'écrit, négligeant la perception auditive, perturbe la composante prosodique et entrave la mémoire à court terme /MCT/. La MCT, dite aussi "mémoire de travail" est extrêmement volatile : les informations qu'elle capte s'effacent au bout de quelques sec. Elle est très rapidement saturée, ne pouvant contenir plus de 7 à 8 items. Plus spontanée et phonétique que la mémoire à long terme, ses performances sont améliorées par les répétitions et freinées par les interférences. Son rôle est capital en situation de production et de réception d'un énoncé.

En production, l'élève doit écouter et répéter en ayant souvent le texte sous les yeux. Le processus de lecture s'accomplit ainsi : l'oeil est fixe pour déchiffrer les mots, puis se déplace, s'arrête à nouveau, repart etc. Pour un lecteur normal, le temps de pause de l'oeil est de 1/4 de sec, le temps de déplacement de 1/40 de sec.. Le cerveau opérant une reconnaissance globale du mot et non un découpage lettre par lettre pour l'appréhender, anticipe le déchiffrement de l'oeil. L'élève français débutant ou faux débutant est souvent un lecteur lent en /R/ en raison de :

1) l'hésitation provoquant un contrôle oculaire inadéquat se traduisant par des régressions et des rectifications

rendant la lecture pénible ; 2) la subvocalisation, incitant au mot à mot, entravant la compréhension globale et nuisant à l'anticipation. La lecture laborieuse est fondée sur la saisie d'une dizaine de lettres au maximum -MCT- et aboutit au pire à du mot à mot. Ce risque est aggravé par l'élaboration et la présentation des exercices dans les manuels. Ils vont du simple au complexe : listes de mots isolés (parfois de syllabes logatomiques), de 2 mots (nom + verbe, adj. + nom, etc.), de phrases mono puis bi-syntagmatiques... Les composantes de ces listes sont séparées par des virgules, des tirets, voire des points de suspension.

Tout ceci incite le Français à déplacer l'accent en fin de mot et à le placer souvent sur sa dernière syllabe. Les gammes d'exercices (1 mot, 2 mots...) mettent en vedette l'accent lexical au détriment des accents secondaire et de groupe. L'élève est également obnubilé par le poids sémantique de la désinence qu'il souligne par l'intensité et l'élévation du ton ; il est impatient d'atteindre la fin du mot où une pause virtuelle est possible ; enfin, il obéit à la règle accentuelle "à droite toute" du /F/.

En réception, la fonction d'écoute suppose l'interaction de 2 systèmes fonctionnant simultanément. Le 1er extrait du continuum sonore des éléments servant à élaborer des hypothèses ; le 2ème vérifie ces hypothèses perceptives et/ou linguistiques en comparant continuellement les éléments perçus et les éléments attendus. La compréhension résulte de la concordance entre les différentes hypothèses. Les enregistrements des manuels proposent un rythme aux environs de 3 syllabes/sec. (soit entre 10 et 12 phonèmes), ralenti par des pauses délimitant les groupes de mots. Ce rythme assez lent facilite théoriquement l'écoute mais peut provoquer à terme la monotonie et donc un relâchement de l'attention. Il y a de plus, hiatus complet avec les actuelles méthodes de russe où le débit des enregistrements est généralement rapide.

Toutes ces données affectent la MCT. Elle ne peut conserver un son au-delà de 2 sec., après il se volatilise. Elle intègre également les pauses sur le temps temporel ; il en va de même pour les blancs et les signes de ponctuation en lecture. En outre, l'oreille perçoit globalement et non son par son, de même que l'œil ne lit pas lettre par lettre. L'élève habitué à entendre par les yeux est soumis à des interférences visuelles et motrices. La subvocalisation provoque des phénomènes de coarticulation sur la base de la langue maternelle et engendre des articulations tendues anihiliant le rythme du /R/. Une MCT non entraînée à la perception des sonorités du /R/ et reposant essentiellement sur la lecture ne peut accroître ses facultés de discrimination auditive et entretient la surdité prosodico-phonologique de beaucoup d'élèves.

Ceci est flagrant lorsque l'élève devant répéter, par ex. , "moloko" prononce 3 [o] ou produit un groupe sifflante + chuintante au lieu de [Z':] dans "prieZZaet" : il lit le mot dans sa tête avant de le prononcer. Ceci nous amène à nous interroger sur les stratégies qu'il met en œuvre pendant son apprentissage du /R/. Nous n'entrerons pas dans les classifications certainement réductionnistes et probablement abusives distinguant des apprenants "globalistes" ou "sérialistes" ou considérant des élèves à dominante visuelle, auditive ou kinesthésique. En fait, l'élève doit traiter simultanément diverses composantes de la langue en interrelation étroite ; ses capacités d'analyse n'y suffisant pas, il effectue forcément des choix, focalise son attention sur tel ou tel aspect de la structure du signal. Au fur et à mesure de l'apprentissage, il automatise certains processus qui au départ étaient contrôlés, plus particulièrement ceux qui se rapportent aux aspects formels de la langue -prosodique, phonologique, morphologique, syntaxique...-.

Par contre, il ne peut automatiser certains processus plus complexes qui varient sans cesse -lexical, sémantique-. La déstructuration prosodico-phonologique évoquée en supra est une conséquence directe de cette gestion : elle constitue pour l'élève une sorte de régularité rentable (économique) lui permettant de se concentrer sur les opérations de haut niveau au détriment de celles de bas niveau. Un enseignement de type "grammaire-traduction" fondé sur l'écrit ne peut que l'encourager dans cette voie. Il en va de même avec le recours à une méthode de correction phonétique très analytique qui ne favorise pas une approche globalisante, structurante et dynamique de la matière phonique de la langue.

[6] KELLER, E. (1985), "Introduction aux systèmes psycholinguistiques" Chicoutimi, Québec, Canada : Gaëtan Morin ed.

[7] MOISEEV, A.I. (1975), "Russkij jazyk. Fonetika, Morfologija, Orfografija" Moskva : Prosveščenie

[8] MUXANOV, I.L. (1983), "Posobie po intonacii", Moskva : Russkij jazyk

[9] "Sovremennij russkij jazyk. Teoretičeskij kurs. Fonetika" (1985), pod red. V.V.IVANOVA, Moskva : Russkij jazyk

[10] "Urovni jazyka v rečevoj dejatel'nosti" (1986), pod red. L.V. BONDARKO, Leningrad : Izd-vo Leningr. un-ta.

[1] BILLIERES, M. (1989), "Les apports de la phonétique expérimentale à l'enseignement de la prononciation du russe", Actes du Vème colloque de Linguistique russe, Poitiers, 13-16 mai 1987, 249-260.

[2] BILLIERES, M. (1989), "La réduction vocalique en russe moderne", Mélanges de Phonétique générale et expérimentale. Publications de l'Institut de Phonétique de Strasbourg, 63-78.

[3] BILLIERES M. (1989), "Non verbal, phonétique corrective et didactique des langues", Revue de Phonétique Appliquée, n° 90, 1-16

[4] FRAISSE, P. (1979), "Des différents modes d'adaptation au temps". Du temps biologique au temps physique. Symposium de l'Association de psychologie scientifique de langue française, 1977, Paris : PUF, 9-20.

[5] GAONAC'H, D. (1987), "Théories d'apprentissage et acquisition d'une langue étrangère", Paris : Hatier-Crédif

Contrastive Phonetics and Teaching Foreign
Language Pronunciation: Theory and Practice

Mukhamedjan K. Isaev

Teachers Training Institute of Foreign
Languages, Alma-Ata, USSR

ABSTRACT

Foreign language teaching especially in classroom conditions can be effective if its methodology is based on solid scientific grounds. Practical needs of teaching languages have given rise to theoretical contrastive study of languages that come into contact. Experimental investigation of the development of phonetic characteristics of bilinguals in L_2 makes a significant contribution into the theory of contrastive studies as well as into the applied aspects of this problem.

It is evident that contrastive studies of languages came into being and developed in many countries of the world as a reaction to practical needs of teaching foreign languages (L_2). Contrastive analysis of two languages with the aim to predict pronunciation errors in L_2 are defined as Applied Contrastive Studies. Theoretical Contrastive Studies involve various methods and models of contrastive analysis and is a part of typological linguistics.

The aims of these two studies are so tightly

interpenetrated with each other and are so interdependent that it is difficult to find the borderline between them. Any contrastive study is supposed to be theoretical. Its theoretical status is high provided the results of such studies are of some practical use. At the same time the discussion and linguistic analysis of pronunciation errors of bilinguals in L_2 belong to the sphere of theoretical linguistics. So, we shall use the term "Contrastive Phonetics" (Linguistics) only. The data obtained form the foundation of the theory of interference. The latter is capable of explaining the nature of each error. In this paper an attempt is made to suggest the principles and models of the wider application of contrastive studies in linguistics.

The necessity to work out the principles of contrastive studies relevant both for theoretical and particularly for practical purposes of teaching L_2 made us carry out multi-step experiment. The procedure of the experiment is the following. It consisted of five steps and lasted for five years. The subjects ($n=45$) were

the students of English at the Teachers Training Institute of Foreign Languages in Alma-Ata, USSR. Their native language is Kazakh. When the experiment began (step 1) the subjects were at the age of 17. Recording of the test material took place in a special studio at the end of the academic year during five years. Step 1 corresponds to the end of the first year of learning English; Step 2-to the second year; Step 3-to the third year; Step 4-to the fourth year; Step 5-to the fifth year.

The test experimental material was compiled on the basis of the data of the contrastive study of the phonetic systems of English and Kazakh. It consisted of English isolated words, sentences of different communicative types, texts of various functional styles.

The staff of subjects and test material remained unchanged throughout all the five steps of the experiment. In order to obtain the most natural and objective data no correction of the subjects' errors in the process of recording was allowed. The same test experimental material was recorded by native speakers of English (n=4).

The recordings of all subjects of each step were listened and transcribed by non-native speakers of English-teachers of English Phonetics (n=5).

Slightly shortened version of the test material was listened and analysed by native speakers of English-non-linguists (n=10) and, finally, rather short test recording was listened and analysed by a na-

tive speaker of English-professor of English Phonetics.

For the acoustical experiment still shorter test material was selected. The experiment meant to study such prosodic properties as duration, intensity, fundamental frequency. A special study of spectral characteristics of English vowels was carried out.

Procedure and Results of Contrastive Phonetic Study of the Development of the English Speech of Kazakh-English bilinguals

1. Contrastive study of the phonetic systems of English and Kazakh, the languages that are in contact in classroom conditions.

The results of the contrastive study are used in two ways: a) practical-the list of potential errors is the starting point for compiling teaching material for the incipient bilinguals, in our case, for the students of the first year; b) theoretical-the list of potential errors is the starting point for compiling test language material for the experimental investigation of the development of phonetics of the bilinguals' English speech.

2. Auding Analysis of the recordings of each step of the experiment helps us to obtain the data of real (actual) pronunciation errors (or real interference). Contrastive study of the data of each step and the data of the potential errors are valid in two ways: a) practical-the results are taken into account when compiling teaching material for the second, third, fourth, fifth year students; b) theoretical-linguistic

analysis of the data obtained makes a considerable contribution into the theory of interference. Phonetic errors are classified and tabulated according to the types of interference mistakes (U. Weinreich); underdifferentiation, overdifferentiation, reinterpretation, substitution, plus segmentation, minus-segmentation. A certain amount of pronunciation errors are beyond these types, forming specific groups of errors. Types of errors by U. Weinreich are rather easily classified in Steps 1, 2, 3. At the advanced steps no definite types are to be found.

3. Contrastive study of the data of auditory analysis of all five steps of the experiment and the data of auditory analysis of native speakers' recordings revealed the points of coincidence and the points which to this or that extent, differentiate the pronunciation of the subjects from that of the native speakers.

a) Practical application of the results of this study is in supplying the teachers with recommendations on how to prevent pronunciation errors of bilinguals depending on the stage of acquiring English.

b) Theoretical-contrastive description of bilinguals' errors and the model pronunciation of the native speakers contributed into typology of errors, problem of the development of phonetic interference (segmental and prosodic) throughout the five steps of the experiment.

4. Contrastive study of the data of the two neigh-

bouring steps of the experiment which made it possible to form a general view of the phonetic characteristics of the English speech of the bilinguals in the process of acquiring L₂.

6. Contrastive study of the data of the acoustic experiment and the data of the auditory experiment.

7. Contrastive phonetic study ended with the evaluation of the recordings of each step by the native speakers of English (non-linguists and linguists).

CONCLUSIONS

The procedure of contrastive study presented in this paper is worth accepting because:

1. Thorough contrastive study of two languages in contact is necessary for linguistics in general, and in particular-it supplies the teaching process with the proper language material specially processed and selected for this very stage of teaching L₂, for this very group of L₂ learners in classroom conditions. The data obtained stimulate preparation of textbooks, manuals, tapes.

2. The term "contrastive phonetics" may have a wider sphere of application. The principles of contrastive analysis presented here enrich and develop the theory of language contact, the theory of interference, contrastive linguistics and typological studies.

3. The procedure of contrastive analysis presented in this paper can be used and development further on the material of other languages in contact.

LES MODELES RYTHMIQUES ET LA FREQUENCE
DES SUBSTANTIFS AVEC L'ACCENT MOBILE EN RUSSE

E. Jasová

Cours du soir pour adultes, Finlande

ABSTRACT

This paper presents the results of our research of Russian nouns with a mobile accent on the basis of Frequency Dictionary [1]. From the point of view of a declension of nouns and in accordance with their frequency we specify 15 types and 10 subtypes of rhythmical patterns. In the declension we determine the following rhythmical patterns: 1) fundamental, 2) reduced and 3) extended. In the same type or subtype can be situated: a) an isosyllabic and b) polysyllabic rhythmical patterns. The isosyllabic rhythmical patterns may exist with an identical or an inidentical accent, but a difference in the grammar.

1. INTRODUCTION

La place de l'accent du mot russe n'est lié ni à une syllabe, ni à un morphème déterminés. Il est libre. Cette propriété rend difficile l'enseignement de la langue russe comme langue étrangère. La recherche de la place de l'accent a deux différents principaux points de vue. L'un examine l'accent sur les morphèmes l'autre sur les syllabes.

Du point de vue syntagmatique, horizontal, en extension, "début - milieu - fin" du mot, on examine la distribution de l'accent à des places différentes seulement au nominatif des substantifs. Du point de vue paradigmatique l'accent russe peut être constant et mobile.

Constant: il se trouve sur la même syllabe, ou sur le même morphème à toute la déclinaison du mot. Ex.: po-gô-da ≠ pogôd/a (le temps) po-gô-dy ≠ pogôd/y (du temps), po-gô-de ≠ pogôd/e (au temps) etc. Mobile: il peut se déplacer; avancer (progresser), ou reculer (regresser) d'une syllabe ou d'un morphème. Ex.: Sg. Nom. go-râ ≠ gor/â (le mont), Sg.Ac. gô-ru ≠ gôr/u (le mont), Pl.Instr. go-râ-mi ≠ gor/âmi (avec, par les monts) etc.

La syllabe est l'unité de la parole et le morphème est l'unité de la langue. Ils sont étroitement unis et, à notre avis, il faut les étudier ensemble.

2. PROCEDURE

Nous avons examiné les 3 935 substantifs du Dictionnaire Fréquent [1] des deux points de vue: syntagmatique et paradigmatique. Sur la base de la distribution de l'accent aux formes fondamentales (au Nom. Sg.) nous déterminons les types des modèles rythmiques dans les classes des substantifs: I - monosyllabiques, II - dissyllabiques. Le premier modèle est avec l'accent sur la première syllabe du mot et l'autre avec l'accent sur la deuxième syllabe. III - trisyllabiques etc. Sous la notion de modèle rythmique nous comprenons le nombre des syllabes du substantif et l'accent sur la syllabe déterminée [2,3]. Nous indiquons l'accent des substantifs d'après le dictionnaire-manuel [4].

Tableau 1

Classes des subst.	Types de modèles ryth.	Nombre	%	Fréquence	%
I	—	411	10,45	42 512	16,92
II	— —	712	18,09	51 270	20,41
	— —	683	17,36	49 961	19,89
III	— — —	202	5,13	12 561	5,00
	— — —	552	14,03	30 054	11,97
	— — —	317	8,06	18 820	7,49
IV	— — — —	12	0,31	521	0,21
	— — — —	251	6,38	14 721	5,86
	— — — —	200	5,08	6 667	2,66
	— — — —	57	1,45	2 339	0,93

Le tableau 1 montre que la plupart des substantifs russes sont dissyllabiques et trisyllabiques. En général entre les deux types de modèles rythmiques il y a très peu de différence en nombre et en fréquence. Ex.: — — = sf-la (la force - 20,41%) - — = na-rôd (le peuple - 19,89%). Mais on observe

des différences remarquables entre les genres masculin, féminin et neutre. Entre les trois types de modèles rythmiques des substantifs trisyllabiques (voir tableau 1, III) le type — — — = pa-gô-da (le temps) monte très nettement au premier plan (11,97%). Mais il y a beaucoup de différences entre les genres.

Tableau 2

Substantifs	Nombre	%	Fréquence	%
Déclinés	3 917	99,54	250 648	99,77
Avec l'accent constant	3 245	82,46	169 352	67,41
Avec l'accent mobile	672	17,08	81 296	32,36
Indéclinés	18	0,46	591	0,23
Au total	3 935	100,00	251 239	100,00

Le tableau 2 montre que la plupart des substantifs russes ont l'accent constant dans la déclinaison. Mais les substantifs avec l'accent mobile ont une très haute fréquence relative. Le nombre de substantifs avec l'accent mobile est approximativement cinq fois inférieur en nombre de substantifs avec l'accent constant, mais leur fréquence relative l'est seulement deux fois 32,36% (voir tableau 2).

Ainsi nous distinguons les modèles rythmiques des substantifs d'après leur fréquence: 1. fondamental, sur la base de la supériorité numérique au paradigme. Par ex.: Pour toutes les formes des 7 cas il existe un modèle rythmique — — —, qui se compose d'une 1ère syllabe atone et d'une 2ème syllabe accentuée (v. figu.1). 2. réductible, sur la base de la disparition du nombre des syllabes au paradigme. Ex.: Gen. Pl.

rūk = (des mains) se compose d'une syllabe accentuée et d'une syllabe annulée.

3. extensible, sur la base de l'augmentation du nombre des syllabes au paradigme. Ex.: Instr. Pl. ru-kā-mi = - - - (avec, par les mains), se compose de trois syllabes, une première syllabe atone, une deuxième syllabe accentuée et une troisième syllabe atone, ajoutée.

Ensuite nous remarquons qu'il existe pour le même paradigme des modèles rythmiques: a) isosyllabiques et b) polysyllabiques. Le même type ou le sous-type peut présenter des modèles rythmiques isosyllabiques avec des accents identiques ou non identiques, mais

une différence grammaticale. Ex.: Nom. Sg. ru-kā (la main), Gen. Sg. ru-kī (de la main) etc. (7 cas) présente le modèle rythmique dissyllabique - - avec l'accent identique, mais une différence grammaticale (voir figu.1). Mais par ex. le Gen. Sg. ru-kī (de la main) - - et le Nom. Pl. rū-ki (les mains) - - présentent des modèles rythmiques dissyllabiques avec l'accent non identique. Le déplacement d'accent à l'Ac. Sg. et aux Nom. et Ac. Pl. est régressif [5]. Les modèles rythmiques polysyllabiques sont caractéristiques de la déclinaison flexible en russe (voir figu. 1).

Segmentation des substantifs

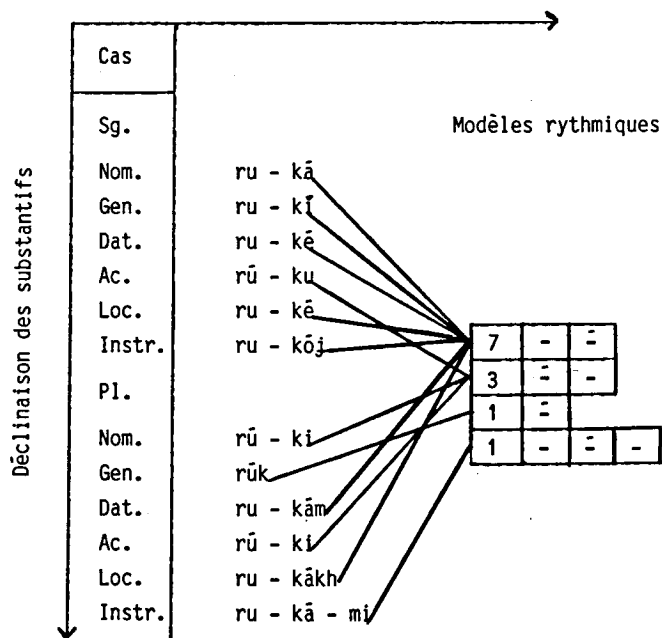


Figure 1 Type de modèle rythmique F3

Le modèle rythmique fondamental - - présente les formes des 7 cas Nom. Sg. rukā (la main), Gen. Sg. rukī (de la main), Dat. Sg. ruke (à la main), Loc. Sg. ruke (sur, dans la main), Instr. Sg. rukōj

(avec, par la main) et aussi Dat. Pl. rukām (aux mains), Loc. Pl. rūkākħ (sur, dans les mains): Le type de modèle rythmique F3 présente seulement 8 substantifs féminins: rukā (la main), porā

(le temps), nogā (le pied), stenā (le mur), rekā (la rivière), gorā (le mont), sredā (le milieu), mais leur fréquence est très haute.

3. CONCLUSION

Dans l'ensemble des substantifs avec l'accent mobile (672) nous limitons les types et sous-types

Tableau 3

Substantifs dissyllabique avec l'accent mobile								
Modèle rythmique - -					Modèle rythmique - -			
Gen.	Nombre		Fréquence		Nombre		Fréquence	
	abs.	%	abs.	%	abs.	%	abs.	%
Masc.	63	57,80	6 855	35,18	134	57,27	7 594	32,31
Fem.	18	16,51	1 625	8,32	75	32,05	12 231	52,04
Neutr.	24	22,02	9 103	46,74	25	10,68	3 679	15,65
Pl. t.	4	3,67	1 902	9,76	-	-	-	-
Tot.	109	100,00	19 477	100,00	234	100,00	23 504	100,00

Nous limitons cinq types et quatre sous-types de substantifs monosyllabiques. En premier lieu en nombre et en fréquence il y a le type M₂ - -, qui présente les substantifs masculins: nōs (le nez) avec l'opposition d'accent Sg. ≠ Pl.

Nous limitons dix types et six sous-types de substantifs dissyllabiques. En premier lieu en fréquence (voir tableau 3) il y a le type F₄ - -, qui présente les substantifs féminins: rukā (la main), bien que le nombre soit tout petit (8). En général, les types de modèles rythmiques des substantifs masculins les plus fréquents sont ceux avec l'opposition d'accent. Au singulier l'accent se trouve sur la première syllabe, mais au pluriel sur la terminaison. Les types de modèles rythmiques des substantifs féminins les plus fréquents sont ceux avec l'accent régressif à l'accusatif Sg. et au Nom., au Gen. et à l'Ac. Pl. Les types de modèles rythmiques des substantifs neutres les plus fréquents

de modèles rythmiques d'après leur fréquence.

Notre recherche montre que l'accent mobile en russe est caractéristique avant tout des substantifs monosyllabiques et dissyllabiques.

sont aussi les types avec l'opposition d'accent.

A notre avis cet index de fréquence est très important pour l'enseignement de l'accent mobile en russe à aide de modèles rythmiques.

4. REFERENCES

- [1] ZASORINA, L. N. (1977) "Cha-stotnyj slovar ruskogo jazyka", Moskva.
- [2] JASOVA, E. (1984) "Die syllabisch-Akzentologischen Modelle der russischen Substantive", Proceedings of the Tenth International Congress of phonetic Sciences, Holland, 696-699.
- [3] ZLATOUSTOVA, L. V. (1975) "Rhythmic Structure Types in Russian Speech", Auditory analysis and perception of speech, London, New York, San Francisco, 477-484.
- [4] AVANESOV, R. I., OZHEGOV, S. I. (1959) "Russkoje literaturnoje proiznosnenije i udarenije", Slovar-spravochnik, Moskva.
- [5] STRAKOVÁ, V. (1978) "Ruský přízvuk v přehledech a komentářích", Praha.

TOWARDS DESIGNING AN INTONATION TRAINING DEVICE BASED ON SPEECH SIGNALS CLUSTERING

Leonid A. Kanter, Alexander V. Savin
and Ksenia G. Guskova

Dept. of Phonetics (English), Lenin Pedagogical State University, Moscow, USSR

ABSTRACT

The paper proposes a technique for automatic intonation type recognition for the purpose of designing an intonation training device to facilitate the acquisition of prosody by foreign language learners as well as to promote intonation improvement of those who have a speech or hearing disorder. The technique is applicable to any utterance with a fixed number of syllables and was initially tested for disyllabic speech samples of Russian and English.

1. INTRODUCTION

Broadly stated, the object of this study is to provide a means for computer-assisted phonetic instruction through which sound effects produced by a spoken voice (or voices) of a strictly standardized character, free of regional accents, may be visually compared with more or less related sound effects, produced by another speaking voice of a non-standardized character, i.e. a voice of a foreign language learner or a verbally/perceptually handicapped patient. Comparing the user's pronunciation with the stored pronunciation makes it possible

to measure a similarity or likelihood degree between the two types of speech patterns. More specifically, the paper proposes a technique for automatic intonation type recognition for the purpose of designing a visual display device for intonation training. The technique is applicable to any utterance with a fixed number of syllables (initially it was tested for disyllabic utterances of Russian and English).

2. CONCEPTUAL FOUNDATION

As a conceptual foundation the zonal principle of basic intonation types realization is adopted. In accordance with this principle, manifestations of a given intonation pattern are not just one-to-one reproductions of the underlying archetype. They are rather variants within the range of tolerance set by a particular invariant in the space of acoustic parameters (F_0 , I, T) and in the perceptual space.

3. GENERAL DESIGN

Intonation contours clustering results, reported for Russian and English in [1-3], were used as source data for the present project.

With implementation of the training device in question visualization of intonational zones is attained, thereby enabling the learner not only to hear intonation but also "to see" it without being tied to a specific intonation curve or a set of curves. In this case the language learner has to deal with points representing the curves on a display. A cluster analysis algorithm, proposed in [4], is used to reduce every curve to a point which is mapped on a plane. In order to make the initial cluster more convenient for teaching all extraneous data (i.e. points belonging to other clusters) are eliminated. As is shown in Fig. 1, the remaining points of each cluster are linked up in straight lines. It is assumed that the interpoint distance within the cluster does not exceed an empirically found fixed value. The procedure helps the learner to assess the cluster structure for the reference samples storage zone. When a new, prosodically distorted realization is mapped on a display, a decision as to which cluster it belongs is made, judging by the two neighbouring points between which it is located. The method is known as "the nearest neighbour strategy". Some possible results are exemplified in Fig. 2. The learner can modify his intonation realization as many times as required until the point on a display reaches the right cluster, that is to say, until the learner's point is located between two reference points of the target cluster.

This sort of a technical aid for teaching intonation may be PC-based.

4. CONCLUSION

Displaying visual information for speech training purposes, as reported here, will enable the learner to make a selection from a series of options available in the reference storage zone for a particular intonation type with due account of the individual range of the learner's vocal performance.

5. REFERENCES

- [1] CHIZHOV, A.P., CANTER, L.A. and SOKOLOVA, M.A. (1983), "English intonation in the space of acoustic parameters", Abstr. Xth Int. Congr. of Phonetic Sciences, Dordrecht/Cinnaminson, 360.
- [2] KANTER, L.A., CHIZHOV, A.P., GUSKOVA, K.G. (1987), "A cluster-seeking technique for prosodic analysis (with special reference to Russian sentence intonation)", Proc. XIth Int. Congr. of Phonetic Sciences, Tallinn, vol. 4, 59-61.
- [3] KANTER, L.A. (1988), "The systems approach as a methodological principle for intonation studies" (in Russian), Funktzionalnij analiz Foneticheskikh edinitz anglijskogo jazijka, Moskva, 26-70.
- [4] SAMMON, J.W., Jr., "A non-linear mapping for data structure analysis", IEEE Transactions on Computers, C-18, 5, 401-409, (1969).

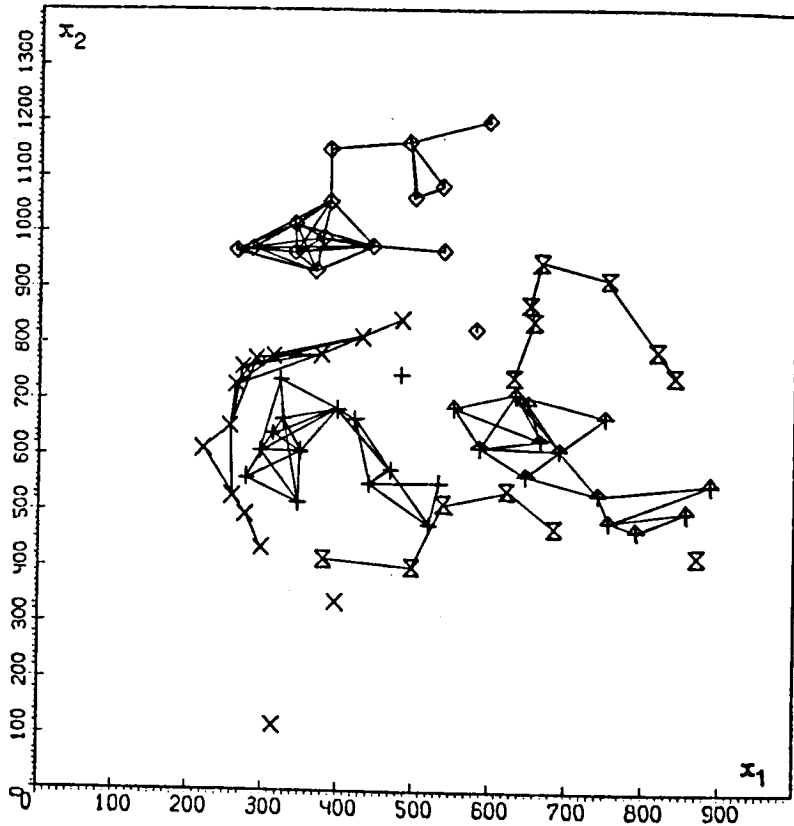


Fig. 1. Internal cluster structure for the reference samples storage zones (interpoint distance ≤ 170). The test phrase OH SHAI, pronounced by 16 male native speakers of Standard Russian was considered.

Symbols for intonation types
read as follows:

- + - final statement
- x - reply statement
- ◇ - general question
- ↑ - exclamation
- X - non-final statement

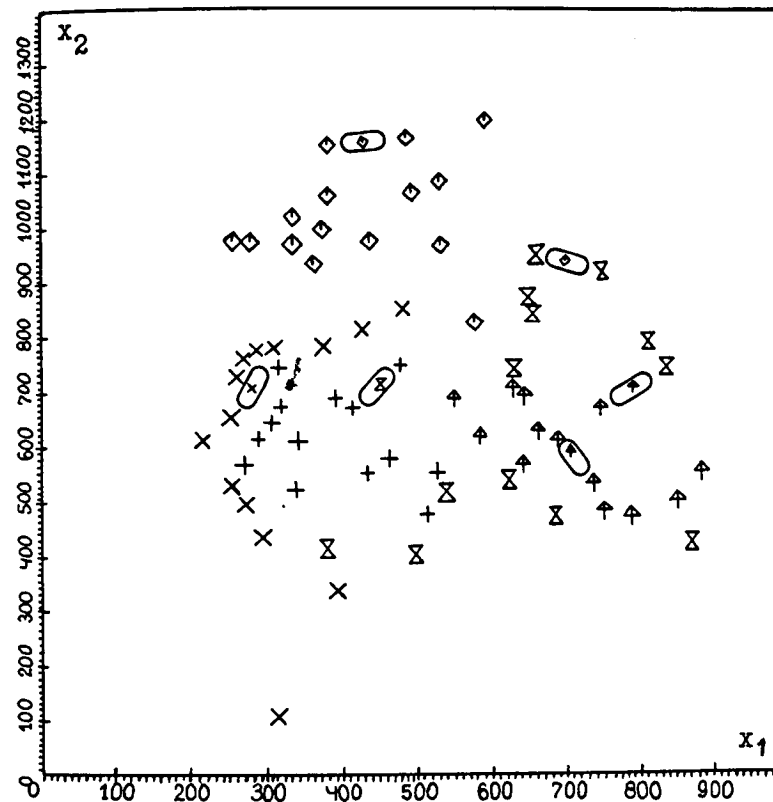


Fig. 2. Reference samples storage zones and non-native imitations of the Russian test phrase OH SHAI, produced by English learners of Russian. Non-native imitations are marked in smaller symbols within an ellipse.

CROSS-SECTIONAL TONGUE MOVEMENT AND TONGUE-PALATE MOVEMENT PATTERNS IN [s] AND [j] SYLLABLES

Maureen Stone,[†] Alice Faber,^{*} and Marc Cordaro[‡]

[†]National Institutes of Health, ^{*}Haskins Laboratories,

[‡]Johns Hopkins University

ABSTRACT

This study examines C-to-V movement patterns to determine the predictability of cross-sectional tongue movements, seen in ultra-sound images, from the movements in tongue-palate contact patterns, observed with EPG. Results indicate that tongue shapes are predictable for consonants, but not for C-to-V movement patterns, due to the confounding effects of jaw and tongue lowering.

1. INTRODUCTION

The tongue is one-third of the vocal tract and it is the major contributor to vocal tract shape. Measuring tongue movements, however, is very difficult, and real-time three-dimensional recording of such movements is not possible at present. One technique, ultrasound imaging, is able to provide two-dimensional scan sequences of tongue surface movements during speech. Scan sequences of the tongue can be imaged in multiple sagittal and coronal planes, and then temporally and geometrically aligned

into a time-varying, multi-sectional composite of the tongue [1].

In addition to tongue movement, the present study examined the relationship between tongue movements and tongue-palate contact. Electropalatography provides important information about how the tongue uses the palate to create various movement patterns. The hard palate is generally ignored in vocal tract models because it does not move. However, it is, in fact, quite important in tongue dynamics. The palate provides the tongue with a solid base of contact for sensory feedback, for light support during rapid or complex movements, and for resistance. When the tongue tip pushes against the palate, various tongue shapes and movements are facilitated.

The tongue is a boneless, jointless structure, yet it can elevate, depress, widen, narrow, extend, and retract. It also can create leverage, torsion, a midsagittal groove, a midsagittal arch, and move differentially, both laterally-to-medially

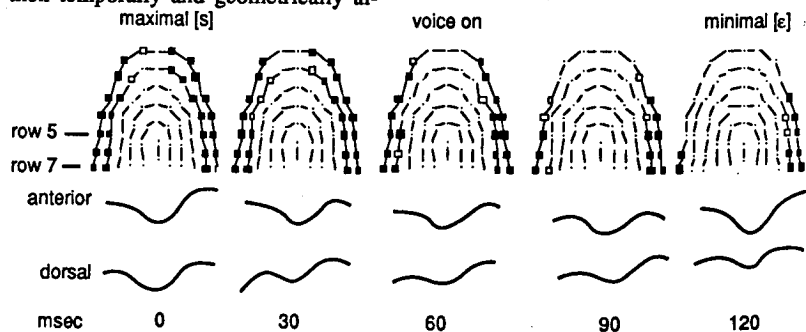


Figure 1: EPG and ultrasound sequences for the [s] to [e] movement. The sequence begins at maximal [s] and ends at minimal [e]. Samples are presented at every 30 ms and the two data sets are not exactly to scale, as the ultrasound images do not include the most lateral margins of the tongue. The ultrasound images show that the anterior tongue lowers as the dorsal tongue elevates. Jaw lowering (not pictured) lowers the tongue so that row 7 on the palate does not reflect dorsal tongue raising.

and left-to-right. These feats are remarkable, and are possible largely because of the tongue's exceedingly complex musculature, the support of the jaw, and the support and resistance of the palate.

Some ultrasound data already exist describing cross-sectional tongue shape [5] and movement patterns [6]. We also know quite a bit about tongue-palate contact patterns (e.g., [2,3,4]). The present study investigates whether patterns seen in coronal tongue movements during CV syllables are reflected in tongue-palate contact patterns.

2. METHODS

The subject was a normal male speaker of American English (New York City dialect) in his mid-40's. The subject's palate has a noticeable but normal degree of post-alveolar asymmetry. Maximal asymmetry occurred 30 mm behind the incisors: The right side was 2.5 mm higher than the left. The phones [s j e a] were embedded in [əCVCə] utterances, with the same C preceding and following the V. These phones were chosen because of their variety of tongue shapes, positions, tongue-palate contact patterns, and C-to-V movement patterns.

2.1 Ultrasound Recording/Analysis

Ultrasound images of the tongue were made in real time at the NIH, using an established recording procedure [5]. Each image was a 90° sector representing a 1.9 mm thick slice of tissue in the coronal plane. Scan sequences were produced at 30 scans per sec.

The ultrasound transducer was placed under the chin 20 mm posterior to the mental symphysis, using a specially designed holder. The beam angle was 110° posterior to the long bone of the man-

dible. The subject wore the electropalate, to create the same oral morphology as in the EPG recording session. The speech materials were recorded 10 times at this scan angle, and then 10 times with the transducer repositioned at 120° posterior, to provide anterior and dorsal coronal scan sequences of the tongue.

The ultrasound and audio signals were recorded on a videotape recorder, and analyzed on Macintosh IIfx and Compaq 386 microcomputers, using custom software [1,7]. For each coronal scan sequence, the video fields containing the movements from maximum C to minimum V position were entered into the computer. The tongue surface profile was extracted for each field and stored as xy coordinates for later use in graphics and other software applications.

2.2 EPG Recording and Analysis

EPG data were collected at a later date at Haskins Labs, using a custom fitted RION artificial palate containing 63 electrodes. The sweep rate for the palate was 64 frames/sec, and the electrodes are c. 4 mm apart. The speech materials were repeated 20 times, while the EPG and audio signals were recorded on an FM tape recorder. After the EPG data had been digitized, tokens of each utterance type were identified from the audio signal, and were aligned at the onset of the stressed vowel. A software generated composite containing electrodes contacted in 80% of the repetitions was created for each utterance. The pre-stress C and the stressed V were selected on the basis of maxima or minima of palatal contact, as appropriate for each sound.

The ultrasound scan sequences and the EPG frame sequences for each C-to-V

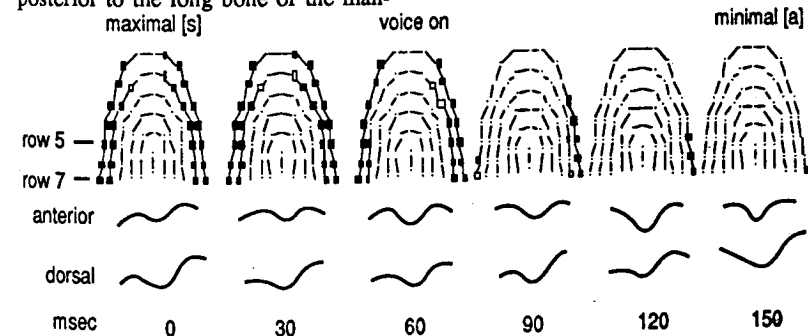


Figure 2: EPG and ultrasound sequences for the [s] to [a] movement. A small amount of anterior tongue lowering and dorsal tongue elevation can be seen. Jaw opening (not pictured) further lowers the tongue.

utterance were aligned using a time-warping algorithm that referred to the speech signals from each of the two datasets. The aligned C-to-V sequences appear in the accompanying figures.

2.3 Movement Patterns

In the ultrasound scan sequences, two patterns of C-to-V movement were observed and defined. The first, *midsagittal lowering*, was the fall of the mid-tongue to a greater degree than the lateral tongue, either creating a deeper groove or changing a convex (arched) tongue contour into a concave (grooved) one. The second pattern, *groove narrowing*, was the inward movement of the lateral sections of the tongue by at least one mm. Groove narrowing changes the width of the groove without necessarily deepening it; the two patterns can co-occur.

In the EPG frame sequences, two movement patterns were also defined during the C-to-V movement. The first, *medial contact decrease*, involved the loss of medial tongue-palate contact during the C-to-V transition. The second pattern, *anterior contact decrease*, was a decrease in the number of tongue-palate contacts in the four anterior rows of the electropalate. Also of interest were changes in medial contact and symmetry at rows 5 and 7 of the electropalate, the calculated location of the anterior and dorsal ultrasound scans.

3. RESULTS

All EPG frames showed asymmetry, reflecting the subject's palatal shape. EPG patterns showed bilateral tongue-palate contact during both [s] and [ʃ] (Figs. 1-4, frame 1). At maximal [s]

contact was symmetrical. At rows 5 and 7 tongue palate contact was approximately 8 mm wide on both sides. At maximal [ʃ], contact was asymmetrical. At row 5 contact was 8 mm wide on the left and 19 mm wide on the right. At row 7, contact was 13 mm wide on the left and 18 mm wide on the right. Ultrasound images showed that at maximal [s] the tongue was grooved midsagittally both anteriorly and dorsally (Figs. 1-2, frame 1, bottom). During [ʃ] (Figs. 3-4, frame 2, bottom), the tongue was arched and oblique with a higher right contour.

Movement into the following vowels was accompanied by a decrease in anterior and medial EPG contact, indicating anterior and medial tongue lowering, with retention of lateral contact. For minimal [e] and [a] (last frame) there was more asymmetry and greater medial contact following [ʃ] than [s] and more contact at row 7 than at row 5. During [s]-to-V movement, anterior-to-posterior tongue rotation was evident in the ultrasound scan sequences. Groovenarrowing and deepening appeared anteriorly during both [s]-to-V movements. During the [ʃ]-to-V movement, the tongue rotated left-to-right. Midsagittal lowering changed the arched shape into a groove during both vowels.

There is a fairly direct relationship between tongue-palate contact and cross-sectional tongue shape during the consonants for this subject (frame 1). Lateral tongue-palate contact 18 mm wide on the right accompanied the arched tongue shape seen in [ʃ]. The lesser, more lateral contact observed during [s] accompanied a lower tongue and a midsagittal

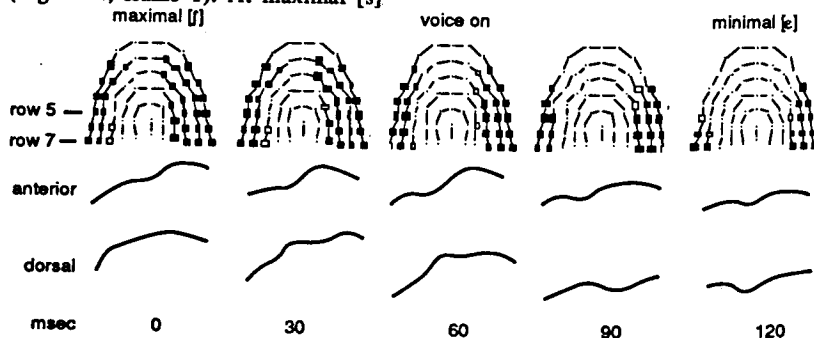


Figure 3: EPG and ultrasound sequences for the [ʃ]-to-[e] movement. The dorsal tongue lowers to a greater extent than the anterior tongue. The tongue grooves midsagittally as it lowers.

groove. Tongue movement features like groove onset, narrowing, deepening, and increased symmetry were not visible in the EPG data. Nor was anterior-to-posterior rotation.

4. DISCUSSION

The EPG patterns predicted the ultrasound cross-sectional tongue shapes, to some extent. Beyond some degree of medial tongue-palate contact, sibilants no longer used a grooved coronal tongue shape, as during [s], but rather an arched one, as during [ʃ]. Fewer medial EPG contacts reflected a checking of the upward force of the tongue and a midsagittal groove. Palatal asymmetries were reflected in tongue shape when the tongue approximated the palate more medially. [ʃ] used a wide area of tongue palate contact at the location of the asymmetry. [s]'s narrow, lateral tongue-palate contact occurred where the palate was level.

It was proposed earlier that tongue-palate contact provides the tongue both with sensory feedback and with resistance to help it assume various shapes. The correlation between tongue shape and EPG pattern for the consonants suggests that a high tongue/jaw position allows great force in the tongue-palate contact. This would facilitate maintaining a precise sibilant airstream. During movement, however, the jaw removes the support from the base of the tongue. EPG contact then may serve more for sensory feedback than for resistance, and, as such, reflects tongue shape to a lesser degree.

In conclusion, tongue-palate contact predicted cross-sectional tongue shape

during sounds that used a high jaw position. During vowels, tongue shape was not as predictable because jaw opening made tongue shape less directly a function of palatal contact. Finally, tongue-palate contact patterns did not reflect either overall tongue movement patterns like anterior-to-posterior tongue rotation or midsagittal movement patterns like groove narrowing and deepening because local tongue lowering and jaw opening removed these activities from the sphere of the palate.

5. ACKNOWLEDGEMENTS

The second author gratefully acknowledges support from NIDCD grand DC-00016 to Haskins Laboratories.

6. REFERENCES

- [1] CORDARO, M., M. STONE, M. GOLDSTEIN, & M. UNSER. 1991. A multi-sectional representation of the tongue surface based on ultrasound scans for time-varying vocalizations. *JASA*, 89, Sup.1.
- [2] FARNETANI, E., K. VAGGES, & E. MAGNO-CALDOGNETTO. 1985. Coarticulation in Italian VTV sequences. *Phonetica*, 42, 78-99.
- [3] FLETCHER, S. 1989. Palatometric specification of stop, affricate, and sibilant sounds. *JSHR*, 32, 736-748.
- [4] RECASENS, D. 1984. V-to-C coarticulation in Catalan VCV sequences. *J. Phon.* 12, 61-73.
- [5] STONE, M., T. SHAWKER, T. TALBOT, & A. RICH. 1988. Cross-sectional tongue shape during vowel production. *JASA*, 83, 1586-1596.
- [6] STONE, M. 1990. A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data. *JASA*, 87, 2207-2217.
- [7] UNSER, M. & M. STONE. 1990. Computerized extraction of the tongue surface from sequences of ultrasound images. *JASA*, 87, S122.

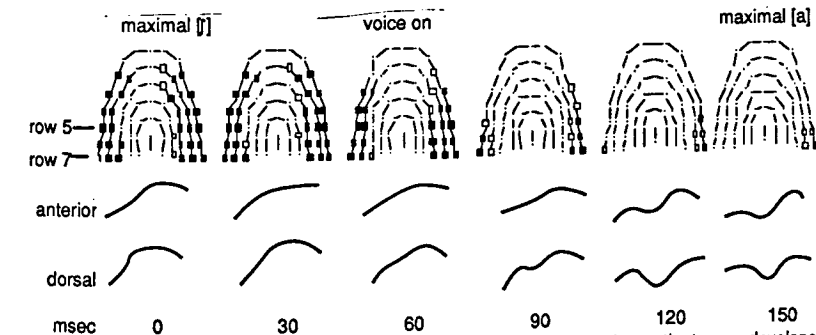


Figure 4: EPG and ultrasound sequences for the [ʃ]-to-[a] movement. In both planes, the tongue develops a midsagittal groove and lowers in height. Dorsally, it rotates left-to-right. Jaw lowering (not pictured) further lowers the tongue.

AN EXAMINATION OF THE JAW'S CONTRIBUTION TO LINGUAL STABILITY

Eric Vatikiotis-Bateson †, Maureen Stone and Michael Unser ††

†ATR Auditory & Visual Perception Research Laboratory, Japan

†† National Institute of Health, USA

ABSTRACT

Analysis of ultrasound, electropalatography, and jaw motion data for production of VCVC_a sequences has shown the coupling among the tongue, jaw, and maxilla to be quite different for /s,ʃ/ and /l/ [6]. In the current study, two sensors are used to transduce jaw motion affording better distinction among the rotational and translational components of the movement; and a new technique for extracting tongue surface contours [5] is used to correlate loci of curvature with jaw motion and palatal contact patterns. [Supported in part by NIH grant DC-00121 to Haskins Laboratories.]

1. INTRODUCTION

In speech production, we try to make quantifiable observations of complex structures and events without trading scope for precision. The tongue is certainly the most interesting and complicated articulator structure to observe because, while difficult enough to examine alone, it continually interacts with other structures. Production of alveolar consonants, for example, entails interaction among the tongue — the whole tongue — the jaw, and the maxillary arch. Unfortunately, no single transduction method affords simultaneous observation of all the major components of this interaction. Tongue-palate contact is largely non-sagittal and often asymmetrical [2] and the jaw's motion is not a trivial rotation around a pivot [1].

In recent studies we have tried to improve our understanding of tongue behavior and its interaction with the jaw and maxilla during alveolar production by combining data from a variety of sources. Examples are the combined use of ultrasound (US) imaging and x-ray microbeam [3] or US, electropalatography (EPG), and jaw motion, [6]. Although mixing such techniques makes data analysis more elaborate, especially since the data cannot all be recorded simultaneously, it has given us insight into tongue behavior and its functional coupling with other structures. In what follows, we further consider these issues aided by two technical improvements: extraction of tongue surface contours from digitized ultrasound and the use of two position sensors in tracking jaw movement.

2. METHODS AND PROCEDURE

In this paper, we discuss a small sample of ultrasound (US), electropalatography (EPG), and jaw motion data taken from a much larger set of utterance types and speakers. The sample consists of one speaker's VCVC_a utterances, where C is s, ʃ, or l and V is a. Each utterance type was repeated 10 times in succession at a rapid, but unprompted rate.

Ultrasound images were recorded at NIH using an ATL ultrasound unit and 30 msec sector scanner transducer. The transducer was mounted under the subject's chin so as to maintain a precise angle of tilt and to minimize transduction of jaw motion (For details, see

[4]). The ultrasound images were digitized and spatially smoothed using Wayne Rasband's IMAGE program running on a Macintosh II equipped with a Data Translation Quick Capture Board (DT 2255). Tongue surface contours were extracted using software developed by Michael Unser [5].

Jaw movement, EPG, and acoustic data were recorded at Haskins Laboratories. Vertical and horizontal (Anterior-posterior) movement of the

jaw was transduced from two infrared LEDs placed 4 cm apart on a rigid splint attached to and extending mid-sagittally from the jaw. Tongue palate contact was transduced at 64 frames per second via a Rion flexible palate and electropalatograph. Movement data were digitized at 200 Hz, numerically smoothed at 40 Hz, corrected for head movement, and differentiated to obtain instantaneous velocity.

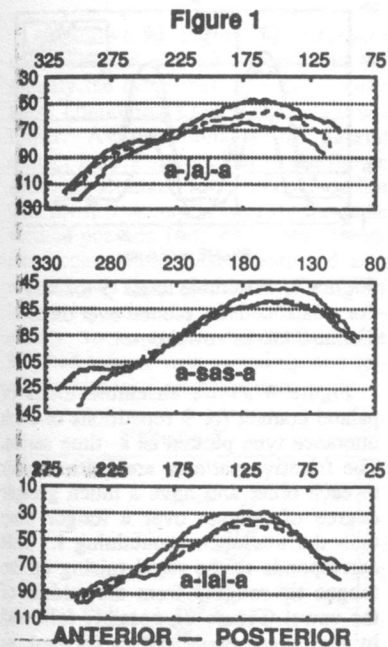


Figure 1. Extracted tongue contours taken from C1 (solid), V (dotted), and C2 (solid). Plot coordinates are rotated 35 degrees relative to the head.

3 RESULTS

3.1 Tongue Surface

In Figure 1, extracted tongue surface contours corresponding to maximum positions achieved during the C1, V and C2 portions of an utterance are shown for each consonant. The fricatives, s and ʃ, behave in similar fashion. Anterior blade is high and forward for

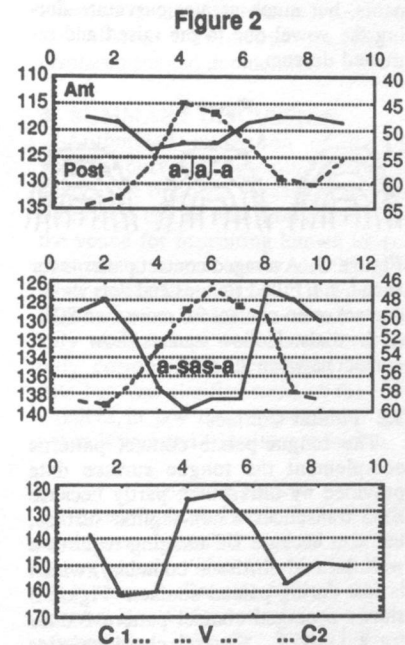


Figure 2. Top traces: overlaid time series measures for anterior blade (solid) and dorsum (dashed). Bottom: radius of circle fit to tongue surface over time.

C1, lowers during the C1-V transition as the dorsum raises and retracts, and then reverses the process in the V-C2 transition. Thus, the tongue surface rocks back and forth around an anterior-posterior pivot. This is also shown in the top two panels of Figure 2, where measures of blade and dorsum height, calculated by the curvature extraction

program, are overlaid for the CVC time series. For *l* and despite the fact that the jaw is raising and lowering, the scans show little change of position for anterior blade and the extreme posterior of the tongue. Finally, tongue curvature is greater during the vowel for all 3 utterances. The bottom panel of Figure 2 shows an example of how the radius of the circle that was fit to the tongue surface decreases for the vowel. The larger radii for C1 and C2 suggest a flatter tongue surface for the consonants, but much greater curvature during the vowel due to the raised and retracted dorsum.

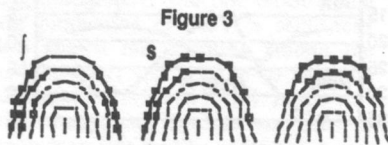


Figure 3. Averaged contact patterns for each /s/, /ʃ/, /l/. Filled squares and dots show contact or no-contact for more than 80% of the trials. Hollow squares show contact between 20-80%.

3.2 Palatal Contact

The tongue-palate contact patterns complement the tongue surface data provided by ultrasound; partly because EPG transduces a non-sagittal surface, but also because US imaging requires a well defined air/tissue boundary, which is lost during palatal contact. Figure 3 shows averaged contact pattern frames for /s/, /ʃ/, and /l/. Several characteristics are noteworthy. First, there is less contact along the sides (parasagittally) for /s/ than /ʃ/. This could be due to the more grooved sagittal channel of /ʃ/. Second, there is more anterior, including midsagittal, contact for /s/ than /ʃ/. This suggests that the tongue tip for /ʃ/ is angled down and/or retracted more than for /s/. Third, contact for /l/ is restricted to the anterior portion of the artificial palate. Although this speaker's contact pattern is more bilaterally symmetrical than many we have seen, the absence of extensive posterior bilateral contact is typical. Previously,

we have suggested that the tongue tip may serve as an anchor for the lateral post-alveolar release. However, in this speaker's case, it could indicate that the tongue tip is angled upward relative to the rest of the blade. Finally, these patterns are very stable across repetitions as shown by the small number of hollow squares, indicating that a given electrode is either on or off for more than 80% of the repetitions.

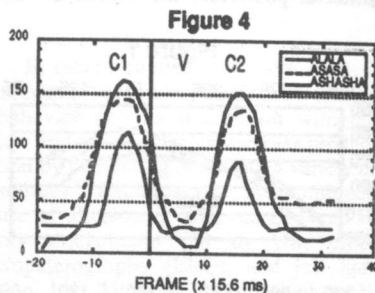


Figure 4. Ensemble totals (9 tokens) for palatal contact, plotted over time. Frame 0 marks vowel onset.

Figure 4 shows ensemble totals of palatal contact for 9 repetitions of each utterance type plotted as a time series. The fricative patterns are quite similar to each other and have a much greater degree of contact over a longer time than the utterances containing /l/. This corresponds to the rapid raising of the tongue tip roughly from the middle of the vowel (Figure 2), possibly followed by the tongue blade sliding forward as the jaw raises into position. Unlike contact for /l/ which has substantially less contact for C2 than C1, the greater precision required for production of fricatives might explain why there is only slightly less contact for post-tonic C2. Finally, the relatively abrupt onset and offset of contact for /l/ may result from articulation of the more agile tongue tip, hence the absence in /l/ of the anterior-posterior pivoting observed for /s/ and /ʃ/ (Figure 2).

3.3 Jaw Motion

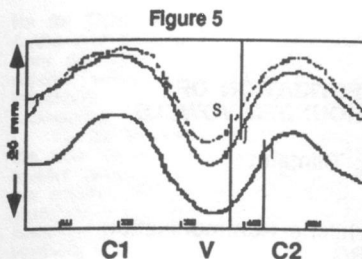


Figure 5. Averaged jaw position over time for each utterance type.

As shown in Figure 5, utterances containing /s/ and /ʃ/ are produced in roughly the same vertical region, while those containing /l/ are produced much lower. Average onsets of palatal contact for C2 are marked in the figure. Although contact for /s/ occurs earlier than for /ʃ/, it occurs at roughly the same vertical position (diff < .4 mm). /l/ contact occurs 40-50 msec later and at a much lower jaw position. The jaw's lowered position for /alala/ may be necessary to accommodate the more bunched tongue with raised tip and relatively high and retracted dorsum.

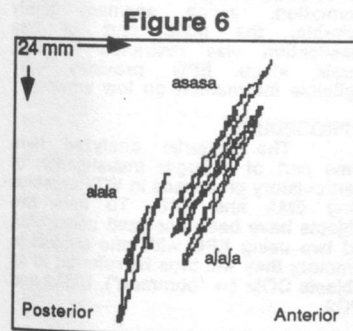


Figure 6. Lissajous plots show 2 dimensional motion of the jaw for each utterance type.

When two dimensional motion of the jaw is examined (Figure 6), we see that the jaw also is more retracted for /l/ relative to /s/ and /ʃ/, and that there is much less jaw rotation — i.e., the jaw is translated vertically. Analysis of the jaw kinematics corroborates the quali-

tative differences seen here. When jaw displacement measures (C1-V, V-C2) are compared for the two position sensors, the displacement difference (sensor 1 - sensor 2) was reliably less for /l/ than the two fricatives, indicative of a greater degree of vertical translation. Furthermore, comparing the sensor displacement difference for C1-V and V-C2, there was even more vertical translation during the raising V-C2 movement for /l/. Finally, the jaw moved with fairly constant average velocity for the fricatives, but there was no linear relation between displacement and duration for /l/.

4. SUMMARY DISCUSSION

Automated extraction of tongue surface curvature greatly facilitates analysis of digitized ultrasound images. More data can be analyzed quickly and the venue for measuring known tongue parameters as well as the identification of new ones is enhanced. Transduction of jaw motion which can be more reliably analyzed in two dimensions not only corroborates the tongue data but also clarifies some of the differences in jaw movement control between lateral and fricative production.

REFERENCES

- [1] Edwards, J., & Harris, K.S. (1990). Rotation and translation of the jaw during speech. *JSHR*, 33, 550-562.
- [2] Hamlet, S., Bunnell, H.T., & Struntz, B. (1986). Articulatory asymmetries. *JASA*, 79, 1164-1169.
- [3] Stone, M. (1990). A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data. *JASA*, 87, 2207-2217.
- [4] Stone, M., Shawker, T., Talbot, T., & Rich, A. (1988). Cross-sectional tongue shape during vowel production. *JASA*, 83, 1586-1596.
- [5] Unser, M., & Stone, M. (1990). Computerized extraction of the tongue surface from sequences of ultrasound images. *JASA*, 87, S122.
- [6] Vatikiotis-Bateson, E., & Stone, M. (1989). In search of lingual stability. *JASA*, 86, S115.

AN ARTICULATORY INVESTIGATION OF FRONT ROUNDED AND UNROUNDED VOWELS

P. Hoole and H.G. Tillmann

Institut für Phonetik und Sprachliche Kommunikation
Munich, FRG

ABSTRACT

Using electromagnetic and palatographic techniques stable differences in lingual articulation for the pair of German front rounded/unrounded vowels /ʌ/ and /y/ were found. This result was related to S. Wood's hypothesis that the articulatory adjustments for such pairs help to maintain each vowel in regions of acoustic stability and enhance their distinctiveness. A second aim was to test the hypothesis that German's complex set of front vowels leads to less variability in the articulation of /ʌ/ compared with the other point vowels /u/ and /a/. This expectation was also confirmed.

1. INTRODUCTION

Many languages, including German, contrast front rounded and unrounded vowels. Yet there is evidence that the distinction may also involve differences in tongue configuration [3]. Of particular interest is Wood's [6] study in which in a large number of languages he typically finds that /y/ has a lower tongue-body position than /ʌ/. He argues that the labial, lingual and also laryngeal manoeuvres for such pairs represent a balanced set of adjustments that maintain each vowel in regions of acoustic stability and enhance their distinctiveness. Wood's evidence is based largely on radiographic data with, of necessity, a restricted range of utterances. In this study we focus on just one aspect of the /y/ contrast, namely the potential lingual differences, but examine it in a wide variety of consonantal contexts in order to determine how robust the distinction is. This would then make it possible to assess the relative importance of the lingual adjustments within the bundle of features contributing to the rounded-unrounded distinction. Ultimately, in combination with studies of the other aspects of the distinction, this work should give a better understanding of the extent to which articulation actually

takes into account considerations of acoustic stability derived on the basis of acoustic theory.

The second aim of this study is dependent on the first one: If stable differences were found this would mean that German may make up to an 8-way distinction in tongue position for front vowels. It then becomes pertinent to ask whether this complexity results in reduced contextual variability for front vowels compared with back vowels.

Electromagnetic articulography (EMA) and electropalatography (EPG) were used to collect the relevant data on lingual configuration. Both techniques readily allow recording and analysis of a large number of utterances.

In order to be able to use EMA and EPG as complementary sources of information, which seemed highly desirable, the main part of this investigation was restricted to high vowels since EPG provides only negligible information on low vowels.

2. PROCEDURE

The material analyzed here forms part of a larger investigation of coarticulatory processes in VCV syllables using EMA and EPG. To date two subjects have been analyzed using EMA and two using EPG with one subject in common; they will thus be referred to as subjects COM (= "common"), EMA2 and EPG2.

2.1 EMA recordings

A commercially available system for electromagnetic movement transduction (Carstens Medizintechnik) was used to monitor movement of tongue and jaw [4], [5], [2]. To this end 3 receiver coils were mounted on the midline of the tongue at locations ranging from 1 to 5 cm from the tongue tip, together with one on the lower incisors (jaw) and upper incisors (reference). The x/y coordinates at these five positions were recorded by a dedicated PC at a sample rate of 193.5

Hz for COM and 250 Hz for EMA2. Audio and synchronization information were recorded on DAT tape and all signals were then transferred to a laboratory computer for further processing.

For the purposes of this experiment the data were placed in a coordinate system whose origin is at the average jaw position for the high front vowels examined here, and with the principal component of jaw movement oriented vertically.

2.2 EPG recordings

The EPG recordings were made on the Reading University multichannel data acquisition system, the EPG sample rate being 200 Hz. The EPG and accompanying audio signal were also transferred to the lab computer for further analysis together with the EMA data.

2.3 Material

A large corpus of VCV nonsense items was recorded. The corpus for the EMA recordings consisted of words of the form /bV1CV2/, the consonants being /p, b, m, f, v, t, d, n, l, s, sh, k, g, h/ and the vowels /i, y, u, a/. All combinations of /i, u, a/ were used but /y/ was only combined with /a/, giving 154 forms in all. The corpus for the EPG recordings was basically the same but without the consonants /m, f, v, g, h/ and also without the initial /b/. The EMA corpus was spoken with a carrier phrase "sage — bitter" whereas the EPG recordings were not.

For both techniques 5 repetitions of each item were aimed for, which was slightly overachieved for EPG and slightly underachieved for EMA due to coils becoming detached prematurely.

3. ANALYSIS AND RESULTS

A waveform editor was used to locate the beginning and end of each vowel in the audio signals. All articulatory analyses were carried out at the mid-point of the vowels so defined.

Since the EPG results are rather more clear-cut they will be discussed first.

3.1 EPG

The electrodes on the artificial palate can be regarded schematically as being arranged in 8 rows and 8 columns. A measure of the location of the articulation on the front-back dimension can be derived by summing the number of contacts in each row and then determining the centre of gravity ("CG") of this vector of 8 values. One additional parameter proved sufficient to capture the difference between /ʌ/ and /y/, namely the grand total ("TOTAL") of the

number of contacts. For similar values of CG different values of TOTAL will indicate differences in tongue height. The results are accordingly presented in Fig. 1 with CG on the x-axis and TOTAL on the y-axis. This figure shows all V1 tokens of /ʌ/ and /y/ with $V2 = /a/$, the data for each vowel being enclosed by a 2-sigma ellipse. The /y/ distinction is obviously very clear-cut for both subjects. COM distinguishes the vowels solely on the basis of TOTAL indicating a lower tongue position for /y/, whereas EPG2 also distinguishes in terms of CG. In fact, the shift of CG to a more rearward value is probably also the reason for less overall contact (an increasing number of anterior rows becoming devoid of contact), so that in contrast to COM the primary mechanism in the /y/ distinction can be assumed to be tongue retraction for /y/.

3.2 EMA

The EMA results for the two subjects are shown in Fig. 2, each plot showing both subjects at one tongue measurement position. In parallel with Fig. 1 larger values on the x-axis indicate more posterior tongue position.

Although the distinction between /ʌ/ and /y/ is less sharp than in the palatographic data clear tendencies remain. COM shows overall a lower tongue position for /y/, thus reinforcing the interpretation placed on his palatographic data, while speaker EMA2 has a more retracted position for this vowel, thus patterning like the second EPG speaker. The distance between the centres of the /ʌ/ and /y/ ellipses, averaged over the three coils, amounts to 1.75 mm for COM and 1.85 mm for EMA2. Although the main purpose of this study was to determine whether the /y/ contrast is stable over many contexts we nonetheless examined to what extent the contrast is enhanced in a more restricted context. The tongue position was accordingly re-evaluated in labial consonantal contexts (p, b, m, f, v). The average distance between /ʌ/ and /y/ increased to 2.3 mm for COM and 2.2 mm for EMA2. Of greater significance was a reduction in the size of the 2-sigma ellipses (averaged over the 3 coils and 2 vowels) from 15.1 to 7.7 mm² for COM and 14.5 to 11.8 mm² for EMA2, the net result being virtually no overlap between the /ʌ/ and /y/ ellipses.

The question arises as to whether the differences in tongue position are a passive effect of differences in jaw position. Both speakers had for /y/ a slightly higher (by 0.4 mm for COM and by 0.7 mm for EMA2) and more retracted (by 0.9 mm for COM and 0.6 mm for EMA2) jaw position. Thus the low tongue position

In COM cannot be explained by jaw influence, but the more retracted tongue position for EMA2 might be partly explained in this way.

Comparison of the EMA and EPG results gives some indication of the robustness of the /i/ distinction. As seen above, the separation is sharper in the EPG data. One possible reason is that the EPG recordings were spoken without a carrier phrase, thus encouraging a more deliberate style of speech. This was certainly the case for COM whose EPG vowels were almost 100 ms longer than the EMA vowels. On the other hand the vowels of EMA2 proved somewhat longer than those of EPG2 so speaker-specific traits may also play a role as well as influences of the carrier phrase not reflected in vowel length.

3.3 Overall vowel variability

Accepting that there are consistent differences in tongue position in /i/ and /y/ and thus potentially a large number of lingual distinctions to be made among German front vowels we aimed to determine whether this is reflected in different ranges of contextually inducible variability for different vowels.

As a first approach to this question we simply measured in the EMA data the areas of the 2-sigma ellipses for each of the point vowels /i, u, a/ in the corpus over all consonantal and vocalic contexts. A highly consistent picture emerged. Both subjects showed a variability order of $i < u < a$ at back and mid coils and $i < a < u$ at the front coil. The results averaged over all coils are given in Table 1, clearly showing the lesser degree of variability for /i/.

4. OUTLOOK

The aim of this paper was to provide a foundation on which to generate hypotheses for future work. Little would be gained from a more comprehensive investigation of the front rounded vs. unrounded vowels if the lingual differences proved highly unstable. But this was not the case. One important point that emerges from this investigation is that speakers appear to differ somewhat in the articulatory adjustments they use to distinguish the rounded/unrounded pairs. In particular, the tendency towards retraction in two of the three speakers was slightly unexpected as this pattern was not found by Wood, and in fact he suggests on the basis of his modelling studies that it is a pattern that is not conducive to maximum acoustic distinctiveness between /i/ and /y/. Acoustic analysis that is in preparation, and lip-movement recordings that are planned, should help throw more light on this issue.

It also remains to be demonstrated that an 8-way distinction in front vowel tongue position actually occurs in German. Based on the results found here for the effect of vowel length on the distinction one specific expectation is that these potential oppositions will be effectively neutralized in the short, lax rounded/unrounded pairs.

Finally, support was also found for the suggestion that the crowded front vowel space will constrain the amount of contextual variation for these vowels. However, there may well be a bias towards low variability in /i/ in this experiment. Firstly, none of the coils were really placed very far dorsally. Secondly, it has been suggested [1] that high vowels tend to benefit from the proximity of the hard palate to reduce the range of variability. These issues could be easily resolved in future work with a more comprehensive sample of the German vowel system.

Acknowledgment

Thanks to Claudia Mässiggang for help with data analysis and to Bill Hardcastle, Fiona Gibbon and Katerina Nicolaidis of Reading University for managing the EPG recordings. Work supported by ESPRIT/BRA 3279 ACCOR.

5. REFERENCES

- [1] FLEGE, J.E. (1989), "Differences in inventory size affect the location but not the precision of tongue positioning in vowel production", *Language and Speech*, 32, 123-147.
- [2] HOOLE, P., GFROERER, S., TILLMANN, H.G. (1990), "Electromagnetic articulography as a tool in the study of coarticulation", *FIPKM (=Forschungsberichte des Instituts für Phonetik, Munich) 28*, 107-122.
- [3] MEYER, E.A. (1907), "Röntgenographische Lautbilder", *Monatsschrift für die gesamte Sprachheilkunde* 17, 225-243.
- [4] SCHÖNLE, P., MÜLLER, C., WENIG, P. (1989), "Echtzeitanalyse von orofacialen Bewegungen mit Hilfe der elektromagnetischen Artikulographie", *Biomedizinische Technik*, 34, 128-130.
- [5] TULLER, B., SHAO, S., KELSO, J.A.S. (1990), "An evaluation of an alternating magnetic field device for monitoring tongue movements", *J. Acoust. Soc. Am.* 88(2), 674-679.
- [6] WOOD, S. (1988), "The acoustical consequences of tongue, lip and larynx articulation in rounded palatal vowels", *J. Acoust. Soc. Am.* 80, 391-401.

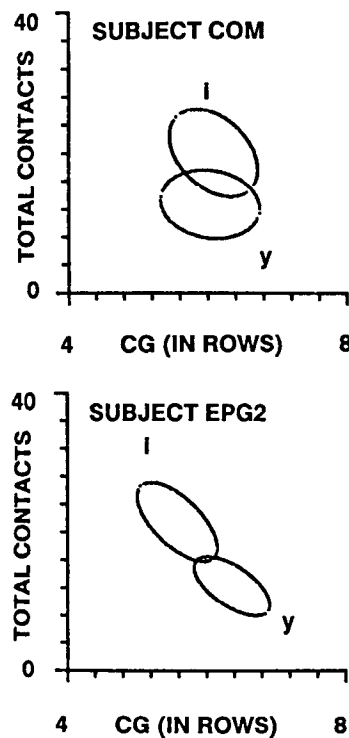


Fig. 1: Results for the two EPG subjects. Higher values of CG indicate more posterior tongue position. COM: N = 70; EPG2: N = 48.

Table 1

	/i/	/u/	/a/	
COM	14	31	34	N = 167
EMA2	14	32	33	N = 210

2-sigma area of variability in mm², averaged over the 3 tongue coils.

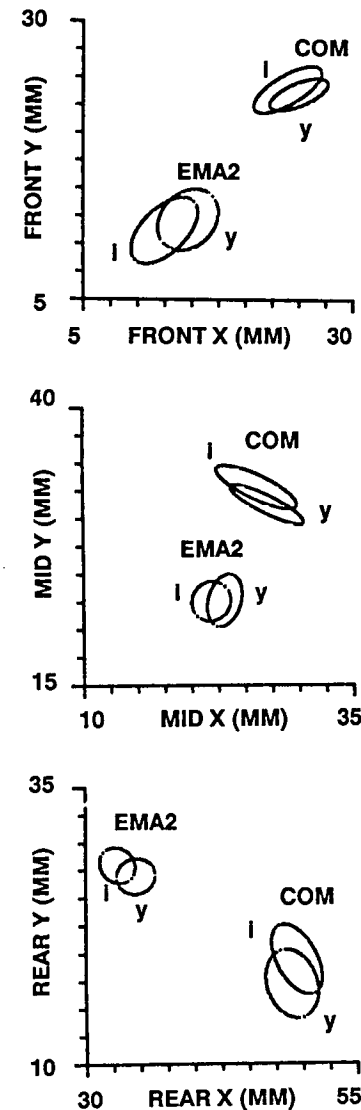


Fig. 2: Results for the two EMA subjects, displayed separately for each tongue measurement position. Higher values on the x-axis indicate more posterior tongue position. COM: N = 56 (Front and Mid coils), N = 56 (rear coil); EMA2: N = 70.

TRACT SHAPES OF TENSE AND LAX VOWELS: A COMPARISON OF X-RAY MICROBEAM AND EMG DATA

Katherine S. Harris,^{†,‡} Eric Vatikiotis-Bateson,^{††}
and Peter J. Alfonso^{†,*}

[†]Haskins Laboratories, New Haven, CT; [‡]Graduate School,
City University of New York, New York; ^{††}ATR Research
Laboratories, Kyoto, Japan; ^{*}University of Connecticut, Storrs.

ABSTRACT

X-Ray microbeam data for simple /əpVp/ utterances for a single speaker were compared with EMG measures from muscles of the tongue and jaw. For the six vowels /i, ɪ, e, æ, u, ʊ/ pellet displacements at vowel center were roughly proportional for front tongue X and Y pellets and a rear X pellet. The rear Y pellet showed somewhat different relationships among the vowels. The relationships between several of the pellet displacements and normalized EMG appeared to be monotonic.

1. INTRODUCTION

Although there is a tradition of classifying the tongue shapes for vowels along the dimensions of high vs low, back vs front, it is well known that these dimension labels do not accurately mirror any conventional geometrical space (see e. g., [4]). Furthermore, as has been pointed out by Wood [8] these dimensions do not mirror the dimensions of vocal tract shape as it is set primarily by the anatomy of the tongue muscles. While these dimensions have been modelled at least twice [5; 7], the empirical information necessary for the refinement of such models is lacking. The present experiments are a modest beginning towards filling this gap, and are a continuation of the work of Baer, Alfonso, and Honda [2] and Alfonso and Baer [1] on the same topic.

2. METHODS

The experiment was done in two parts. The talker for both was TB, an author of the previous papers in this series. The speech consisted of isolated utterances

of the form /əpVp/. While the total set was larger, data from the vowels /i, ɪ, e, æ, u, ʊ/ only will be reported here. In all cases, multiple tokens were averaged with respect to a common lineup point at consonant release.

For Part One of the experiment, hooked wire recordings of the electromyographic signal were made from the muscles anterior and posterior genioglossus, (GGA and GGP), hyoglossus, (HG), styloglossus (SG), geniohyoid (GH), mylohyoid (MH), and orbicularis oris (OO). Simultaneously, acoustic recordings were made, as were recordings from jaw movement in the Y dimension, using an optoelectric movement transducer. These data have been previously reported [2].

For Part Two of the experiment, x-ray microbeam data were taken on the same subject for the same inventory. Pellets on mid- and rear-tongue, lower lip and jaw were analyzed with a system then at the Institute of Logopedics and Phoniatrics at the University of Tokyo [6]. An effort was made to keep the utterance rate similar across the two experiments. Success in obtaining comparability could be assessed by comparing audio recording and jaw Y tracings.

3. RESULTS

3.1 X-Ray Microbeam Analysis

Pellet tracings for the averaged tongue front movement in the X dimension is shown in Fig. 1. The results conform fairly well with expectations, in that the extreme vowels fan out forwards and backwards from the initial neutral syllable.

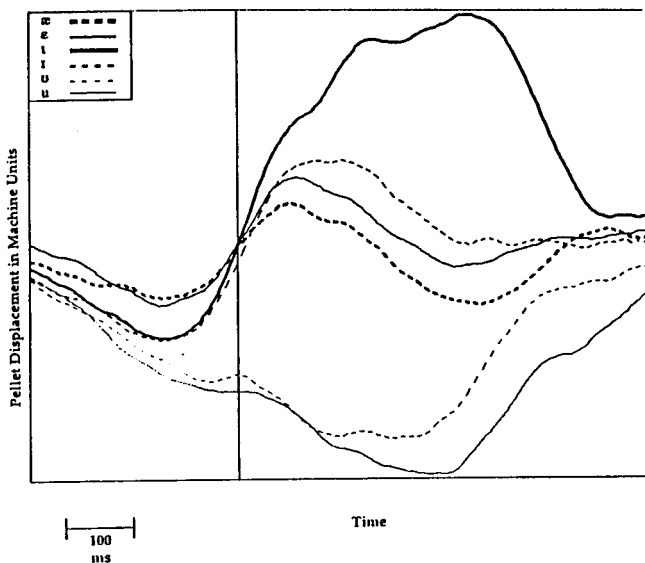


Figure 1. Averaged Front X pellet movement as a function of time. The line up point, /p/ burst release, is indicated by a solid vertical line.

The trajectory shape for the lower-lip Y pellet was used to locate vowel midpoints. Midpoint values were then used in comparisons with the EMG data. The overall relationship of the vowels to each other at midpoint is quite similar for X and Y front pellets, and X rear pellet. Relative positions for the extreme vowels are different for the rear Y pellet.

3.2 Comparison of EMG and x-ray data

In order to compare vowel midpoint positions and EMG data, we assumed a mechanical response time of the muscle of 100 msec, a value calculated in an earlier study [1]. A point in each EMG files 100 msec earlier relative to the lineup than the vowel midpoint was located. The EMG values were normalized so that the largest value observed for that muscle was treated as 100%. These values were used for scatter plots relating (front X, rear X, front Y) or rear Y position to the tongue muscle value, or GH value to jaw Y position. Because the tongue pellet positions have not yet been corrected for jaw position [3], nor have we attempted to adjust the files for the small differences in speaking rate over the two parts of the experiment, as we intend before final presentation, we will present only partial results at this time.

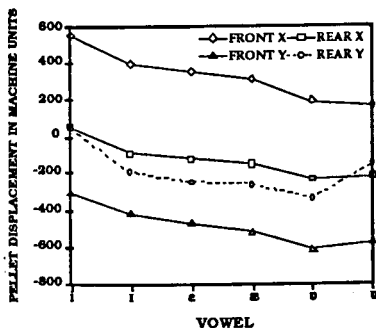


Figure 2. Tongue pellet position at vowel midpoint for X and Y dimensions of front and back pellets.

The sole muscle in the experimental set concerned with jaw opening, GH, correlates quite well with jaw Y position.

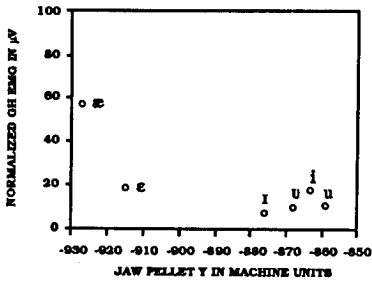


Figure 3. Scatter plot of Jaw Y position as a function of GH EMG activity.

With respect to correlations with the tongue X position group, the strongest relationship is with SG EMG activity.

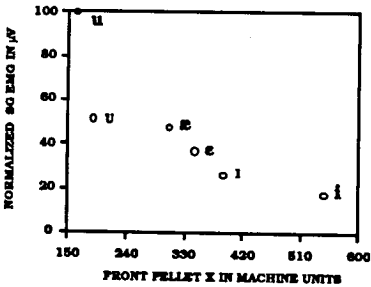


Figure 4. Scatter plot of Front Tongue X position with SG EMG activity.

The strongest relationship to rear Y position is that of GGP EMG.

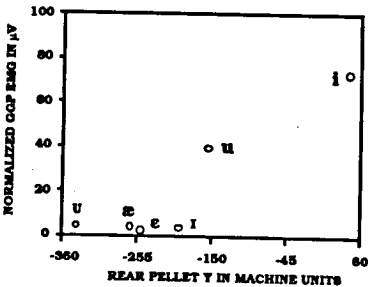


Figure 5. Scatter plot of Rear Tongue Y position with GGP EMG activity.

The activity of MH relates most strongly to Front Pellet X EMG activity. However, a more quantitative assessment must await file correction.

4. ACKNOWLEDGMENTS

This work was supported by NIH Grant DC-00121 to Haskins Laboratories. We are grateful to the staff of the Research Institute for Linguistics and Phoniatrics for their collaboration with Drs. Baer and Alfonso, who collected the x-ray microbeam data there. The x-ray data was analyzed by Arlyne Russo. Joseph Kalinowski and Haralambia Kollia prepared the figures for this abstract.

5. REFERENCES

- [1] ALFONSO, P. J. & BAER, T. (1982). "Dynamics of vowel articulation." *Language and Speech*, 25, 151-173.
- [2] BAER, T., ALFONSO, P. & HONDA, K. (1988). "Electromyography of the tongue muscles during vowels in /epVp/ environment". *Annual Bulletin Research Institute of Logopedics and Phoniatrics*, 22, 7-19.
- [3] EDWARDS, J., & HARRIS, K.S. (1990). "Rotation and translation of the jaw during speech". *Journal of Speech and Hearing Research*, 33, 550-562.
- [4] FISCHER-JØRGENSEN, E. (1985). "Some basic vowel features, their articulatory correlates, and their explanatory power in phonology", in V. A. Fromkin, (Ed.), *Phonetic Linguistics*. New York: Academic Press.
- [5] KAKITA, Y., FUJIMURA, O., & HONDA, K. (1985). "Computation of mapping from muscular contraction patterns to formant patterns in vowel space". In V.A. Fromkin, (Ed.), *Phonetic Linguistics*. New York, Academic Press.
- [6] KIRITANI, S., ITOH, K., & FUJIMURA, O. (1975). "Tongue-pellet tracking by a computer-controlled x-ray microbeam system". *Journal of the Acoustical Society of America*, 57, 1516-1520.
- [7] PERKELL, J.S. (1974). "A physiologically-oriented model of tongue activity in speech production". Unpublished doctoral dissertation, Massachusetts Institute of Technology.
- [8] WOOD, S. (1979). "A radiographic study of constriction location for vowels". *Journal of Phonetics*, 7, 25-43.

MODELING VOWEL ARTICULATION / MODÉLISATION DE L'ARTICULATION DES VOYELLES

Michel T. T. Jackson

Speech Communication Group, MIT, and
Speech & Hearing Science, Ohio State University

ABSTRACT

This paper discusses a cross-linguistic articulatory model of vowels based on sagittal plane x-rays of vowels from Akan, Arabic, Chinese, and French. The model gives a mapping from a universal space of articulatory parameters to the specific vowels of each language. It contains an explicit parametrization of speaker variation, based on data from 13 speakers. We show here that the model also generalizes to certain consonants. The results suggest that the phonological description of these consonants can be changed profitably.

1. INTRODUCTION

The aim of this paper is to present an articulatory model of vowel production. Vowel articulation is simpler than consonant articulation because there is relatively little contact-related deformation in the shape of the tongue and other articulators and so measurements of tongue position in vowels do not show non-linear "ceiling effects" due to the contact of the tongue with the hard palate.

Several articulatory models have been proposed (e.g. [9], [11], [12]), but few are based on multi-speaker data, and even fewer are based on cross-linguistic data ([15] and [16] are exceptions, see [6] for a review). This model fills the gap with an articulatory model based on multi-

subject, cross-linguistic data from vowels in four unrelated languages. With a cross-linguistic articulatory model, we may discuss cross-linguistic differences in vowel articulation quantitatively.

1.1 Measurement of articulators

This is a model of the mid-sagittal profile of the vocal tract during vowel production, since vowels only involve central articulations. Furthermore, there are algorithms for approximating the shape of the tongue given the mid-sagittal profile ([5], [8]). The scheme used to approximate the mid-sagittal profile of the vocal tract is described fully in [7]. 34 measures of articulator position are used: two of the epiglottis, 5 of the dorsal wall of the pharynx, 13 of the tongue, 12 of the velum and uvula, and the x- and y-coordinates of the lower incisor.

1.2 Speech Materials

The data used in this study were measured from x-rays in the literature. The data used to construct the model are from Akan (described in [10]), Arabic [1], Chinese ([13], [17]), and French [4]. The Akan vowels used are /i e e a o o u/ (one token of each vowel from four speakers), the Arabic vowels - pharyngealized and non-pharyngealized allophones of /i e a o u:/ (two tokens, two speakers), the Chinese vowels - /y e a u o/ (one, three), and the

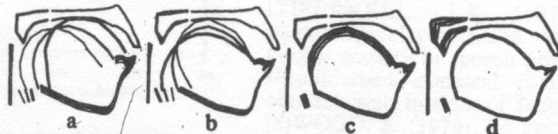


Figure 1. (a) The pattern of articulatory displacements produced by Front-Raising. (b) Back-Lowering. (c) Tongue Root Advancement. (d) Nasalization.

French vowels - /i e y # a o u o a e e o u a/ (one, four). Further details are given in [7]. Overall, the model accounts for 85% of the variance in articulator positions in these vowels.

1.3 Parameters of the model

The articulatory model has four parameters, called Front-Raising (FR), Back-Lowering (BL), Tongue Root Advancement (TRA), and Nasalization (N). The articulatory effects of these four parameters are shown in Figure 1, in the vocal tract outline of the second Akan speaker in [10]. The mean articulatory position is plotted with a medium-weight line, the articulatory configuration given by a negative displacement is plotted with a fine line, and the articulatory displacement is in bold.

The vowels of Arabic and French are plotted in the parameter space defined by FR, BL, and TRA in Figures 2 and 3 (axes normalized to unit variance). The left-hand plots show the vowels projected into the FR/BL plane, and the right-hand plots show them in the FR/TRA plane.

Articulatory FR is similar to traditional front/back. The pharyngealized vowels (open symbols) of Arabic and the low vowels of French tend to positive contributions from BL, whereas non-pharyngealized and high vowels have small contributions from BL. BL is thus similar to acoustic height. TRA patterns similarly to height in Arabic, but does not show a clear pattern in French. At best, it tends to separate nasalized and/or lax vowels from their oral and/or tense counterparts. In Akan, this parameter separates the [+ATR] vowels /e o u/ from the [-ATR] vowels /e a o u/ (see [7]).

This model embodies various articulatory and phonological constraints which we will not discuss here. [7] shows that FR and TRA involve no velum displacement, but there is an apparently irreducible correlation between BL and velum displacement. The TRA parameter is not used in Chinese; BL and TRA are used differently in Akan and Arabic.

2.0 GENERALIZABILITY

Data from consonants is used to test the generalizability of the model. X-ray

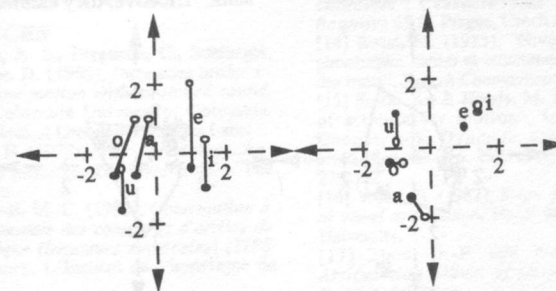


Figure 2. (left) Arabic vowels in the FR/BL plane; pharyngealized allophones in open circles, non-pharyngealized in solid. (right) Arabic vowels in the FR/TRA plane.

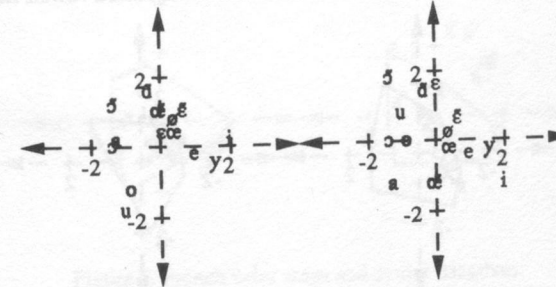


Figure 3. French vowels in the model articulatory space.

tracings of velar and uvular stops in several contexts were measured and fit to the model. Arabic uvular and velar stops from [2] & [3] in the contexts /ka ki ku qa qi qu/ (two speakers), Arabic uvular fricatives in the contexts /ra ri ru xa xi xu/ (one), and French velar stops and uvular fricatives in the contexts /ku ky gu gy ru re/ (four) were used. Speaker normalization factors were determined by fitting the articulatory model to the vowels /aa: ii: uu:/ for the two Arabic speakers; for the French speakers, speaker normalization was determined earlier (see [7]).

If the model generalizes well to these consonants, then it should describe them as well as it describes vowels. Quantitatively, it should account for the same proportion of variance as it accounts for in the vowels.

2.1 Arabic back consonants

When fit to the vowels from the two Arabic speakers in [2] and [3], the model accounts for 87% of the variance in articulator position. It fits 90% of the variance in the velar and uvular stops, but only about 80% in the uvular fricatives.

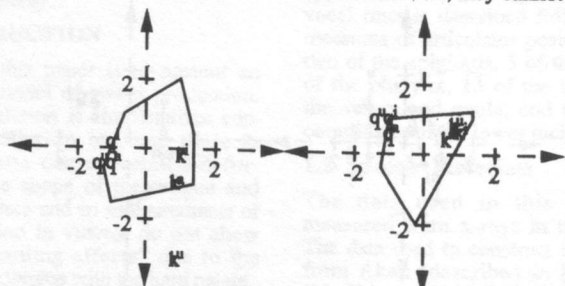


Figure 4. Arabic velar and uvular stops. Superscripts indicate vowel context. The polygons outline the area in which the modeled vowels lie.

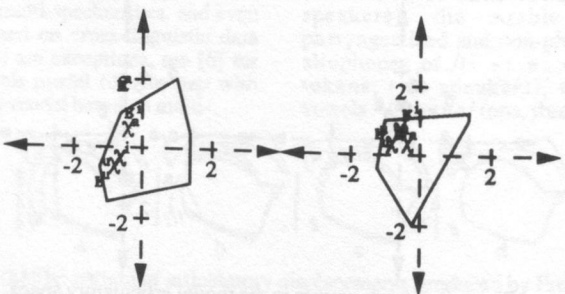


Figure 5. Arabic uvular fricatives.

The parameterization of the stops is shown in Figure 4. The stops fall in the outskirts of the articulatory space occupied by vowels. The fricatives, in Figure 5, although towards the periphery, generally lie within the vowel space.

2.2 French back consonants

In French, the model accounts for 72% of the variance in the vowels. This is rather low, but on the other hand, these vowels are part of the sample which was used to construct the original model in [7]. The model also fits 72% of the variance in the velar stops, but only fits 56% of the variance in the uvular fricatives. The articulatory parameterization of these consonants is shown in Figure 6.

3.0 DISCUSSION

Uvular and velar stops are well-described by the model in both Arabic and French. Their articulation is vowel-like, but more extreme.

The uvular fricatives pattern differently. The French ones have substantial individual variation (see [4], p. 229). Since the speakers did not produce "the same" fricative, they cannot be modeled

uniformly by this model.

The Arabic uvular fricatives use a mode of velum displacement that is not found in the other data. In the x-rays, the velum appears to "bulge" and descend (see [3], p.100 ff.) This mode of velum displacement cannot be modeled by the N parameter of the model. If the measures of velum position are excluded, the model fits 87% of the variance in articulator position - comparable to the fit to vowels. We conclude that tongue, etc., positions in the uvular fricatives should be described with the same phonetic and phonological features as used for vowels.

Coarticulatory variation in the articulation of stops and fricatives in different vowel contexts is clearly visible in Figures 4-6. This model gives us a method for articulatorily quantifying contextual coarticulation, which despite its name has usually been measured acoustically.

ACKNOWLEDGEMENTS

This work was supported in part by NSF Grant BNS-8713522, and PHS Grant 8 T32 DC 000005 15 to the Speech Communication Group at MIT.

REFERENCES

- [1] Abramson, A. S., Ferguson, C., Schlaeger, R., & Zeichner, D. (1962), *Damascus arabic x-ray film in slow motion with stretched sound*. New York, Columbia University, Columbia Presbyterian Medical Center and Haskins Labs.
- [2] Al-Ani, S. H. (1970), *Arabic phonology: An acoustical and physiological investigation*. The Hague, Mouton.
- [3] Boffi-Dkhissi, M.-C. (1983), *Contribution à l'étude expérimentale des consonnes d'arrière de l'arabe classique (locuteurs marocains) (TIPS 15)*. Strasbourg, L'Institut de Phonétique de Strasbourg.
- [4] Bothorel, A., Simon, P., Wioland, F., and Zerling, J.-P. (1986), *Cinéradiographie des voyelles et consonnes du français*. Strasbourg, L'Institut de Phonétique de Strasbourg.
- [5] Hashimoto, K., and Suga, S. (1986), "Estimation of the muscular tensions of the human tongue by using a three-dimensional model of the tongue", *J. Acoust. Soc. Japan (English)* 7, 39-46.
- [6] Jackson, M. T. T. (1988), *Phonetic theory and cross-linguistic variation in vowel articulation*. Ph.D. thesis, Los Angeles, UCLA.
- [7] Jackson, M. T. T. (1991), "A cross-linguistic articulatory model of vowels", submitted to *J. Acoust. Soc. Amer.*
- [8] Kakita, Y., Fujimura, O. and Honda, K. (1985), "Computation of mapping from muscular contraction patterns to formant patterns in vowel space", in *Phonetic linguistics*, ed. V. Fromkin, Orlando FL, Academic Press, 133-144.
- [9] Liljencrants, J. L. (1971), "A Fourier series description of the tongue profile", *STL / QPSR* 4, 9-18, Stockholm, KTH.
- [10] Lindau, M. (1979), "The feature expanded", *J. Phonetics* 7, 163-176.
- [11] Lindblom, B. & Sundberg, J. (1971), "Acoustical consequences of lip, tongue, jaw, and larynx movement", *J. Acoust. Soc. Amer.* 50, 1166-1179.
- [12] Mermelstein, P. (1973), "Articulatory model for the study of speech production", *J. Acoust. Soc. Amer.* 53, 1070-1082.
- [13] Ohnesorg, K., and Svarny, O. (1955), "Études expérimentales des articulations chinoises", *Ceskoslovenské Akademie Ved: Rozpravy* 65 5. Prague, Czech Academy.
- [14] Rossi, M. (1983), "Niveaux de l'analyse phonétique: nature et structuration des indices et des traits", *Speech Communication* 2, 91-106.
- [15] Shirai, K., & Honda, M. 1978. "Estimation of articulatory motion", in Sawashima & Cooper, eds. *Dynamic aspects of speech production*, Tokyo, University of Tokyo Press, 279-302.
- [16] Wood, S. (1982), *X-ray and model studies of vowel articulation*, Ph.D. thesis, Lund, Lund University.
- [17] Zhou, D.-F. and Wu, Z.-J. (1963), *Articulation album of putonghua*. Beijing, Commercial Press.

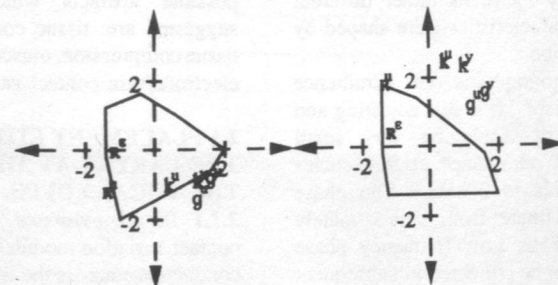


Figure 6. French velar stops and uvular fricatives.

WHAT DOES THE LARYNGOGRAPH MEASURE?

D. Miller(1), S. Nevard(2), R. Brun(2) and A.J. Fourcin(2)

(1) Laryngograph Ltd (2) University College London

ABSTRACT

A review of the literature reveals, in the case where the electrodes are placed on the wings of the thyroid cartilage, that this question has been investigated and broadly answered on several levels: electrically, physiologically and acoustically. We note results obtained from the Laryngograph at a new site: the nose. We focus on the boundaries of the Laryngographs use in order to highlight areas of possible improvement to the Laryngograph. We try to focus on questions of acoustic/phonetic (and to a lesser extent clinical) relevance.

1. ELECTRICALLY

The Laryngograph was originally developed as a means to accurately and non-invasively measure and display, in real time, the fundamental frequency of voicing, the acoustic correlate of intonation. Developed on the basis of an impedance measuring technique pioneered by Fabre its rather different electrical characteristics were shaped by this application.

The Laryngograph is an admittance variation sensor. It is self adjusting and is capable of displaying very small variations of admittance at frequencies between 10Hz to 20kHz. The phase response is linear from approximately 50Hz to 20kHz. Low frequency phase distortion can be corrected in subsequent processing. Versions are available with a phase and frequency response down to d.c.. Admittance variations can be

caused by resistive and reactive effects and the relative importance of each depends on what is between the electrodes e.g. simply holding the electrodes in air and shaking them will give a "waveform" but here the very high resistive path between the electrodes will ensure the capacitive effect dominates; when the electrodes have a relatively low conductive path between them (such as on the neck) the resistive component will dominate. The current laryngograph is not designed to distinguish resistive and capacitive effects, nor is it able to make absolute measurements.

2. PHYSIOLOGICALLY

When applied to the body our question can only be answered in relation to a knowledge of what is physically, physiologically and where applicable, phonetically possible between the electrodes at a specific site. Some possible artifacts which have been suggested are: tissue contact, vibration, tissue compression, muscular contraction, electrode/skin contact variation.

2.1 PLACEMENT EITHER SIDE OF THE LARYNX AT THE LEVEL OF THE VOCAL FOLDS.

2.1.1 Direct evidence that vocal fold contact variation modulates the electrical conductance across the larynx was found in a study of the transglottal impedance prior to the development of the Laryngograph: there was greater

conductance across the larynx in a static closed glottal state than when the folds were apart (contrary to Fabre's observations). The percentage change was increased when a three terminal (guard ring) electrode system was used to reduce the effect of skin conduction. The most conclusive evidence that the Laryngograph measures vocal fold contact area variation is provided in an experiment by Gilbert [6]. A 5mm wide insulating polyethylene strip was slowly withdrawn from between the vibrating vocal folds during phonation; the simultaneously recorded Laryngograph output waveform (Lx) increased as the area of obstruction to vocal fold contact decreased, the speech waveform remained relatively constant throughout. This experiment effectively eliminates all other variables except those which relate to the nature and area of vocal fold contact variation, for a steady state vowel in normal conditions, it falsifies Smith's [10] claim that tissue compression consequent on the acoustic pressure wave and the relative tenseness of the contracted musculature are the sole causes of Lx; these factors obviously can produce waveforms but all explanations must be site specific.

2.1.2. Correlations between the laryngographic waveform and direct observation of the detail of vocal fold contact cycle have been made.

Viewing down on the vocal folds: Fourcin, Donovan and Roach using a cine camera and stroboscope synchronised to the Lx waveform; Childers [2] using synchronised ultra high speed laryngeal films and Laryngograph waveforms.

Viewing the vocal folds from the front Noscoe et. al.[7] used X-flash imaging with a specially developed high voltage x-ray technique.

The results of these analyses broadly confirm the schematisation made by Fourcin [5] and, in a slightly different way, by Rothenberg [8]. Childers [2] cautions that the model's features tend to be inferred from research observations rather than from a compilation of statistical data extracted from experimental measurements, but their own work reaches the same broad conclusion.

Computer simulations of the vocal fold area contact (VFCA) variation have produced waveforms similar to Lx, and enable the effect of various features of vocal fold action to be modelled independently Titze [11], Childers [2].

2.2. ELECTRODES ACROSS THE BRIDGE OF THE NOSE

If the electrodes are placed either side of the bridge of the nose a larynx-periodic waveform can be observed which varies in amplitude during phonation according to the degree of nasalisation [1]. We currently have no direct evidence of the physiological cause of this waveform. It would be reasonable to suggest the amplitude is modulated by the opening of the velum, and that the coupling of the nasal cavities causes them to respond and change their geometry.

3. ACOUSTIC CORRELATIONS OF LX

Correlations have been made with the inverse filtered flow waveform by Fourcin and by Rothenberg. These confirm an approximately antiphase relationship between Lx and the glottal flow waveform. Acoustically the abrupt cessation of the airflow is of primary significance and this is clearly correlated with the rapid closure of the vocal folds evident in the Lx waveform.

4. BOUNDARIES OF LX APPLICATION

Vocal fold vibration is a complex three dimensional wavelike motion in response to which the laryngograph can only give an integrated two dimensional representation of the varying effects of the nature and area of vocal fold contact. This sets the limits to any physiological and acoustic interpretation of the Lx waveform although for some analyses its simplicity is a positive advantage. In addition the vocal fold element has to be extracted from other gross physiological movements such as of the larynx as a whole.

For its original purpose of providing an accurate means to measure and display the fundamental frequency (Fx) of voiced speech on a period by period basis the Laryngograph has become a reference. The limitations of Lx as the basis for a measure of Fx and voicing appear in cases where an oscillatory volume velocity flow is observable but Lx is not. This can occur in laryngeal co-articulation (e.g. [ehe]) where the vocal folds adduct to such an extent that they no longer actually contact although they still produce an oscillatory airstream.

It is generally agreed that closure is one of the most clearly defined events in the Lx waveform [5][8], although it must be emphasised that an Lx peak does not necessarily denote full closure along the length of the vocal folds. A "chink" may remain in breathy voice. There is interest in the measurement of vocal fold closed phase (or some related ratio to open phase, or related duty cycle): for closed phase formant analysis; in the study of the structured variability of voiced excitation (for analysis and synthesis of more natural speech [4]; in the related study of relative vocal fold abduction [9]; in the study of voice quality and pathology (Fourcin 1990); and in the study of vocal efficiency in trained singers (Howard.D 1990).

The problem here is the definition

of vocal fold opening. By its nature this is a less well defined event: typically the vocal folds peel apart from the bottom up and then split along the upper edge. Correlations have been made with the "knee" which is sometimes apparent in the opening edge of the waveform and the breaking of the mucus bridge on the upper edge of the folds and with the onset of airflow. This is the justification for one of the practical measures that have been applied to the measurement of opening [4]; (1) the maximum rate of change on the opening edge. Other measures chosen are: (2) when the voltage on the opening edge reaches the same level as existed when the closure point was chosen; (3) the point where the opening edge falls to some fixed percentage of the peak voltage or (4) some average level [9]. Given the indeterminacy of this feature, decisions are necessarily pragmatic. Generally a percentage fall from a peak value is more reliable on the opening edge than the differential.

In pathology all problems are likely to be compounded. Fourcin [5] suggests five key features of the Lx waveform, useful when using Lx as an analytical tool in the physical interpretation of aspects of voice: Uniformity of peaks = uniform output; Sharply defined Lx contact = good excitation of vocal tract; long closure duration = well defined formants; regular, sharply defined contact periodicity = well defined pitch; progressive change in sharply defined period lengths = smoothly changing pitch.

Childers [2] appears to be investigating more of a "template" approach. The use of computer simulation is obviously attractive here. Colton and Conture [3] warn of the dangers of interpreting pathological data on the basis of a model of normal vocal fold vibration. All

waveform interpretation must be based on a knowledge of the phonetically structured variability which is a property of normal speech.

5. CURRENT WORK

Our original intention, prompted by the nasal results and a desire to quantify effects at the larynx, was to review what the Laryngograph was measuring by making some comparisons of its sensitivity to factors which can affect the impedance between the electrodes at various sites on the body. This could be seen as a quantitative version of Smiths experiments. It is evident that the following factors need to be taken into account:

1. All explanations must be site specific.
2. Gilberts experiment and the X-flash observations show the overwhelming effect of vocal fold contact.

3. Further consideration of the problems of using signals in the existing Laryngograph circuit to measure absolute or even relative impedance would make intersite comparisons difficult.

We have therefore decided to concentrate on the neck and nose.

At the neck the main limitation to Laryngographic measurement is that in some cases the signal to noise ratio is poor. We are conducting a series of measurements relating the output of the Laryngograph on the neck to absolute impedance measures on a range of subjects in order to determine the correlations between larynx size, the effects of varying amounts of subcutaneous fat and the effects of different current levels.

Measures on the nose can similarly be made to clarify the result here.

Results of these and related studies will be presented at the conference.

REFERENCES

- [1] BRUN R, SPENCER C, FOURCIN A, "Nasalisation detection using the

electrolaryngography principle", *Speech, Hearing and Language, Work in Progress*, (4),59-62.UCL, London.

[2] CHILDERS, D G, HICKS, D M, MOORE, G P, ESKENAZI, L, and L A L W A N I, A L. (1990) "Electroglottography and vocal fold physiology", *Journal of Speech and Hearing Research*, 33, 245-254.

[3] COLTON R H, CONTURE G C, 1990, "Problems and pitfalls of electroglottography", *Journal of Voice*,4,10-24.

[4] DAVIES P, LINDSEY G, FULLER H, FOURCIN A.(1986),"Variation in glottal open and closed phase for speakers of English", *Proc. Inst. Acoustics*.

[5] FOURCIN A J,(1981), "Laryngographic assessment of phonatory function" in C L LUDLOW

Conference on the Assessment of Vocal Pathology, Maryland:ASHA Reports 11.

[6] GILBERT,H R, POTTER, C R, HOODIN, R:(1984) "The Laryngograph as a measure of vocal fold contact area", *Journal of Speech and Hearing Research*,27,178-182.

[7] NOSCOE, N J, FOURCIN, A J, BROWN, M A, and BERRY, R J. (1983), *British Journal of Radiology*,56, 641-645.

[8] ROTHENBERG, M.(1981),"Some relations between glottal airflow and vocal fold contact area"

in C L LUDLOW *Conference on the Assessment of Vocal Pathology*, Maryland:ASHA Reports 11.

[9] ROTHENBERG M, MASHIE J,(1988),"Monitoring vocal fold abduction through vocal fold contact area"*Journal of Speech and Hearing Research*,31,338-351.

[10] SMITH, S:(1981),"Research on the principle of electroglottography", *Folio Phoniatica*, 33:105-114.

[11] TITZE R T,(1990),"Interpretation of the Electroglottographic signal"*Journal of Voice*,4, 1-9.

DES PARAMETRES FORMANTIQUES AU PROFIL ARTICULATOIRE

P. Jospa

Institut de Phonétique, Université Libre de Bruxelles,
50, av Franklin Roosevelt, 1050-Bruxelles,
Belgique.

ABSTRACT

A method for calculating the parameters of an articulatory model [6] is presented. This method takes the formant values and data of labial articulation as its starting point. The method proceeds via the minimisation of the "articulatory effort" under the constraints imposed by articulatory and acoustic data. The expression of these acoustic constraints is based on a variational formulation of the relation linking the resonance modes (formants) to the area function [5]. This formulation takes account of the dissipative effects of the tract and of the lip-spread. Four variants of the procedure are evaluated, with articulatory profiles published in the literature serving as references. In most cases it seems that the simplest version gives the most satisfactory results.

1. INTRODUCTION.

Un modèle articuloire qui rend compte de la position des principaux articulateurs, présente, vis-à-vis de la fonction d'aire, une redondance de degrés de liberté nécessaire pour produire les phénomènes de coarticulation et de compensation entre articulateurs. Un calcul fiable des paramètres du modèle à partir du signal de parole ne peut se concevoir sans disposer d'informations complémentaires de nature acoustique ou articuloire. Dans quelle mesure l'usage d'une stratégie générale de coordination articuloire autorise-il un tel calcul, lorsqu'à côté du signal de parole, la donnée de l'articulation labiale est également fournie?

Nous avons développé une procédure de calcul des paramètres d'un modèle articuloire à partir des trois premières fréquences formantiques, de la donnée optionnelle des largeurs de bande des 2ème et 3ème formants, de la donnée de l'ouverture labiale, ainsi que d'un principe

de moindre "effort articuloire". Le problème est posé en ces termes: satisfaire le principe de coordination articuloire (moindre effort), les contraintes imposées par les données acoustiques et articuloires étant au mieux respectées. Notre méthode de calcul fait un usage direct des données formantiques. Outre la donnée de l'ouverture aux lèvres, la protrusion labiale et l'ouverture de la mâchoire peuvent également être fournies, auquel cas la procédure se limite à la détermination du profil lingual. La distance glotte-incisives est supposée connue. Le modèle articuloire choisi est celui de Maeda [6]. Il comporte 6 paramètres: le degré d'ouverture de la mâchoire, les degrés d'ouverture et de protrusion labiales, et trois paramètres contrôlant l'articulation linguale, à savoir: la position (avant/arrière) du corps de la langue, la forme (arquée/plate) du dos de la langue, et la position (élevée/abaissée) du sommet de la langue.

2. CALCUL DES PARAMETRES FORMANTIQUES.

Le modèle acoustique adopté est celui du conduit vocal dissipatif proposé par Sondhi [10]; il est caractérisé par une admittance des parois proportionnelle à la fonction d'aire et un facteur de forme constant. La distribution d'amplitude ψ et la fréquence ω d'un mode de résonance satisfont l'équation suivante:

$$(1) \quad \partial_x(A\partial_x\psi) + \frac{1}{c^2}(\omega^2 - \omega_p^2)A\psi = 0,$$

avec $A(x)$: la fonction d'aire, c : la vitesse du son dans le conduit vocal, et ω_p : la fréquence de vibration des parois ($2\pi \cdot 200$ hz environ). A l'extrémité glottique, nous imposons la condition:

$$(2) \quad \partial_x\psi = 0, \quad \text{en } x=0,$$

qui correspond à une impédance infinie à la glotte. A l'extrémité labiale, soit en $x=L$, le traitement classique consiste à calculer une correction de longueur ΔL :

$$(3) \quad \Delta L = \frac{3}{8} \sqrt{\pi A(L)}$$

et à imposer à l'amplitude ψ du mode, la condition:

$$(4) \quad \psi(L') = 0, \quad \text{avec } L' = L + \Delta L.$$

Cette condition est bien fondée dans le cas du premier mode. Un traitement différent, fondé jusqu'au troisième mode, consiste à imposer la condition suivante aux lèvres [5]:

$$(5) \quad A\partial_x\psi + q\sqrt{A}\psi = 0, \quad \text{en } x=L,$$

avec q une constante de l'ordre de 2,1.

Le calcul des fréquences ω et des fonctions ψ qui satisfont à l'équation (1), sous les conditions (2) et (4) ou (5), représente un problème aux valeurs propres classique. Ce problème peut recevoir une formulation variationnelle équivalente [4],[5] qui présente certains avantages. Etant de nature intégrale, cette formulation variationnelle conduit à des solutions qui sont liées plus au caractère général qu'au comportement détaillé de la fonction d'aire. Elle permet l'usage des méthodes directes du calcul des variations [3] --telle la méthode de Rayleigh-Ritz utilisée dans ce travail--, qui substituent au problème initial, un problème de valeurs propres de l'algèbre linéaire, de dimension généralement faible, plus simple à résoudre. Enfin, la formulation variationnelle permet d'approcher analytiquement le lien qui uni la fonction d'aire aux fréquences des formants. Dans le cadre de la méthode de Rayleigh-Ritz, et pour un modèle articuloire donné, ce lien prend la forme d'équations algébriques non linéaires reliant les fréquences formantiques f_n aux paramètres articuloires $\{a\}$ [4],[5]:

$$\det\{K(\{a\}) - \lambda_n V(\{a\})\} = 0,$$

$n=1, \dots, N$

avec: $\lambda_n = (2L/c)^2 \cdot (f_n^2 - f_p^2)$

où f_n désigne la fréquence du n ème formant, et où K et V , sont des matrices carrées, symétriques, ne dépendant que

des paramètres articuloires. La définition et le calcul de ces matrices sont exposés ailleurs [5]. L'ordre M de ces matrices peut être relativement faible: $M=8$ dans nos calculs, pour chacun des trois premiers formants. Une fois construites les matrices K et V , la détermination des premières fréquences propres se ramène au calcul des premières racines de l'équation:

$$(6) \quad \det\{K_{l,m} - \lambda V_{l,m}\} = 0.$$

Le calcul des largeurs de bande exige de connaître non seulement la fréquence, mais aussi la distribution spatiale du mode considéré. Une fois calculée la racine λ_n de l'équation (6), le calcul de ψ_n ne fait guère de difficulté [5].

3. CALCUL DES PARAMETRES ARTICULOIRES.

Désignons par a^0_k les premières approximations (ou valeurs initiales) des paramètres articuloires, par $f_n(\{a\})$ et $B_n(\{a\})$ les paramètres formantiques calculés pour un profil articuloire $\{a\}$, et par f_n^* et B_n^* les données formantiques. Désignons encore par K le nombre de paramètres articuloires à calculer, et soit, finalement, p_k et q_k ($k=1, \dots, K$), des pondérations convenablement choisies. La méthode consiste à minimiser une fonctionnelle E des articulateurs ("l'effort articuloire"), de la forme:

$$E(\{a\}) = \sum_{k=1}^K (p_k |a_k - a^0_k| + q_k |a_k|)$$

sous les contraintes acoustiques suivantes:

$$(f_n(\{a\}) - f_n^*) / f_n^* = 0, \quad n=1,2,3$$

$$(B_n(\{a\}) - B_n^*) / B_n^* = 0, \quad n=2,3$$

et sous les contraintes articuloires:

$$-3 \leq a_k \leq 3 \quad k=1, \dots, 4$$

$$0.2 \text{ cm} \leq a_5 \leq 2.4 \text{ cm}$$

$$A_{\text{min}} > 0.3 \text{ cm}^2$$

La procédure adoptée pour résoudre ce problème d'optimisation sous contraintes repose sur une méthode de minimisation séquentielle sans contrainte, faisant usage de fonctions de pénalisation [1]. Soit:

$r_j = (0.1)^j r_0$, ($j=0, 1, \dots, J$), une séquence décroissante de nombres positifs ($r_0 = 5$), telle que $r_j > r_{\text{min}}$ ($r_{\text{min}} = 0.001$).

Nous construisons une suite de $J+1$ systèmes de K équations non linéaires à K inconnues, à résoudre séquentiellement, de

manière optimale au sens des moindres carrés. L'algorithme de Powell [9], utilisé à cet effet, s'est avéré rapide et robuste. La solution obtenue au terme de cette procédure, optimise le respect des contraintes, tout en minimisant l'effort articuloire E . Des exemples de résultats sont illustrés par les figures 1 à 7.

4. EVALUATION.

Une évaluation informelle de la procédure a été réalisée en utilisant les données (fonctions d'aire et fréquences formantiques) publiées par Fant [2] (6 voyelles russes) et par Mrayati [7] (11 voyelles françaises). En l'absence de données fiables relatives aux largeurs de bande, nous utilisons comme données, les largeurs de bande calculées pour les fonctions d'aire publiées. Les données des fréquences formantiques sont celles qui sont publiées par les auteurs: mesurées sur sonogrammes pour Fant, calculées selon le modèle de la ligne de transmission pour Mrayati. Les tableaux I et II résument les résultats obtenus selon que l'une ou l'autre des conditions (4) et (5) est appliquée aux lèvres et selon que les largeurs de bande des formants 2 et 3 sont prises ou non en compte. Quatre versions de la procédure sont ainsi obtenues; leurs résultats sont comparés en fonction de leur degré d'accord entre les fonction d'aire calculées et les fonctions d'aire données. La première colonne indique la version de la procédure, soit:

- Version 1: la condition (4) est adoptée aux lèvres et les largeurs de bande ne sont pas prises en compte.

- Version 2: la condition (4) est adoptée aux lèvres et les largeurs de bande sont prises en compte.

- Version 3: la condition (5) est adoptée aux lèvres et les largeurs de bande ne sont pas prises en compte.

- Version 4: la condition (5) est adoptée aux lèvres et les largeurs de bande sont prises en compte.

La colonne 2 indique la moyenne sur toutes les voyelles v de l'écart logarithmique moyen entre la fonction d'aire calculée A_v et la fonction d'aire de référence A_{vr} :

$$\frac{1}{N_v} \sum_{v=1}^{N_v} \frac{1}{N_p} \sum_{i=1}^{N_p} |\log(A_v(x_i)) - \log(A_{vr}(x_i))|,$$

N_v désignant le nombre de configurations et N_p le nombre de sections par configuration. La colonne 3 indique l'écart absolu moyen, en hertz, pour toutes les voyelles, entre les valeurs calculées et les valeurs données des 3 premières fréquences formantiques. La colonne 4 indique un écart relatif moyen pondéré, en pour-cent, entre les mêmes valeurs formantiques.

Tableau I: données de Fant [2], moyennes sur 6 voyelles russes.

Version:	Ecart log. aires	Fréquences:	
		Eabs(hz)	Erel(%)
1	0.41	87.3	5.30
2	0.46	98.4	7.83
3	0.41	87.1	6.92
4	0.42	95.0	7.40

Tableau II: données de Mrayati [7], moyennes sur 11 voyelles françaises.

Version:	Ecart log. aires	Fréquences:	
		Eabs(hz)	Erel(%)
1	0.44	71.6	4.66
2	0.37	88.1	6.83
3	0.44	59.6	4.95
4	0.47	58.8	4.83

On ne constate pas différences nettement significatives entre les différentes versions. La version 1 réalise l'une des meilleures performances. Etant la moins coûteuse en temps de calcul, et ne demandant pas la connaissance des largeurs de bande, elle semble devoir s'imposer.

Les résultats présentés appartiennent au cas où l'aire d'ouverture et la protrusion labiales sont données; le nombre K de paramètres articulatoires calculés est donc égal à 4, soit l'ouverture de la mâchoire et les trois paramètres de la langue. Des résultats similaires ont été obtenus lorsque seule l'aire de l'ouverture aux lèvres est donnée ($K=5$), et lorsqu'en plus des paramètres labiaux, l'ouverture de la mâchoire est donnée ($K=3$).

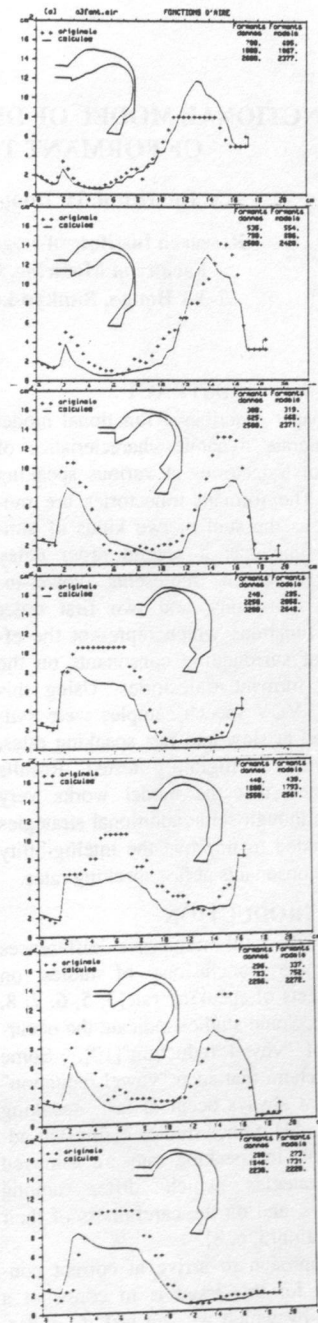
L'influence de la première approximation sur la fonction d'aire calculée a été testée, en adoptant, systématiquement, comme première approximation de l'articulation linguale, les valeurs du profil neutre. Les performances moyennes observées ne sont pas significativement modifiées.

5. CONCLUSION.

Nous pensons avoir développé une méthode robuste et relativement économique en temps de calcul pour estimer les paramètres d'un modèle articuloire, à partir des fréquences des trois premiers formants et de la donnée de l'articulation labiale. Appliquée au modèle de Maeda, elle permet une estimation des paramètres de l'articulation linguale. La donnée des largeurs de bandes formantiques ne semble pas utile, et peut, dans certains cas, avoir un effet perturbateur. La condition aux lèvres (5), plus complexe à traiter que la condition (4), ne conduit pas à des performances supérieures; la version 1 de la procédure semble donc pouvoir s'imposer.

REFERENCES:

- [1] ADBY, P.R. et DEMPSTER, M.A.H. (1974), "Introduction to Optimizaton Methods", London: Chapman and Hall.
- [2] FANT, G. (1960), "Acoustic theory of speech production", The Hague: Mouton & Co.
- [3] GOULD, S.H. (1966), "Variational Methods for Eigenvalue Problems", London: Oxford University Press.
- [4] JOSPA, P. (1972) "Forme approchée du conduit vocal déduite des fréquences de résonance. Théorie des perturbations et méthode variationnelle", Lannion: JEP, GALF, 225-262.
- [5] JOSPA, P. (1977), "Théorie acoustique du conduit vocal de forme variable dans le temps", Thèse de Doctorat, Université Libre de Bruxelles.
- [6] MAEDA, S (1979), "Un modèle articuloire de la langue avec des composantes linéaires", 10èmes JEP, GALF, 152-164.
- [7] MRAYATI, M. et CARRE, R. (1976), *Phonetica*, 33, 285-306.
- [9] POWELL M.J.D. (1970), "A fortran subroutine for solving systems of nonlinear algebraic equations", in: "Numerical methods for nonlinear algebraic equations", RABINOWITZ, Ph. Ed., London: Gordon and Breach.
- [10] SONDDHI, M.M. (1974), *J. Acoust. Soc. Am.*, Vol.55, 5, 1070-1075.



A FUNCTIONAL MODEL OF DYNAMIC CHARACTERISTICS OF FORMANT TRAJECTORIES

S. Imaizumi, H. Imagawa and S. Kiritani

Research Institute of Logopedics and Phoniatrics
Faculty of Medicine, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 JAPAN

ABSTRACT

This paper describes a functional model to generate dynamic characteristics of formant trajectories at various speaking rates. The formant trajectories are modelled as the sum of two kinds of temporal functions: a second order delay function which represents vowel-to-vowel transitions, and two first order delay functions which represent the effects of surrounding consonants on the vowel formant trajectories. Using this model, VCV speech samples were synthesized at slow and fast speaking rates, and their intelligibility tested. Results suggested that the model works very well, although some additional strategies are needed to improve the intelligibility of the consonants at fast speaking rates.

1. INTRODUCTION

There are still numerous differences among the conclusions of studies on the effects of speaking rate [3, 5, 6, 7, 8, 10, 11]. Some studies indicate the occurrence of "vowel reduction" [10]. Some others claim that such "vowel reduction" does not always occur at fast speaking rates [5, 11]. Other studies claim that adjustments in speaking rate are achieved by strategies which differ among speakers, and on the carefulness of their articulation [3, 6, 8].

One approach to arrive at correct conclusion for this issue is to construct a model, by which we can test if under-shoot or reorganization is necessary or

not to generate high quality speech at various speaking rates. Although some models have been proposed to generate formant transitions [1, 2, 7, 9], there are still problems remaining to be solved to generate natural formant trajectories.

In this paper, we proposed a functional model which describes the formant transitions as the sum of two kinds of temporal functions: one represents vowel-to-vowel transitions, and the other represents CV or VC transitions. The model was assessed via an intelligibility test.

2. METHOD

2.1 Formant Transition Model

The trajectory of n th formant, $F_n(t)$, in a vowel segment is assumed to be expressed as

$$F_n(t) = U_n(t) - C_{nf}(t) - C_{np}(t) \quad (1)$$

Here, $U_n(t)$ is the step response of a second order delay function which represents vowel-to-vowel transition, $C_{np}(t)$ is the first order delay function which represents the effect of a preceding consonant, and $C_{nf}(t)$ is the first order delay function which represents the effect of a following consonant.

To generate $U_n(t)$, the putative target frequency R_{ij} of each vowel in the sequence $V_1 C_p V_2 C_f V_3$, ($i=1, 2, 3, j=1, 2, 3$) is assumed to be set at t_i as a step input. The suffix i represents vowel number, j formant number. For the back vowels /a, u, o/, j represents j th lower formant frequency.

For the front vowels /i, e/, $R_{i,1}$ is the lowest, $R_{i,2}$ the third, and $R_{i,3}$ the second. Let $W_j(t)$ represent the step response of a second order delay function, $W_j(t)$ can be expressed as

$$W_j(t) = R_{ij} + a_j(t)(R_{ij} - R_{i,j}), \quad (2)$$

$$a_j(t) = 1 - \{1 + b_j(t)\} \exp(-b_j(t)) u(t-t_i),$$

$$b_j(t) = (t-t_i)/g_j,$$

$$u(t-t_i) = 1 \text{ for } t > t_i, = 0 \text{ for } t < t_i,$$

g_j : time constant representing transition speed.

For transitions from a back vowel to a front vowel or vice versa, $W_2(t)$ and $W_3(t)$ intersect each other. Such intersection never occur in actual speech due to the coupling between two resonance frequencies. Therefore the resonance frequencies $W_j(t)$ are modified accounting for the coupling between $W_2(t)$ and $W_3(t)$ as follows [4].

$$U_1 = W_1, U_2 = c\sqrt{(W_2 W_3)}, U_3 = \sqrt{(W_2 W_3)}/c, \quad (3)$$

$$c = \sqrt{e}, d = (W_2 W_2 + W_3 W_3)/W_2 W_3,$$

$$e = d - (dd - 4(1 - kk))/2(1 - kk), k = 0.2.$$

Two functions representing the effect of a preceding consonant $C_{np,i}(t)$ and of the effect of a following consonant $C_{nf,i}(t)$ upon the formant trajectories in the segment V_i are assumed as follows.

$$C_{np,i}(t) = c_{np,i} \exp\{(t-t_{p,i})/g_p\}, \text{ for } t_p < t_i, \quad (4)$$

$$C_{nf,i}(t) = c_{nf,i} \exp\{-(t-t_i)/g_f\}, \text{ for } t_{p,i} < t_i, \quad (5)$$

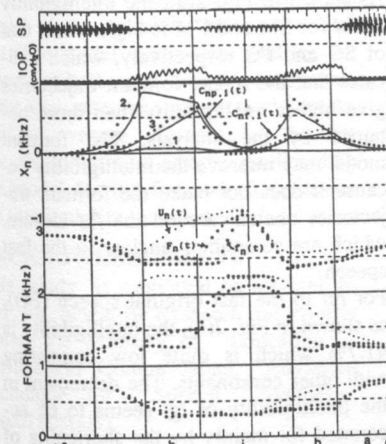


Fig.1(a). An example of formant trajectories $F_n(t)$, and those obtained by analysis $f_n(t)$ for a slow utterance of /abiba/.

$t_{p,i}$: initial time of vowel V_i ,
 $t_{f,i}$: final time of V_i ,
 g_p, g_f : time constant representing the decay speed.

In this paper, only the temporal parameters, t_i : onset time of the targets for vowel V_i , $t_{p,i}$: initial time of V_i and $t_{f,i}$: final time of V_i are changeable depending on the speaking rate. This means that this model does not take account of possible changes or "reorganization" in the vowel targets or other parameters such as g_p and g_f .

2.2 Model Parameters Estimation

The speech material used here consisted of $V_1 C_p V_2 C_f V_3$ samples, where V_1, V_2 were one of /a, i, u, r/, $V_1 = V_2$, and C_p or C_f was one of /b, d, g, p, t, k, r/. Those samples were recorded from two male speakers spoken at slow (S) and fast (F) speaking rates. The original utterances were /korewa $V_1 C_p V_2 C_f V_3$ desu/, that means, "this is $V_1 C_p V_2 C_f V_3$ ". Among five Japanese vowels, only three vowels /a, i, u/ were used as the typical examples of low-back, high-front, and intermediate vowels.

The details of recording, analyses, parameter estimation method were reported

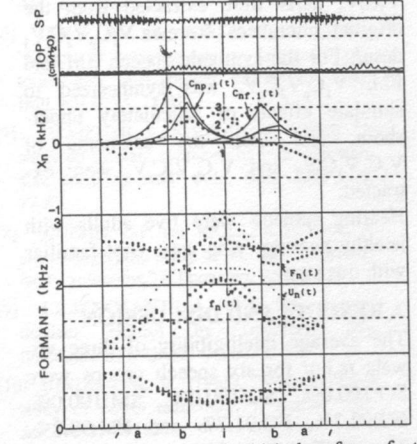


Fig.1(b). Same as Fig.1(a), but for a fast utterance of /abiba/.

in other place [6,7]. The formant trajectories were estimated mainly using the closed phase LPC analysis. The estimation of the model parameters was carried out based on a least square error method and interactive modification using only speech uttered slowly and clearly.

2.3 Intelligibility Test

To assess how well the model could generate formant trajectories, an intelligibility test was carried out for two kinds of synthetic speech (G and M), and the original speech samples (O) from which model parameters were extracted. Here, synthetic speech samples G were generated using the formant frequencies $f(t)$ obtained from SO by the analysis and the glottal source obtained from the polynomial glottal source model. Synthetic speech samples M were generated using the model formant trajectories $F(t)$ and the model glottal source. The speech samples were synthesized or recorded at two speaking rates, slow (S) and fast (F). Each group consisted of 84 $V_1C_pV_2$ samples, where V_1, V_2 were one of /a, i, u/, $V_1=V_2$, and C was one of /b, d, g, p, t, k, r/. For SO and FO, $V_1C_pV_2$ and $V_2C_pV_3$ parts were extracted from the original utterances /korewa $V_1C_pV_2C_pV_3$ desu/. For the synthetic speech SM and FM, $V_1C_pV_2C_pV_3$ was synthesized to simulate effects of articulately under-shoot, and then the segments of $V_1C_pV_2C_pV_3$ and $V_1C_pV_2C_pV_3$ were extracted.

Hearing subjects were five adults with healthy hearing who were not familiar with this study.

3. RESULTS AND DISCUSSION

The average intelligibility of three vowels /a,i,u/ for six speech groups were SO:100.0%, SG:100.0%, SM:100.0%, FO:92.7%, FG:91.7% and FM:93.8%. The intelligibility of FM is 93.8% which is better than those of FO and FG. Concerning the vowels, it is suggested that the formant model maintains or even

slightly improve the intelligibility compared to the original speech in slow and fast speaking rates.

On the other hands, the average intelligibility of the consonants /b,d,g,r/ were SO:91.7%, SG:81.3%, SM:83.3%, FO:79.2%, FG:68.8% and FM:62.5%. For the consonants, the use of model voicing source without plosion decreases the intelligibility about 10%. The use of the formant model slightly increases the intelligibility by about 2%(SM-SG). For the consonants in the slow speech, the formant model works well on average, and even slightly improve the intelligibility comparing with that of SG. However, for the fast speech, the formant trajectories predicted by the model for the fast speech decreases the intelligibility by about 6%.

There were, however, some differences among the intelligibility of the consonants, or the goodness of the model. Fig. 2 shows the intelligibility of the consonants /b,g/ for the six speech groups. The box-whisker graph in this figure shows the minimum, 25%-tile, median, 75%-tile, and maximum of the intelligibility scores.

As shown in Fig. 2(a), the intelligibility of /b/ for SM and FM is higher than that of SG and FG respectively, which indicates that the model formant trajectories give higher intelligibility than those obtained by the analysis. The formant model may improve the intelligibility because it does not make the formant trajectories unclear around the /b/ closure, which are sometimes unclear in the fast speech.

For /g/ in the fast original speech (FO), as shown in Fig. 2(c), the intelligibility is 41.7% which is quite low comparing with other consonants. The decrement in the intelligibility of /g/ seems to be accounted for mainly by the shortening of segments due to the speaking rate. Because the formant transition of /g/ is slower than that of other stops, the speak-

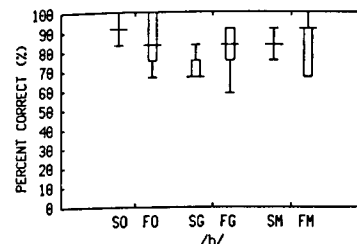


Fig.2(a) The intelligibility of /b/.

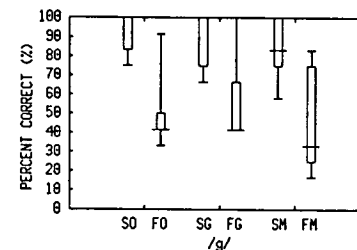


Fig.2(b) The intelligibility of /g/.

ing rate change may affect largely the intelligibility of /g/.

In this model, only the timing parameters are changeable depending on the speaking rate. It should be noted that the intelligibility of the consonants at fast speaking rates might be improved by taking account of possible changes in some parameters such as those representing transient speed which are set constant in this report.

4. CONCLUSION

This paper proposes a model of formant trajectories at various speaking rates, and reports on the intelligibility of VCV speech samples synthesized based on the model at two speaking rates, slow and fast. The model works very well for the vowels at both rates. For the consonants at the slow rate, the formant model works well on average. However, at the fast rate, the formant trajectories predicted by the model for the fast speech decreases the intelligibility by about 6%.

5. ACKNOWLEDGMENTS

This study is supported by a Grant-in-Aid for Scientific Research on Priority Areas, the Ministry of Education, Science and Culture, Japan, "Advanced Technique for Speech Synthesis" (No.01608003).

6. References

- [1] BROAD, D. J. & FERTIG, R. H. (1970), "Formant-frequency trajectories in se-

lected CVC utterances," J. Acoust. Soc. Am., 47, 1572-1582.

- [2] BROAD, D. J. & CLERMONT, F. (1987), "A methodology for modeling vowel formant contours in CVC context," J. Acoust. Soc. Am., 81(1) 155-165.
- [3] FLEGE, J.E. (1988), "Effects of speaking rate on tongue position and velocity of movement in vowel production," J. Acoust. Soc. Am., 84(3), 901-916.
- [4] FUJISAKI, H., YOSHIDA, M., SATO, Y. and TANABE, Y. (1974), "Automatic recognition of connected vowels using a functional model of the coarticulatory process," J. Acoust. Soc. Jpn, 29, 636-638.
- [5] GAY, T. (1978), "Effect of speaking rate on vowel formant movements," J. Acoust. Soc. Am., 63(1), 223-230.
- [6] IMAIZUMI, S., KIRITANI, S. (1989), "Effects of speaking rate on formant trajectories and inter-speaker variations," Ann. Bull. RILP, 23, 27-37.
- [7] IMAIZUMI, S., KIRITANI, S. (1990), "A study on formant synthesis by rule with variable speaking rate," Ann. Bull. RILP, 23, 77-87.
- [8] KUEHN, D.P. and MOLL, K.L. (1976), "A cineradiographic study of VC and CV articulatory velocities," J. Phonetics, 4, 303-320.
- [9] LILJENCRAFT, J. (1970), "Speech synthesizer control by smoothed step functions," STL-QPSR 4/1969, 43-50.
- [10] LINDBLOM, B. (1963), "Spectrographic study of vowel reduction," J. Acoust. Soc. Am., 44, 1773-1781.
- [11] van SON, R.J.J.H. & POLS, L.C.W. (1990), "Formant frequencies of Dutch vowels in a text, read at normal and fast rate," J. Acoust. Soc. Am., 44, 1683-1693.

COMPARISON OF THE MODIFIED HERMITE TRANSFORMATION WITH OTHER UNITARY TRANSFORMATIONS IN A PTC SCHEME

Victoria E. Sanchez, Jose C. Segura, A.M. Peinado,
Juan M. Lopez and Antonio J. Rubio

Dpto. de Electronica y Tcno. de Computadores
Facultad de Ciencias, Univ. de Granada
Granada (SPAIN)

ABSTRACT

Recently a Predictive Transform Coding (PTC) scheme has been proposed. This is a transform coding scheme with a strong link to the LPC model of speech production. In this paper several unitary transformations are studied within this scheme. These are the Discrete Cosine Transform, a unitary transformation resulting from applying the Singular Value Decomposition to the impulse response matrix of the LPC filter, the identity transformation and the recently developed Modified Hermite Transformation. We determine the number of parameters needed in this scheme for each transformation, in order to have a high quality synthesis and make both objective and subjective measures.

1. INTRODUCTION

In the last years residual speech coders have had a great development. In these kind of coders the speech signal is represented by the LPC filter and by the LPC residual as the excitation. Different representations of the LPC residual lead to different schemes, such as multipulse, CELP and others which attempt to represent the residual in a simpler manner [3].

Recently a unified framework for LPC excitation representation in residual speech coders has been presented [3]. Within this framework a new scheme has been proposed called Predictive

Transform Coding (PTC), which is a transform coding scheme with a strong link with the LPC model of speech production.

In this scheme we are going to study the performance of several unitary transformations, among them the recently developed Modified Hermite Transformation, and to determine the number of parameters needed for the excitation in order to have a high quality synthesis.

2. VECTORIAL EXPRESION OF THE LPC EXCITATION

Let's consider the excitation $x(n)$ of the LPC filter as a linear combination of a subset of vectors L taken from a given set V of M vectors of length N [3]

$$x = \sum_{i=1}^l b_i v_i, \quad v_i \in V, \quad l < N \quad (1)$$

and in matrix form

$$x = Lb \quad (2)$$

We have for the synthesized signal $\hat{s}(n)$

$$\hat{s}(n) = \sum_{k=0}^n h(n-k)x(k) + r(n) \quad (3)$$

where $h(n)$ is the impulse response of the LPC filter, $x(n)$ the excitation and $r(n)$ is the contribution to the present frame of the excitation in the previous frames. The error is given by

$$e(n) = s(n) - r(n) - x(n)*h(n) \quad (4)$$

Weighting $e(n)$ with a filter of the form

$$W(z) = \frac{1 + \sum_{k=1}^p a_k z^{-k}}{1 + \sum_{k=1}^p a_k \gamma^k z^{-k}} = \frac{A(z)}{A(z/\gamma)} \quad (5)$$

we can tolerate larger errors in the formant regions than in the in-between formant regions. Finally we have for the weighted mean squared error

$$E_w = \sum_{n=0}^{N-1} (lf(n) - x(n))^* h_w(n)^2 \quad (6)$$

where $f(n)$ is the signal resulting from passing $s(n) - r(n)$ through the filter $A(z) = 1/H(z)$ and $h_w(n)$ is the impulse response of the weighted LPC filter. Substituting (1) in (6) and expressing it in matrix form

$$E_w = (f - Lb)^T H_w^T H_w (f - Lb) \quad (7)$$

Let's now minimize E_w with respect to the excitation. For a given subset L the coefficients b that minimize E_w are given by [3]

$$b_m = (L^T H_w^T H_w L)^{-1} L^T H_w^T H_w f \quad (8)$$

and replacing (8) in (7) we have the minimum value of E_w for a given L

$$E_w^m = f^T H_w^T H_w f - [(f^T H_w^T H_w L)$$

$$(L^T H_w^T H_w L)^{-1} (L^T H_w^T H_w f)] =$$

$$f^T H_w^T H_w f - \Delta E_w^m \quad (9)$$

As V is known, we still have to determine which subset L of V is the one that minimizes E_w^m . From (9) it is clear that it will be the one that maximizes ΔE_w^m .

So the excitation is characterized by the indexes of the vectors of V that belong to L and by the values of the coefficients b_m .

Let's take now $H_w^{-1}T$ as V , being T a unitary matrix. As L is a column submatrix of V then

$$L^T H_w^T H_w L = I_l \quad (10)$$

where I_l is the identity matrix of order $l < N$. Substituting in ΔE_w^m we have

$$\Delta E_w^m = (f^T H_w^T H_w L)(L^T H_w^T H_w f) = \|L^T H_w^T H_w f\|^2 \quad (11)$$

We will maximize ΔE_w^m by choosing as

L those vectors of V whose indexes correspond to the l biggest magnitude elements of the vector

$$q = V^T H_w^T H_w f =$$

$$T^T H_w f = T^T y \quad (12)$$

where y is the signal $s(n) - r(n)$ weighted by the filter $W(z)$. The coefficients b_m are given by the values of the elements of q chosen.

As it can be seen in (12) we are applying the unitary transform T^T to the weighted speech signal $s(n) - r(n)$, and considering equal to zero the smallest magnitude elements as common in conventional transform coding. However this scheme is different because the speech signal is weighted by the filter $W(z)$ and it is considered the contribution of the previous frames to the present frame. Due to this link with the LPC model of speech production, this scheme has been called Predictive Transform Coding (PTC) [3].

We are interested in studying the performance of different unitary transforms and, specially, in determining how many elements can be considered equal to zero in this new scheme for each transformation without degrading speech quality.

3. SEVERAL UNITARY TRANSFORMS

We are going to study four different unitary transforms: the identity transform, the unitary transform U resulting from applying the Singular Value Decomposition (SVD) to the matrix H_w , the discrete cosine transform (DCT) and, finally, the recently developed [2] Modified Hermite Transformation (MHT).

3.1 Identity transform

Making $T = I$ results in

$$q = T^T y = y \quad (13)$$

So in this case we just take the biggest magnitude elements of the weighted speech signal, and reconstruct the signal by passing these elements through the inverse weighting filter $1/W(z)$ and adding the signal $r(n)$.

3.2 Unitary transform resulting from SVD

Let's weight (3) and express it in matrix form

$$y = s_w - r_w = H_w x \quad (14)$$

If now we apply the Singular Value Decomposition to the matrix H_w

$$H_w = UDK^T \quad (15)$$

where U and K are unitary matrixes and Σ is a diagonal matrix. Making $T = U$ we finally have

$$q = U^T y = DK^T x \quad (16)$$

3.3 Discrete Cosine Transform

The Discrete Cosine transform of a data sequence $s(n)$, $n = 0, 1, \dots, (N-1)$ is given by [1]

$$c_s(0) = \frac{\sqrt{2}}{N} \sum_{n=0}^{N-1} s(n) \quad (17)$$

$$c_s(k) = \frac{2}{N} \sum_{n=0}^{N-1} s(n) \cos \frac{(2n+1)k\pi}{2N} \quad (18)$$

$$k = 1, 2, \dots, (N-1)$$

This transform has been traditionally used due to its close performance to that of the Karhunen-Loeve transform.

3.3 Modified Hermite Transformation

This unitary transformation has been recently developed by [2] from the binomial and discrete hermite families. It is defined by the unitary matrix A

$$A(r, k) = \frac{[B(0, k)B(0, r)]^{1/2}}{2^{N/2}} H(r, k) \quad (19)$$

where

$$B(0, k) = \binom{N}{k}$$

$$B(r, k) = \binom{N}{k} \sum_{m=0}^r (-2)^m \binom{r}{m} \frac{k^m}{N^m}$$

$$H(r, k) = \frac{B(r, k)}{\binom{N}{k}}$$

$$k = 0, 1, \dots, N-1$$

$$r = 0, 1, \dots, N-1$$

It has also been developed an algorithm for the MHT which is very easy to implement. We are interested in evaluating the performance of this new transform in a PTC scheme.

4. EXPERIMENTAL RESULTS

For performance evaluation we used four Spanish sentences pronounced by two male speakers and two female speakers. The sample was low-pass filtered at 4 kHz cut-off frequency and digitized by a 12 bit A/D converter at 8 kHz sampling.

The length of the analysis frame was considered of 15 msec, 120 samples, and it was divided into 3 sub-blocks of 40 samples each. Synthesis filter order was established to be 10 and the value of $\gamma = 0.8$.

We have studied the performance of the four transformations in three different cases:

- A- We consider different from zero 5 values in each sub-block.
- B- We consider different from zero 10 values in each sub-block.
- C- We consider different from zero 15 values in each sub-block.

In all the cases without quantization.

For objective evaluation we have used the Segmental SNR. In figures 1a and 1b the SNRseg is plotted for male and female sentences respectively. We evaluate the four transformations at the three different cases A, B and C.

For subjective evaluation we have used the Mean Opinion Score (MOS) scale with five categories, ranging from 1 (Unacceptable) to 5 (Excellent). Opinion rating was made by ten listeners over the four Spanish sentences. In figure 2 we have the MOS for the four transformations at the three different cases A, B and C.

The results obtained show a close objective performance of the DCT and of U. Their SNRseg is 1.5 db approximately better than that of the MHT, being the performance of the identity transformation quite worse. As far as the subjective measure is concerned we get a MOS of 3 to 4 for DCT, U and MHT if we consider different from zero

10 to 15 elements. A score of 4.0 on the MOS scale signifies high-quality and 3.5 is an acceptable quality for telephone communication. As it can be seen the MHT also has a worse subjective performance with respect to the DCT and U, however this difference diminishes as the number of elements different from zero increases, being the performance of these three transforms quite close for around 15 elements.

There is another point to take into account when evaluating the different transforms and it is the amount of calculation involved. The quickest to implement is, of course, the identity transform but it gives worse results. U implies a large amount of calculation as we have to do the SVD of the matrix H_w for every frame. For the DCT and the MHT there are fast algorithms developed, being the one developed for the MHT easier to implement than any of those developed for the DCT [2] to the authors knowledge. So the MHT is more advantageous in this aspect.

The results obtained are without quantization of the different parameters, what would evidently introduce a certain amount of degradation over this results.

5. CONCLUSIONS

We have studied the performance of four unitary transforms in a predictive transform coding (PTC) scheme. The identity transformation is very easy to implement but it gives poorer results. The DCT and U have a better objective and subjective performance over the MHT, but for around 15 elements different from zero, the subjective performance of these three transforms is very close, with a MOS of about 4 without quantization. As the MHT is easier to implement, this transform can be a good election when implementation reasons dominate in the development of PTC coders.

REFERENCES

- [1] Ahmed N., Natarajan T., Rao K.R., "Discrete Cosine Transform", IEEE Trans. on Compt., January 1974.

- [2] Haddad R.A. and Akansu A.N., "A New Orthogonal Transform for Signal Coding", IEEE Trans. on ASSP, vol. 36, No. 9, September 1988.
- [3] Ofer E., Malah D. and Dembo A., "A Unified Framework for LPC Excitation Representation in Residual Speech Coders", ICASSP 1989

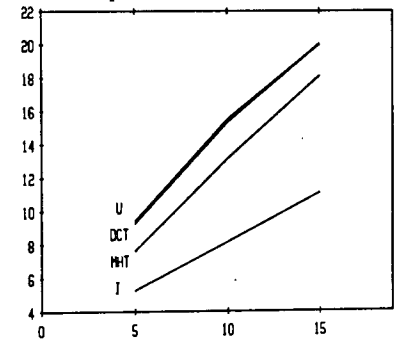


Figure 1a.- SNRseg(db) for cases A,B,C for male sentences

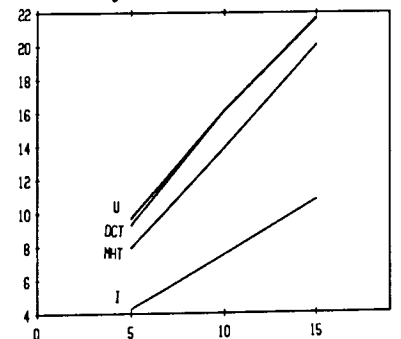


Figure 1b.- SNRseg(db) for cases A,B,C for female sentences

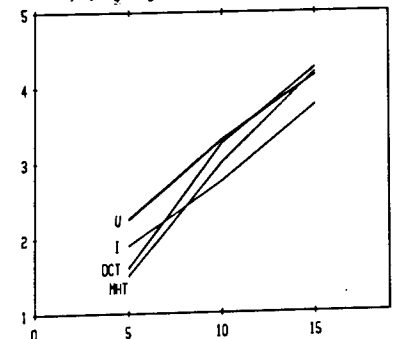


Figure 2.- Mean Opinion Score (MOS) for cases A,B,C

NUMERICAL SIMULATION OF THE GLOTTAL FLOW AND GLOTTAL EXCITATION

Gabriele C. Hegerl

Siemens AG, Otto-Hahn-Ring 6, D-8000 München 83, West Germany

ABSTRACT

Up to now, little is known about the glottal flow while phonated speech is produced. We derive a model of the glottal flow which includes nonlinear flow, the dynamic generation of sound waves by the vibrating vocal cords and the interaction of those two phenomena. Results of a numerical simulation based on this model and the compressible Navier-Stokes equations are given and compared with theoretical values.

1. INTRODUCTION

Up to now, most models for speech production represent the vocal tract by a linear filter with the glottal excitation as source. Nonlinear phenomena like energy dissipation and noise creation by vortices are not taken into account. A crucial point of speech production is the sound generation at the glottis. So far only very crude glottal models are used showing great deviations from reality, as these models do not include nonlinear time-dependent flow and its interaction with the glottal excitation [1]. For investigating the behaviour of the elastic vocal folds, Ishizaka and Flanagan introduced a basic mass-spring model [2], whose driving force is the estimated pressure of the flow. Another approach is, to get insight into the nonlinear flow connected

with speech generation. Since measurements of the flow are difficult and have to be based upon simplified mechanical models [3], numerical simulation is needed. As a first effort, T. Thomas simulated the flow in a two-dimensional model of the supraglottal vocal tract with a coarse grid and steady equations [4]. Recently, Iijima et al. computed the two-dimensional flow at the glottis in different but time-invariant stages of glottal opening [5]. They have shown that complicated vortical flow develops causing pressure variations which may acoustically affect the glottal flow. As sound waves result from the compressibility of air, phonation can not be represented by their incompressible approach. Connecting both approaches, our model of the glottal flow and glottal excitation for the first time includes the dynamic generation of sound waves and their interaction with two-dimensional vortical flow. It is based on the Navier-Stokes equations for compressible viscous gases in a domain with moving boundaries. Due to the high computational effort, we presently restrict our simulation to two space dimensions.

2. A MODEL OF THE GLOTTAL FLOW

The glottal excitation is generated by the unstationary flow of moist air bounded by the moving vocal cords.

This flow is governed by the compressible Navier-Stokes-equations, which are based on the *balance of mass*

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{V}) = 0, \quad (1)$$

and the *balance of momentum* [6]:

$$\frac{\partial}{\partial t} (\rho \mathbf{V}) + \nabla \cdot (\rho \mathbf{V} \mathbf{V} + p \mathbf{I} - \tau) = 0. \quad (2)$$

(t : time, ∇ : gradient in space, ρ : density, \mathbf{V} : velocity vector, p : pressure, τ : viscous tensor for air [6], \mathbf{I} : unit matrix; bold letters: vector-valued functions). The equations describe that no mass can be lost (1) and how mass flow, density fluctuations, pressure and viscosity interact (2). An adiabatic relation between pressure and density is used:

$$p / \rho^\gamma = \text{const.} \quad (3)$$

Although the flow at the glottis is slow compared to the velocity of sound, we cannot neglect compressibility (by assuming constant pressure in (1)), since the generation of sound waves is basically connected with density fluctuations.

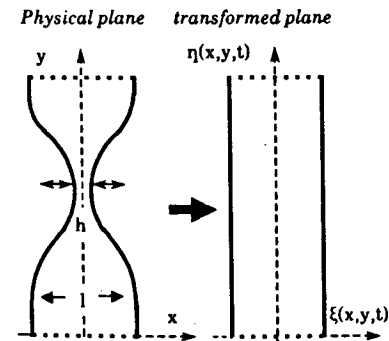


Fig. 1: Time-dependent transform

The geometry of the simulation is a simplified model of the glottal region. To treat the moving boundaries, we use a time-dependent coordinate transform from the physical domain to a rectangular geometry (see [7]). The transform should not distort the grid too much, otherwise the trans-

formed equations and the numerical method get unstable. Thus a simplified geometry neglecting the false vocal folds (fig. 1) is presently used. Our computational domain contains *solid wall boundaries* (fat lines in Fig. 1, here the air moves with the walls of the vocal tract) and *artificial boundaries* (dotted lines in fig. 1), where absorbing boundary conditions are used. Table 1 gives the values of constants used in our computation.

The glottal excitation is induced by the pressure of the flow. A model comparable to Ishizaka's model with one mass and one spring characterizing the movement of the elastic vocal cords [2] can simulate this effect. It leads to a differential equation based on the balance of inertia, friction, the nonlinear elasticity of the vocal cords and the pressure of the flow. Presently, a sinusoidal movement of the vocal cords is imposed from outside, the model for flow-induced excitation is being developed.

Table 1: Values of some constants used in the computation

initial density of air	$\rho = 0.15 \text{ kg/m}^3$
dyn. viscosity of air:	$\mu = 0.19 \cdot 10^{-4} \text{ kg/(m sec)}$
ratio of specific heat capacities:	$\gamma = 1.4$
volume flow:	$V_{\text{vol}} = 50 \text{ to } 500 \text{ cm}^3/\text{sec.}$
length of the constrictions:	$l = 1.2 \text{ cm}$
depth of the glottis:	$l_d = 1.2 \text{ cm}$
glottal opening:	$h = 4.0 \text{ to } 6.7 \text{ mm}$ (realistic: 0 to 3 mm)

3. NUMERICAL SOLUTION

The compressible Navier-Stokes equations are very delicate numerically. We have adapted the ASWR (Asymmetric Separated Weighted Residuals) -method for solving the transformed equations gained from (1), (2) [8]. Although we use an efficient implicit method for time-integration, the computational expense is immense, in the current

implementation a Tera-FLOP-computer would be necessary for a realtime calculation. The results have been verified by simplified fluid dynamic considerations and by comparing the calculated pressure loss for a steady geometry with an approximative theoretical value [2].

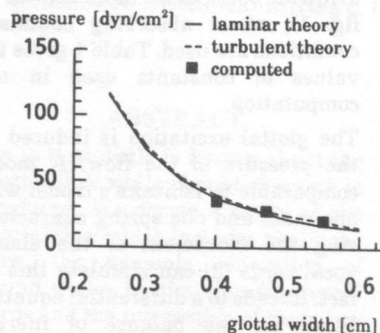


Fig. 2: Translaryngeal pressure loss in the glottis compared to the approximate theoretical value [3].

4. RESULTS

By the vibrating glottal model, the generation of sound waves can be modeled. The constriction moves with a frequency of 100 Hz from a slightly to a maximally constricted tube. The pressure in space and time and the velocity have been watched during several oscillation periods. Results are shown in Figures 3 to 5. After the initial value of zero flow, the flow increases, causing the initial pressure rise in fig. 3. The pressure in time shows a steeper rise and a softer decline than the sine, which governs the movement of the constriction. This is an effect of the finite speed of sound, it agrees with experimental results also using a sinusoidal excitation [4]. The deformation of a sound wave by the flow pressure can be watched in fig. 4. In fig. 5, the vortical flow caused by the moving constriction is depicted. The strongest nonlinear effects develop while

the glottis opens (right).

5. CONCLUSIONS

This simulation is only the first step towards a nonlinear dynamic glottal model, but it already includes substantial effects. A fully nonlinear dynamic vocal tract model is beyond the reach of today's computers, but by coupling a usual linear model of the vocal tract with the nonlinear model at the glottis, simulation of the generation of voiced speech including nonlinear effects can be accomplished.

Acknowledgement

I would like to thank Prof. Sachs and U. Graf for help in the numerical realisation and O. Schmidbauer and E. Marschall for many helpful discussions. The work was supported partly by the EEC in the context of the ESPRIT-project ACCOR.

Literature

- [1] S. Rösler, H. W. Strube: *Measurement of the Glottal Impedance with a Mechanical Model*, J. Acoust. Soc. Am. 86 (5), Nov. 1989
- [2] K. Ishizaka; Matsudaira: *Fluid Mechanical Considerations of Vocal Cord Vibration*. S.C.R.L. Monograph 8, April 1972
- [3] A. M. Barney, C. H. Shadle, D. W. Thomas: *Airflow Measurement in a Dynamic Mechanical Model of the Vocal Folds*, Proc. ICSLP-90, Kobe, Japan, Nov. 1990
- [4] T. Thomas: *A Finite Element Model of Fluid Flow in the Vocal Tract*, Computer Speech and Language 1, pp. 131-151 (1986)
- [5] H. Iijima; N. Miki; N. Nagai: *Glottal Flow Analysis Based on a Finite Element Simulation of a Two-Dimensional Unsteady Viscous Fluid*, Proc. ICSLP-90, Kobe, Japan.
- [6] D. Landau, E. M. Lifshitz: *Lehrbuch der theoretischen Physik: VI, Hydrodynamik*, Akademie-Verlag, Berlin, 1978.
- [7] G.C. Hegerl: *Numerical Simulation of the Glottal Flow by a model based on the Compressible Navier-Stokes Equations*, Proc IEEE ICASSP 91, Toronto, 1991, to appear
- [8] U. Graf, *A Survey of a Numerical Procedure for the Solution of Hyperbolic Systems of 3-D Flow*, Kerntechnik, Vol. 49 (1986)

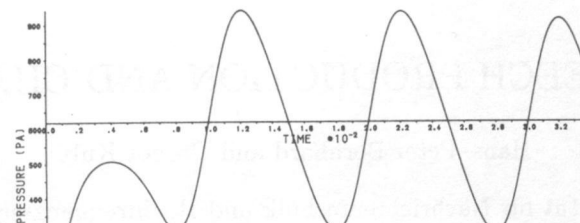


Fig. 3: Pressure at a point behind the glottis during time.

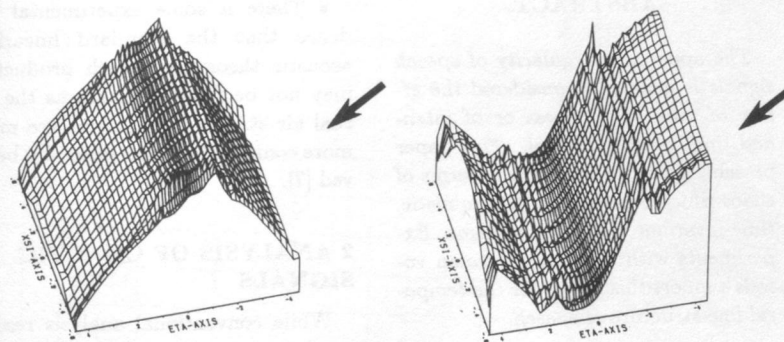


Fig. 4: Spatial pressure distribution in the transformed plane at the slightly constricted (left) and the fully constricted area (right). Arrow: direction of the flow. Maximal pressure difference: 289 dyn/cm² (left) to 330 dyn/cm² (right)

velocity vector.
dot: origin, line: direction and length

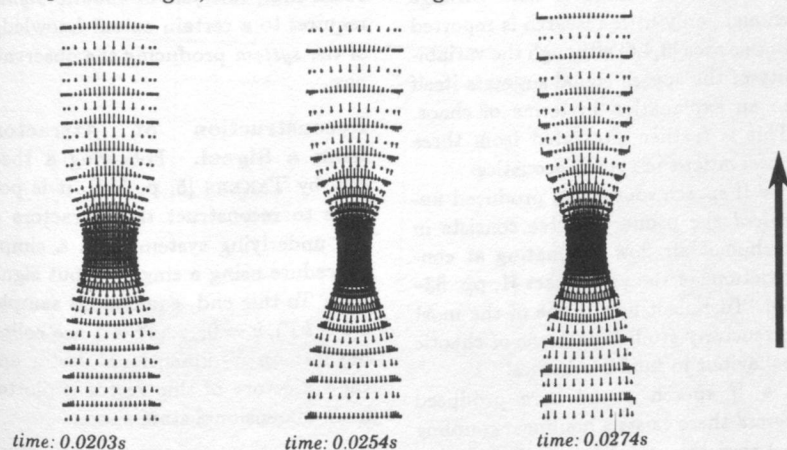


Fig. 5: Velocity vectors in a central section of the geometry drawn in the physical plane. Max. velocity: 5.7 m/sec. (Scaled to 3.5 m/sec)

SPEECH PRODUCTION AND CHAOS

Hans-Peter Bernhard and Gernot Kubin

Institut für Nachrichtentechnik und Hochfrequenztechnik,
Technische Universität Wien,
Gusshausstrasse 25/389, A-1040 Vienna, Austria

ABSTRACT

The apparent irregularity of speech signals is generally considered the effect of pure randomness or of intended time-varying control. This paper presents a new explanation in terms of chaos which originates from deterministic, time-invariant, nonlinear systems. Experiments with sustained German vowels support this model for the temporal fine structure of speech.

1 INTRODUCTION

The present study is devoted to an experimental investigation whether the chaotic paradigm is useful for speech production modeling or not. Strange enough, only little research is reported in this area [3,4,6] although the variability of the speech signal suggests itself to an explanation in terms of chaos. This is further motivated from three observations in speech acoustics:

- If speech sounds are produced unvoiced the primary source consists in turbulent air flow originating at constrictions of the vocal tract [1, pp. 53-58]. Turbulent flow is one of the most extensively studied patterns of chaotic behaviour in fluid mechanics.

- If speech sounds are produced voiced there exists a nonlinear coupling between the sound source and the vocal tract [1, pp. 41-53, pp. 246-259].

- There is some experimental evidence that the standard linearized acoustic theory of speech production may not be fully adequate as the actual air stream mechanisms are much more complicated than generally believed [7].

2 ANALYSIS OF CHAOTIC SIGNALS

While conventional analysis resorts to the assumption of an underlying stochastic process to explain the irregular behaviour of observed signals, chaos is an inherent property of purely deterministic, nonlinear systems. From that, analysis of chaotic signals requires to a certain extent knowledge of the system producing the observations.

Reconstruction of Attractors from a Signal. Following a theorem by TAKENS [5, p. 191], it is possible to reconstruct the attractors of the underlying system from a simple procedure using a single output signal $s(t)$. To this end, equidistant samples $s(t + kT)$, $k = 0, \dots, N - 1$ are collected into an N -dimensional vector and the trajectory of this vector is plotted in N -dimensional state space.

Measurement of Fractal Dimension. There are various definitions for

the fractal dimension of a geometrical object [5, pp. 167-191]. We have implemented an estimator for the correlation dimension of the reconstructed attractor.

Measurement of Information Production Rate. The unpredictable behaviour of chaotic systems corresponds to a steady production of new information which is related to its Lyapunov exponents [5, pp. 73-81]. We follow an independent approach developed by FRASER [2] which computes the mutual information between a signal sample and a vector of delayed samples. This quantity defines the marginal redundancy $R(T, N)$ of the sample, i.e. how many bits of the sample are predictable from the past N samples with delay T . It exhibits three distinctive properties for chaotic signals:

- An increase of N makes the redundancy $R(T, N)$ reach a saturation line.

- This saturation line decreases with increasing delay T . Its slope is the information production rate of the underlying system in bits/sec.

- The delay T can be chosen such that the information conveyed by the reconstructed attractor is maximized.

3 EXPERIMENTS WITH SPEECH SIGNALS

Our experiments cover long vowels [a, e, i, o, u] spoken by three male native speakers of German, of age 24 to 30. Although the vowels were embedded in carrier sentences, they were sustained over several seconds to provide sufficiently long stationary data for further analysis. The recordings were made in an anechoic chamber and digitized with 16 bit resolution and 32 kHz sampling rate. The analysis software is implemented on an IBM

PC-AT 286 for testing and visualization purposes while the more extensive batch-mode analysis runs were done on a MicroVax II minicomputer.

Results. Figure 1 shows a two-dimensional projection of the attractor corresponding to the vowel [i:] by speaker 'F'. The optimal delay time $T = 0.94$ msec. The graph displays three periodicities: (1) the trajectory almost replicates itself after every pitch period; (2) the first formant frequency is roughly twice the pitch frequency, so the overall structure of the graph is similar to a folded 8 (two cycles per 'pitch orbit'); (3) the second formant frequency is about 23 times the pitch frequency, so there are 23 small loops superimposed on the pitch orbit. From the reconstructed attractor, the correlation dimension is estimated as 1.7. This fractional value reflects the observation that the signal is almost periodic (otherwise it would be 1).

The overall structure of the attractor is maintained for other speakers. Figure 2 displays the attractor of vowel [i:] by speaker 'B'. It is characterized by the same looping structure while the amplitude of the second-formant loops is significantly reduced due to power differences in the respective formants of the two speakers.

Figure 3 explains the importance of the right choice for the delay time T . While Figure 1 is obtained with the optimal T -value, this figure shows again vowel [i:] by speaker 'F' but with $T = 1.88$ msec. The optimal T -value produces results which are easier to read and, most interestingly, this value is rather independent of the individual speaker or vowel under consideration. The optimal delay time is on the order of 1 msec.

Redundancy analysis of the attrac-

tor of Figure 1 delivers Figure 4. On display are the marginal redundancies $R(T, N)$ for $N = 1, 2, 3$ and T between 0 and 500 sampling intervals (i.e. 0 to 15.6 msec). The curves saturate already for $N = 2$ so that $N = 3$ does not increase the predictable amount of information significantly. Prior to saturation (for $N = 1$), unmodeled periodicities are still visible as local maxima of the curve, note e.g. the pronounced peak due to the pitch at $T = 7.9$ msec. The slope of the saturation line is 0.94 bit over one pitch period. This is the information production rate of the underlying system.

Figures 5 and 6 show as a further example the reconstructed attractor and redundancy plots for the vowel [a:] by speaker 'G'. While differing slightly in the numerical values, the general picture found in the previous figures is corroborated. For comparison, the pertaining measurement values are listed here: correlation dimension 1.5, optimal delay time $T = 0.94$ msec, information production rate 0.9 bit over one pitch period.

4 CONCLUSION

Albeit the experimental basis is still rather limited, preliminary conclusions may be drawn:

- The investigated vowels can be interpreted in terms of deterministic chaos. This is supported through the positive information production rate of roughly one bit per pitch period and the fractional value of the correlation dimension between 1 and 2.

- As only sustained vowels with constant pitch have been investigated, the observed chaotic behaviour can only be related to the temporal fine structure of speech that changes from pitch period to period. No claim about the dynamic behaviour of spectral or articu-

latory parameters is made.

- The gross shape of the reconstructed attractors can be interpreted in terms of standard phonetic theory. Their 'strangeness', however, goes beyond such theory which offers only explanations in terms of randomness or time-varying control.

REFERENCES

- [1] J.L. FLANAGAN. *Speech Analysis Synthesis and Perception*. Springer-Verlag, Berlin, 1972(2).
- [2] A.M. FRASER. "Information and entropy in strange attractors," *IEEE Trans. on Inf. Theory*, IT-35(2):245-262, 1989.
- [3] A. KUMAR and S.K. MULICK. "Speech signal modeling à la chaos," *IEEE 1990 Dig. Sig. Proc. Workshop*, paper 1.8, New Paltz (NY), Sept. 1990.
- [4] P. MARAGOS and K.L. YOUNG. "Fractal excitation signals for CELP speech coders," *Proc. ICASSP'90*, 669-672, Albuquerque (NM), Apr. 1990.
- [5] T.S. PARKER and L.O. CHUA. *Practical Numerical Algorithms for Chaotic Systems*. Springer-Verlag, Berlin, 1989.
- [6] T.F. QUATIERI and E.M. HOFSTETTER. "Short-time signal representation by nonlinear difference equations," *Proc. ICASSP'90*, 1551-1554, Albuquerque (NM), Apr. 1990.
- [7] H.M. TEAGER and S.M. TEAGER. "Evidence for nonlinear production mechanisms in the vocal tract," in *Speech Production and Speech Modelling*, Kluwer, Dordrecht, 1990.

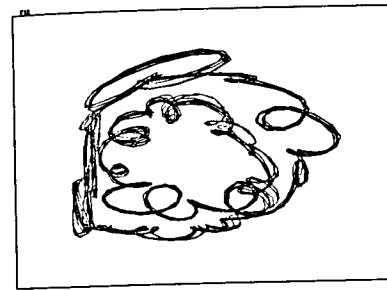


Figure 1: Attractor of vowel [i:] by speaker 'F' ($T = 0.94$ msec).

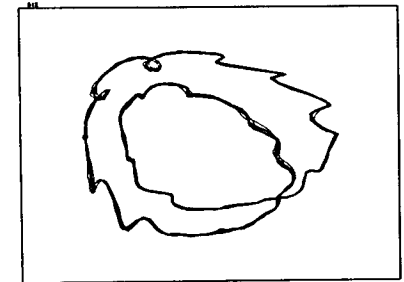


Figure 2: Attractor of vowel [i:] by speaker 'B' ($T = 0.94$ msec).

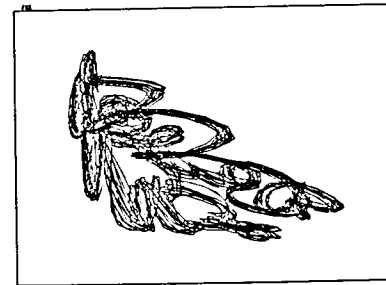


Figure 3: Attractor of vowel [i:] by speaker 'F' ($T = 1.88$ msec).

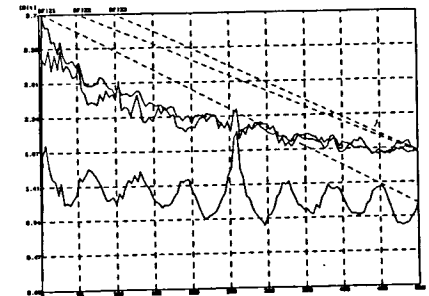


Figure 4: Marginal redundancy $R(T, N)$ of vowel [i:] by speaker 'F'.

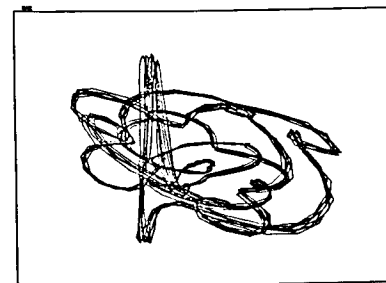


Figure 5: Attractor of vowel [a:] by speaker 'G' ($T = 0.94$ msec).

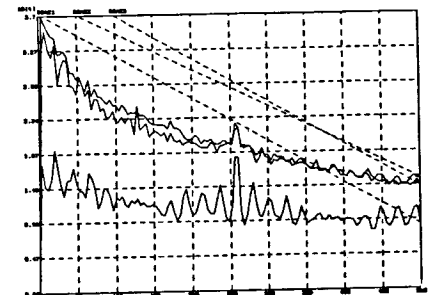


Figure 6: Marginal redundancy $R(T, N)$ of vowel [a:] by speaker 'G'.

EFFECTS OF LANGUAGE CHANGE ON VOICE QUALITY.
AN EXPERIMENTAL STUDY OF
CATALAN-CASTILIAN BILINGUALS

M. BRUYNINCKX *, B. HARMEGNIES *,
J. LLISTERRI ♦, D. POCH-OLIVE ♦.

* Département de Phonétique et Psychoacoustique
Université de Mons-Hainaut, Belgium.

♦ Departament de Filologia Espanyola, Laboratori de Fonètica
Universitat Autònoma de Barcelona, Barcelona, Spain.

SUMMARY

The paper is aimed at determining whether the language spoken by the speaker exerts influences on his or her voice quality. The investigation involves 24 bilinguals (12 males and 12 females) spliced into two subgroups of linguistic dominance (Catalan or Castilian). The Long Term Average Spectrum (LTAS) is used as an acoustic cue to voice quality. The inter-language variability is higher than the intra-language variabilities, irrespective of the sex and language dominance categories. The data moreover exhibit a tendency towards greater intra-language variability in the dominant language than in the non-dominant one.

1. INTRODUCTION

Various experiments have dealt with the idea that the language spoken by a subject could influence his or her voice quality. Studies using Long Term Average Spectra (LTAS) in order to objectivate voice quality have nevertheless resulted in -at least apparently- contradictory results (Nolan, 1983).

In order to by-pass the

contradiction, Harmegnies & Landercy (1985) have developed a specific methodology, aimed at differentiating subject- and language related variabilities. Analysing utterances from bilingual speakers (French/Dutch), they confirmed the existence of language effects on long-term spectral variability but also revealed that these effects are smaller than individual ones.

Harmegnies et al. (1989) have confirmed these findings on the basis of recordings drawn from languages even closer one to another (Castilian and Catalan, that belong to the same Romance group). Their investigations, together with others (Bruyninckx et al., 1990) nevertheless showed that a more refined assesment of bilingualism could be of some use. On the other hand, splitting the samples into sex groups might be more convenient, given the generally important differences between within-subject variabilities in males and females (Harmegnies, 1988, a).

This paper is therefore meant as an extension of previous research, putting emphasis on the kind of bilingualism, and involving

both male and female subjects in a balanced proportion.

2. EXPERIMENTAL SETTING

A sample of 24 bilingual subjects (12 males, 12 females), with a fairly good knowledge of Catalan and Castilian was selected. Each of them nevertheless exhibited dominance either in Catalan or in Castilian; their assesment was performed by means of classical sociolinguistic methodology (Viladot, 1981).

Each speaker was asked to utter 5 times both a balanced Castilian and Catalan text. The recordings and analyses were performed thanks to the experimental setting previously used, i.e. 0-5 kHz LTAS (Harmegnies et al., 1989). The spectra were compared by means of the SDDD dissimilarity index (Harmegnies, 1988, b).

3. COMPARISON PROCEDURE

Two types of intra-speaker comparisons of the stored LTAS were performed for each subject: inter-language comparisons and intra-language comparisons.

In the case of the inter-language comparisons (Catalan *vs.* Castilian) each Catalan LTAS of each subject was matched against each Castilian LTAS of the same subject; in intra-language matchings (i.e. both Catalan *vs.* Catalan, and Castilian *vs.* Castilian), for each one of the 24 speakers, one comparison was performed for each possible non-redundant pair of his or her 5 LTAS in each language.

The whole procedure resulted in a total of 1080 comparisons (480 intra-language and 600 inter-language).

4. RESULTS

As shown in Figures 1 a-b, which summarize the results, the SDDD values derived from inter-language comparisons tend to score higher than those drawn from intra-language matchings. This relationship can be observed for the average values in each sex/dominance group. This tendency moreover appears to be very strong, since the relationship holds true for each one of the 24 speakers in the sample.

A specific statistical analysis (Bruyninckx et al., forthcoming) confirms this observation. It moreover shows that in 2 out of 4 subgroups (Catalan dominant females and Castilian dominant males), the intra-language comparisons exhibit results significantly different one from another.

5. DISCUSSION

At first sight, the results seem in partial agreement only with those previously reported by Harmegnies et al. (1989): in the present research, inter-language differences also appear to be significant; however, the systematic superiority of the inter-language dissimilarity values is no longer to be regarded as the only one source of variation in the design: the differences between intra-language dissimilarities suggest that in some cases, the intra-

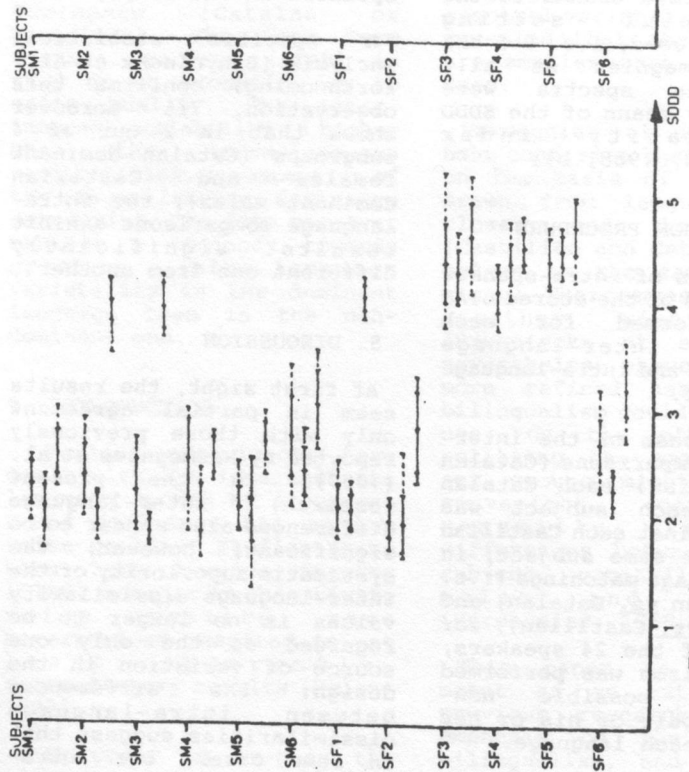


Figure 1a : means, standard deviations, lower and upper values of the SDDD values for each Catalan dominant male (SM) and female (SF) subject in each comparison condition (cat/cat; cast/cast; cat/cast).

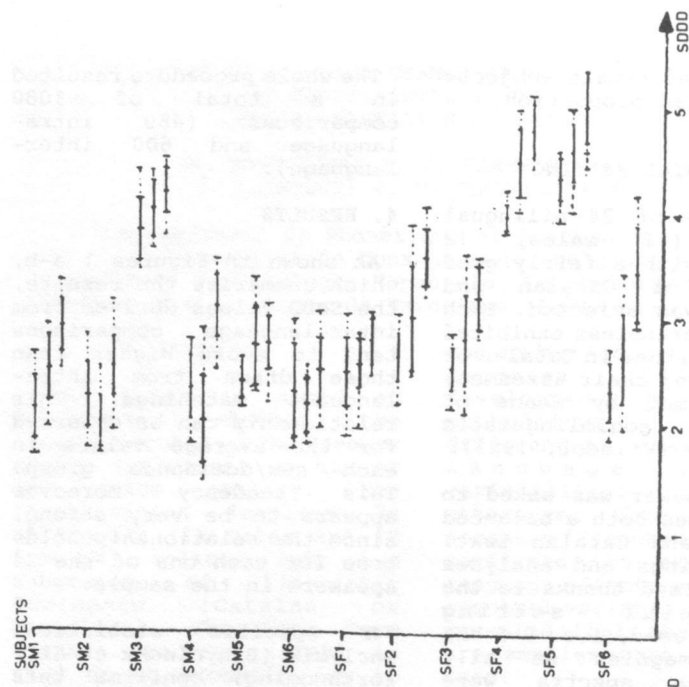


Figure 1b : means, standard deviations, lower and upper values of the SDDD values for each Castilian dominant male (SM) and female (SF) subject in each comparison condition (cat/cat; cast/cast; cat/cast).

language consistency tends to depend upon the language spoken. As shown in figure 1, the significant differences in the Catalan dominant group may be attributed to greater voice consistency in Castilian, as opposed to Catalan. On the other hand, in Castilian dominant speakers, the significant differences are due to greater voice consistency in Catalan.

Thus, there is a tendency towards greater voice quality variability for a given speaker in his or her dominant language, while the degree of voice coherence tends to be higher in the non-dominant language.

Finally, this research not only confirms previous findings (i.e. Catalan LTAS are different from Castilian LTAS), but also suggests a new hypothesis: voice consistency in bilinguals is greater in the non dominant language; inter-language variability is nevertheless greater than intra-language ones.

6. REFERENCES

BRUYNINCKX, M., HARMEGNIES, B., LLISTERRI, J., POCH, D., "Bilinguisme et qualité vocale. Contribution à l'analyse des variations du spectre moyen à long terme sous l'effet du changement de langue", in Mélanges de phonétique et didactique des langues. Hommage au Professeur Raymond Renard, 43-53, P.U.M./Didier Erudition, Paris, 1990.

BRUYNINCKX, M., HARMEGNIES, B., LLISTERRI, J., POCH, D., "Voice quality and language dominance in bilinguals",

forthcoming.

HARMEGNIES, B., Contribution à la caractérisation de la qualité vocale; analyses plurielles de spectres moyens à long terme de parole, Doctoral Dissertation, Université de Mons, 1988, a.

HARMEGNIES, B., "A new dissimilarity index for the comparison of speech spectra", Pattern Recognition Letters, 8, 153-158, 1988, b.

HARMEGNIES, B., BRUYNINCKX, M., LLISTERRI, J. & POCH, D., "Effects of language change on voice quality. An experimental contribution to the study of the Catalan-Castilian case", Proceedings of the First European Conference on Speech Communication and Technology, Paris, 489-492, 1989.

HARMEGNIES, B. & LANDERCY, A., "Language features in the long-term average spectrum", Revue de Phonétique Appliquée, 73-75, 69-80, 1985.

NOLAN, F., The Phonetic Bases of Speaker Recognition, Cambridge, Cambridge University Press, 1983.

VILADOT, M.A., El bilingüisme a Catalunya. Investigació i psicologia. Barcelona, Laia, 1981.

A NEW HIGH RESOLUTION TIME-BARK ANALYSIS METHOD FOR SPEECH

Unto K. Laine

Helsinki University of Technology, Acoustics Lab,
Otakaari 5 A, 02150 Espoo, Finland

ABSTRACT

A new complex auditory filter bank based on a new class of orthogonal functions called FAM functions [2] is developed. The filter bank is used to produce time-Bark spectrograms where both the magnitude and the phase variations of the speech sample in the channels are seen. The temporal resolution is of high quality: even phenomena inside one pitch period can be monitored. An analyzer is implemented with the TMS320C30 digital signal processor, which is programmed in a Macintosh IICI / CLOS Lisp environment [1].

1. INTRODUCTION

Auditory based representation of speech is proven to be an effective way to compress, code, enhance, analyse, describe, and visualize speech signals. Auditory analysis and compression seems to have many promising applications in digital audio, communication technology, and in audiology. The methodology provides also new tools for basic research in the fields of speech acoustics and production as well as speech perception and automatic speech recognition.

The auditory representation of audio signals is typically achieved by a filter bank with each filter having a bandwidth equal to one Bark (one critical band) which forms the limit for the auditory spectral resolution around the frequency in question [5].

Typically, one of the following methods is used to construct an auditory filter bank: a standard linear filter design method is used, modifying the short time Fourier transform (STFT), or applying quadrature filter (QF) techniques [3]. In

the case of the modified STFT a proper frequency dependent window function is included in the Fourier transform in order to achieve the nonuniform frequency resolution in the uniform frequency (Hz) scale. Note that this is the same case as if we had a uniform resolution in a nonuniform frequency scale (in Barks). In the QF-approach a uniform resolution filter bank is first designed according to the highest frequency resolution, and the narrow channels are then combined at the higher frequencies to reduce the resolution according to the Bark scale.

This paper describes a new type of auditory filter bank, which is designed by applying the FAM-method [2]. The method is based on orthogonal functions of a FAM class, which leads to an auditory bank with complex and orthogonal one Bark bandpass channels. Each channel consists of two filters which form a Hilbert pair and give a complex signal as the channel output. This allows to define the channel energy at every sample as the magnitude of the complex output and to formulate a phase measure within each critical band (auditory phase modelling). The orthogonality between the channels means that the impulse responses of the channels do not correlate. In fact, such a bank performs an Orthogonal Auditory Transform (OAT) from the time domain to the Bark domain.

The magnitude calculation improves the time resolution of the bank since magnitude can be calculated immediately at every sample instead of rectification and low-pass filtering the output of one (real) channel. The latter introduce a delay and some unprecision and uncertainty along the time scale.

2. FAM-METHOD

The new class of orthogonal FAM functions which are formed from the circular functions by Frequency and Amplitude Modulation has been earlier published by Laine and Altosaar [2]. In FAM functions the frequency and amplitude modulations of sinusoids are combined in a way so that a set of orthogonal functions are produced. A generative function, $g(\omega)$, which defines the frequency modulation, defines the properties of the FAM function set.

Fig. 1 describes briefly the FAM-method used to form a complex, orthogonal auditory one Bark bank. The method and its use for computation of auditory spectrograms proceeds through the following steps:

1° The desired frequency resolution, or the new warped frequency scale, is fixed by the choice of the generative function $g(\omega)$. In the case of the auditory bank the choice equals to a Hz-to-Bark conversion [4,5].

2° The orthogonal sets of FAMsin(-) and FAMcos(-) functions e.g. $\text{FAMexp}(j n g(\omega))$ are generated in the frequency domain in order to describe the complex spectra of the signal with the new resolution in the linear ω -scale, or with a uniform resolution in the new warped $g(\omega)$ -scale.

3° The FAM functions are inverse Fourier transformed into the time domain. The inverse Fourier transform retains the orthogonality, thus we get an orthogonal set of real time functions. Channel responses of this bank resemble phase distorted impulses. So, we could call the set as the Bark pulse bank (BPB) to emphasise their correspondence with the frequency domain (Bark-warped) FAM functions.

4° The speech signal can be convolved with the BPB to get the coefficients for the corresponding FAM-functions for spectrum composition.

5° The spectral picture in the Bark-domain is produced by Fourier transforming the output samples from the BPB. Note that this part of the spectrum computation can be avoided by forming the final one Bark filter bank as linear combinations of the FIRs in the BPB. Instead of the Fourier transform we can produce the final one Bark FIR filter bank by forming a $\cos(nx)$ weighted sum

of the BPB FIRs to produce the real part of channel n , and $\sin(nx)$ weighted to produce the imaginary part of channel n .

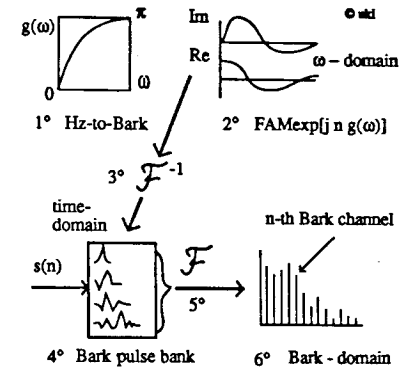


Fig. 1. FAM - method for nonuniform resolution spectral analysis.

6° The Fourier transform of the outputs of the BPB gives the desired complex spectra in the Bark-domain. The same information is available from the one Bark bank made in 5°.

3. IMPLEMENTATION

The one Bark auditory FIR filter bank designed by the FAM-method is implemented by using the TMS320C30 digital signal processor in a Macintosh IICI/ CLOS Lisp environment [1].

The C30 processor takes the speech samples in and performs the numerically intensive filtering, magnitude and phase computations, and finally, sends the data to the MacII for displaying the spectrogram and plotting with a laser printer.

The filter bank covers the frequency band from about 5 Hz to 11 kHz having one DC-type (real only) channel and 21 complex one Bark channels.

The magnitude and phase are processed for every sample in every channel. The prototype realization is relatively slow. However, we have estimated, that a real time implementation of the one Bark bank is possible by using all-pass structures for the filters.

The phase information from the one Bark channels can be processed in different

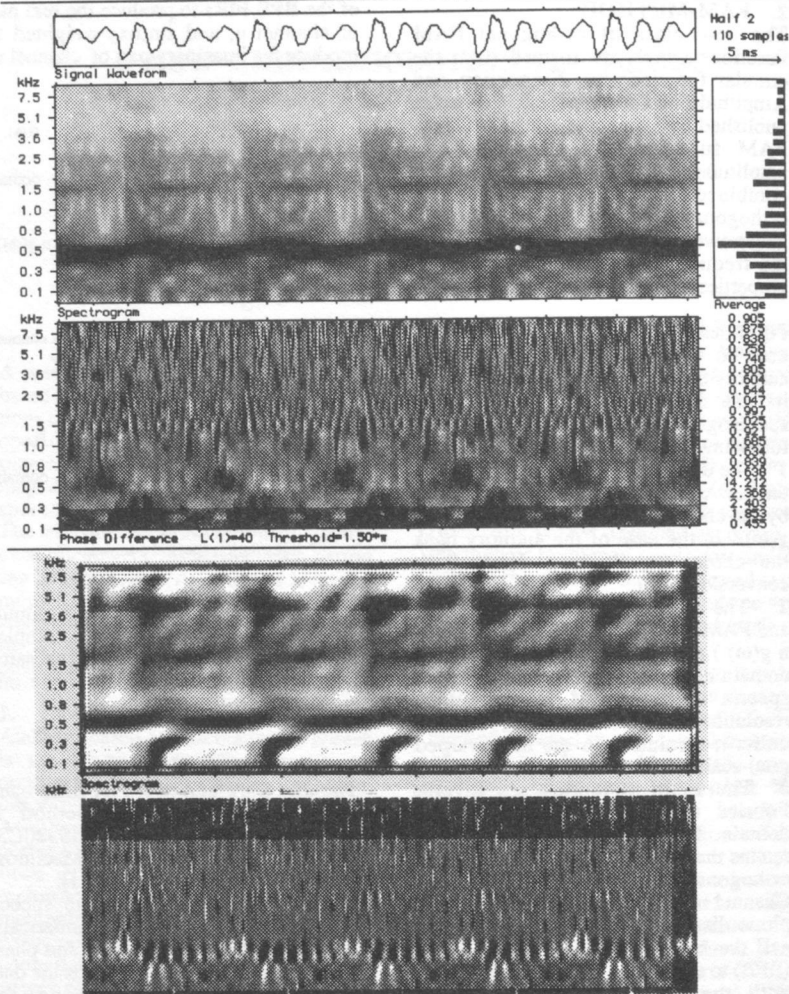


Fig. 2. Vowel /æ/ analysed by the auditory bank (see text).

ways in this auditory spectrograph: in each channel the phase information over time is first dewarped in order to cancel the 2π steps. Then the more or less linear ramp (corresponding to the time flow) is differentiated to remove the linear part and to emphasise the variations in phase. In another method a trivial prediction is made to estimate the coming phase sample from the previous samples. Then the difference of the estimated value and the true value is formed and plotted.

3. SPEECH ANALYSIS WITH THE AUDITORY BANK

Fig. 2 depicts results of the analysis made for the vowel /æ/. The waveform is shown in the topmost panel. Below it can be seen the time-Bark energy distribution and on the right hand side the average Bark-spectra. The auditory phase processed by the trivial predictor is shown in the next panel. In the following frame a graphically processed version of the previous spectrogram is

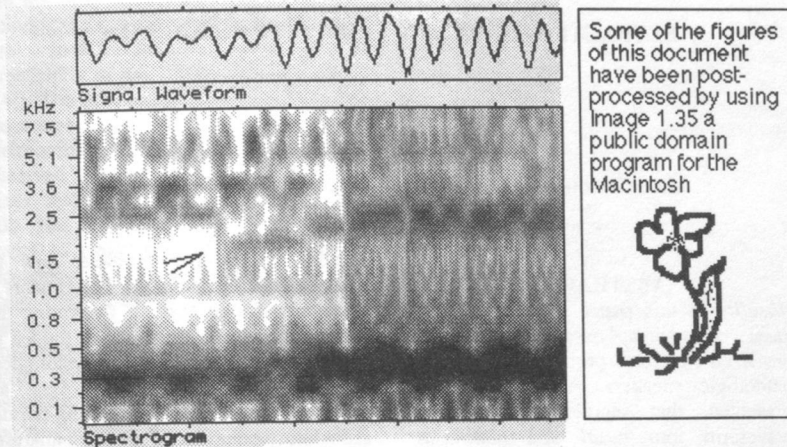


Fig. 3. Spectrogram of syllable /li/.

shown. The differentiated auditory phase is displayed in the lowest frame.

Some pitch-synchronous effects can clearly be seen: the higher formants will be primarily excited at the closure of the glottis. In some cases a clear secondary excitation appears at the glottal opening. Note also the frequency modulation of the first formant.

Fig. 3 demonstrates a transient analysis of the syllable /li/. Typically a clear transient effect can be perceived in this syllable even though the steady state spectra of /l/ and /i/ do differ only very little. The arrow in the figure points to the place where the second formant suddenly appears at the moment the tongue opens the mouth cavity. The formant moves up very fast, about 2 Barks in 12 ms! Note that the contrast of the left hand side is improved by graphical postprocessing.

4. DISCUSSION

Preliminary results from the first implementation of the FAM-based auditory filter bank was presented. Many details can and must be improved. The way the phase information is processed is not yet robust enough. After a better processing it may reveal much more detailed and relevant information of the complicated phenomena called speech acoustics.

5. ACKNOWLEDGEMENTS

The author is grateful to his colleagues at the HUT Acoustics Lab. for the help and assistance during this study. Toomas Altsaar has made a great deal of programming work with Lisp in the QuickSig/Symbolics environment, work which made this study possible. Vesa Välimäki has implemented the algorithms in the C30/MacII CLOS Lisp environment creating a nice, user friendly package. The Lisp software for the environments has been developed by Professor Matti Karjalainen.

6. REFERENCES

- [1] KARJALAINEN, M., "A Lisp-based high-level programming environment for the TMS320C30", *Proc. of IEEE ICASSP 89, Glasgow, Scotland, 1150-1153*.
- [2] LAINE, U., and ALTOSAAR, T. (1990), "An orthogonal set of frequency and amplitude modulated (FAM) functions for variable resolution signal analysis", *Proc. of IEEE ICASSP 90, Albuquerque, New Mexico, 1615-1618*.
- [3] SMITH, M., and BARNWELL, T. (1987), "A new filter bank theory for time-frequency representation", *IEEE Tr. ASSP, 35, 3, 314-327*.
- [4] TRAUNMÜLLER, H. (1981), "Perceptual dimension of openness in vowels", *J. Acoust. Soc. Am., 69, 1465-1475*.
- [5] ZWICKER, E., FELDKELLER, R. (1967), "Das Ohr als Nachrichtenempfänger", Stuttgart: Hirzel Verlag.

EXTRACTING NASALITY FROM SPEECH SIGNALS

Henning Reetz

Max-Planck-Institut für Psycholinguistik, The Netherlands

ABSTRACT

Nasality in this paper refers to voiced nasal consonants, nasal vowels, and nasalized vowels produced by non-pathologic speakers. An algorithm is presented, that segments the speech waveform into *nasal* and *non-nasal* parts. The decision whether nasality exists is based on two features of the speech signal: (1) Presence of vocal cord vibration (voicing) and (2) presence of a resonance in the range 200 Hz - 400 Hz. The frequency of this resonance may vary between speakers but is constant for a single speaker. The algorithm is validated by comparing it to results of a perception experiment. The problems to define and measure nasality are sketched before the algorithm is presented.

1. INTRODUCTION

Nasality is defined in different ways in different fields. It can be determined by underlying or surface structures of the speech material, it can be described by articulatory movements or physiological data, it can be related to signal characteristics that are expected on the basis of theoretical considerations, and it can be determined in perception tests where subjects use their native or experience listening skills. Each field defines *nasality* within its system and the definitions must not necessarily coincide. For example, [10] reported velar opening on (phonetically) non-nasal sounds and [6] found coupling to the nasal cavities even if the sound was not perceived as nasalized by listeners. In addition to the problem how to define

nasality comes the problem how to measure it. In articulatory experiments it is a complicated task to determine the position of the velum that lies hidden in the back of the mouth. Physiological experiments have to measure (nasal) air flow accurately without effecting normal articulatory behavior. Acoustical analyses have to separate properties of the nasal system from the oral and pharyngeal system. In perception tests it is not trivial to create listening conditions that are similar to everyday speech perception situations. If in any field indirect measurements are performed, it can happen that the coherence between the measured quantity and the related quantity does not exist in the expected way. For example, nasal airflow is not necessarily a measure for the size of the velar opening and it has been reported that changes in velar heights occur even though the velar port is closed ([2]).

In this paper an acoustical measure will be used that intends to be related to perceptual nasality. First we will brief some acoustical properties of nasal sounds and ways to measure them. In the following chapter an algorithm to extract low-frequency resonance will be presented. After describing a perception experiment the relation between the low-frequency resonance and the perception results is demonstrated and discussed.

2. SOME ACOUSTIC PROPERTIES AND MEASUREMENTS

Acoustic theory predicts extra anti-resonances and resonances for the additional shunting branch of the nasal cavity system ([3]). These effects can be found in the speech signal, though they

are not easy to identify ([3]). Common acoustical features found in the literature are low-frequency energy and broadening of the overall spectral pattern ([4]). The occurrence of a low-frequency resonance can be explained by a Helmholtz resonating cavity. [7] observed a resonating frequency of the sinus maxillares in the range of 200-800 Hz. [1] computed a resonating cavity of about 800 cm³ for a 250 Hz resonance and proposed (parts of) the scalp as possible resonator. The flattening of the spectrum can be explained by the complicated nasal cavity structure inducing many partial resonating and shunting cavities ([7]).

Spectrographic (or FFT) analyses have the advantage to preserve all information of the acoustical signal. This can be a handicap because the ground frequency (and the window size) is apparent in the spectrogram. Smoothing these spectra reduces the appearance of these effects but also reduces the spectral resolution. Source-filter models separate the properties of the source signal (F_0) from the oral and nasal tract, but add assumptions about the model to the estimated spectra. All-pole LPC does not include shunting cavities in its model and therefore seems to be inadequate. Fitting the model to the analyses purposes (i.e. determining the number of poles) is done by rules of thumb. Spectral estimation methods that include zeros in their models (like analyses-by-synthesis schemes or ARMA models) inherit the problem to estimate the number of poles and zeros. A problem that is non-deterministic because a zero can be approximated by a large number of poles ([8]).

In this study the low-frequency resonance will be used as a gross indicator for a phenomenon that listeners would call *nasality*. This feature is reported in many articles studying the acoustical properties of nasality and seems to us more obvious in the spectrum than the theoretically more founded existence of zeros. We are not

interested in detecting zeros at all and choose LPC analyses as the spectral estimation method. This has several advantages: it is easy to compute, it is commonly available, and it suppresses the effect of the ground frequency in the estimated spectrum.

3. ALGORITHM

In a pre-processing step the speech signal is separated in voiced and unvoiced parts using a pitch extraction algorithm ([11]). This reduces the amount of computing to be performed in the following steps and excludes background or friction noise that might show low-frequency prominence.

In the next step high order autocorrelation LPC analyses with parabolic interpolation is carried out on the voiced parts (25.6 ms Hamming window, 12.8 ms step rate). The order of the LPC must be adjusted manually to the recording and voice quality. For a 10 kHz recording, values between 14 and 28 poles were found to practical.

In the last step a low-frequency resonance is searched that lies below 400 Hz, is constant in frequency, and lasts for at least 40 msec. Constant in frequency means here that the center frequency of the resonance does not vary more than 10% in the adjacent frames.

4. EVALUATION

The performance of the algorithm is compared to the results of a lexical decision experiment carried out by [5]. The study included a gating experiment on British English monosyllabic words of the form CVC and CVN. Listeners were presented gradually increasing information (gates) from word-onset until the entire word was heard. The listeners' task was to write down the word they were hearing. The length of the gates were incremented by about 40 ms from step to step. One gate of each word were set at the offset of the vowel. 20 minimal CVN-CVC pairs (40 words) spoken by one male speaker

were presented with 9 to 13 gates giving a total of 441 gates. Details of the experimental procedure can be found in [5].

The responses of the subjects were classified in the following manner (this classification was not part of the experiment of [5]): If more than 50% of the listeners wrote down a CVN word for a gate then it was called nasal, otherwise oral. Because the listeners heard gates with increasing length, it happens that a gate was judged oral, while the next, longer gate, was judged nasal (or vice versa). By this it is possible to make a statement about the 40 ms incremental part of a gate whether it were or were not perceptually nasal. We will name an incremental part of a gate and the first gate of a series *segment* from now on.

The speech recordings used in the perception test were analyzed by the described algorithm as well. If the bandwidth of the low-frequency resonance as found by the algorithm was below 150 Hz the sounds were classified as nasal, otherwise as non-nasal. If the results of the perception test and the algorithm for a segment did not coincide it was counted as an error.

Of the 441 segments, 322 were perceived as nasal and 119 as oral. 67 (15%) of the segments were mis-labelled by the algorithm. 50 segments perceived as oral were classified nasal, in 17 cases a perceptual nasal segment was classified oral by the algorithm. We will first characterize the errors of the 50 *false alarms* then the 17 *missings*.

Of the 50 *false alarms*, 10 were segments where the algorithm classified them *too early* as nasal in front of perceptually nasal segments. 9 segments were clustered in 2 cases where a vowel in front of a nasal consonant was perceived as oral by the listeners. 29 times classified the algorithm the pre-voicing of voiced stops as nasal. And 2 segments of /l/ in CLAD were classified by the algorithm as nasal.

Of the 17 *missings* were 4 segments, where the algorithm classified them *too late* as nasal at the beginning of perceptually nasal segments. The other 13 nasal segments were missed in 4 clusters inside oral or nasalized contexts: in /l/ in CLOWN, and in the vowels of BRAN, VAIN, and TRADE.

5. DISCUSSION

The errors made by the algorithm can be grouped in 4 categories: (1) *too early* or *too late* indication of nasality, (2) judgment of pre-voicing as nasal, (3) mis-classification of /l/, (4) and mis-classification of nasalized vowels.

The first category of errors can be taken into account as jitter between perceptual and acoustic segmentation.

The second category - the classification of pre-voicing as nasals - is an interesting result. Airflow measurements have observed nasal airflow at the beginning of pre-voicing, either due to insufficient velar closure together with oral closure and increasing intra-oral pressure, or due to an upward push of the uvula extinguishing air from the nasal tract. Whether a nasal coupling exists and in how far a perception test without context information will yield nasal responses in these cases stays unclear.

The third category - mis-classification of /l/ - appears in both directions. This is a frequently mentioned effect ([9]).

The mis-classification of nasalized vowels reflects the more complicated structure of these sounds. It should be taken into account that 50% of all words contained (phonetical) nasalized sounds and that in most of these cases the results of the algorithm went along with the results of the perception test: When the listeners heard a word as nasal from an early stage of the vowel, the algorithm did so. When the listeners perceived the nasality at the end of such a vowel, so did the algorithm. Only in 4 cases we found a mis-labelling.

6. CONCLUSION

The algorithm uses a very gross feature to identify nasality ignoring context effects and formant transitions as used by [9]. He found the mean centroid frequency below 500 Hz to be the most useful contributor to the total categorization score. He observed further that a value for the first resonance near 250 Hz is a necessary but not sufficient condition for the existence of nasal murmurs (because this property is shared by the first formant frequency of high vowels). Most of his false indications result from the confusion of liquids, glides and semivowels with nasals.

The simpler algorithm used here shares obviously some of these results. But to come to a stronger statement about the behavior of the proposed method and to get more evidence for the applicability of the algorithm as a valid nasality classifier more investigation will be performed. A larger set of words of different speakers including a significant set of nasal vowels must be tested. An open question is, whether the algorithm measures a low-frequency resonance originated by the coupling of nasal tract, or whether F0 or a harmonic of it is detected. [4] and [1] observed this resonance independent of the F0 frequency and informal tests with the proposed algorithm yielded the same result. Analyses of high pitched voices have given mixed results: some voices showed good results, while other voices were classified as nasal for most of the voiced segments. Whether these are only artifacts of the algorithm or whether those voices are nasal has not been proved by perceptual tests. Application to running speech gave the impression that nasal consonants and voiced stops are consistently detected by the algorithm.

We are aware of the fact that the algorithm does not have any normalizing mechanisms and does not perform any analyses of transitions. To some extent these features could be modelled by an adaptation strategy for the number of

poles and an investigation of the development of the bandwidths in time. But we first want to study the algorithm at the present state and collect experience about its behavior before extending it.

7. REFERENCES

- [1] Castelli, E., Perrier, P., Badin, P. (1989), "Acoustic considerations upon the low nasal formant based on nasopharyngeal tract transfer function measurements", *European Conference on Speech Technology, II*, Paris, 412-415.
- [2] Counihan, D. T. (1979), "Oral and nasal airflow and air pressure measures", In: Bzoch, K.R. (Ed.) "Communicative disorders related to cleft lip and palate", Little, Brown & Company, Boston, 269-276.
- [3] Fant, G. (1960), "Acoustic theory of speech production", Mouton & Company, The Hague, Netherlands.
- [4] Fujimura, O., Lindqvist, J. (1971), "Sweep-tone measurements of vocal-tract characteristics", *JASA*, 49, 541-558.
- [5] Lahiri, A., Marslen-Wilson, W. (1991), "The mental representation of lexical form: a phonological approach to the recognition lexicon", *Cognition (in print)*.
- [6] Lindqvist, J. (1965) "Studies of the voice source by means of inverse filtering" *Speech Transmission Laboratory - Quarterly Progress and Status Report*, 2, Stockholm, 8-13
- [7] Lindqvist-Gauffin, J., Sundberg, J. (1976), "Acoustic properties of the nasal tract", *Phonetica*, 33, 161-168.
- [8] Makhoul, J. (1975), "Linear Prediction: a tutorial review", *Proceedings of the IEEE*, 63, 561-580.
- [9] Mermelstein, P. (1977), "On detecting nasals in continuous speech", *JASA*, 61, 581-587
- [10] Moll, K.L. (1962), "Velopharyngeal Closure on Vowels", *Journal of Speech and Hearing Research*, 5, 30-37
- [11] Reetz, H. (1989), "A Fast expert program for pitch extraction", *European Conference on Speech Technology, I*, Paris, 476-479.

TWO-FORMANT MODEL OF THE ACOUSTIC DESCRIPTION
OF SPEECH ARTICULATION

N. Degtyaryov

Institute Eng. Cybern. Ac. of Sc. BSSR

ABSTRACT

The present report suggests and consists a hierarchical model for obtaining an acoustic description of speech articulation. The primary description of speech articulation can be given by the parameters of two-formant model spectrum. A secondary, more precise articulation description is achieved by means of formant parameters of submodels, conditioned by different manners of speech sounds formation. A two-formant model of acoustic speech articulation description is suggested.

1. INTRODUCTION

In spite of the great efforts made, the problem of reliability in automatic extracting of formant parameters from speech signal, is far from being solved. The situation makes us think over some new approaches to the problem of formant speech analysis. The extracting methods of formant analysis of speech signals are based, as a rule, on a non-complete model of speech signals generation, the latter extracts amplitudes and frequencies of the first three-four voice formants /1,2/. That is why the most stable results of

analysis can be obtained only on the segments, matching with the given model. The reasons and character of the mistakes found here (loss of the 3rd and 4th formants because of their low level as compared with noises, loss of the 2nd formant because of its shunting during nasalization or low resolving power of spectrum analyser, etc./2/), demonstrate structural limitedness of the formant analysis models used as the result of that different characteristics of speech signal for different speech sounds, as regard the manner of their formation, are not taken into consideration. This brings us to the problem of an acoustic speech articulation description which considers structural characteristics of a formant model of speech signal generation.

2. STRUCTURAL MODEL OF SPEECH SIGNAL GENERATION

The universal theory of speech generation /3,4/ suggests an acoustic or equivalent electric submodels for each manner of speech sounds formation. The structure of each of the submodels is specific as it reflects articulation (speech organs shape, place and type of the excitation

source) of one particular manner. That's why each submodel can be described by means of its particular, different from the other submodels, set of significant formant characteristics. The necessity of consideration, while analysing, of these structural speech signal characteristics produced brings us to the following important conclusions. Firstly, the complete model of speech signal formant analysis must be structural and must include the submodels of speech sounds formation manners. Secondly, during the formant analysis process there must be a controlled commutation of submodels correlating with the nature of articulation manner of an analysed speech sound. Obviously such information can be obtained only by means of phonetic context hypothesizing.

3. HIERARCHY MODEL OF SPEECH ARTICULATION ACOUSTIC DESCRIPTION

Thus, the main conclusion we've come to, is that we can solve the problems of formant analysis of a speech signal only on the basis of a complete structural analysis model by means of synthesis using the information on the current phonetic context, hypothesized from the upper levels of a perception model. Some information on the problem can be found in /1,5/. Besides, the attempts to make a more or less complete mathematic speech signal model also result in a controlled structure /6/. According to our conception we must accept that the notion of a formant is conditioned by a definite phonetic context. So detailed for-

mant description can be given in symbolic and parametric representation and thus is secondary by nature. Then there must be a certain generalized, but unconditioned and in this sense primary acoustic description of articulation, which forms an initial stage in the process of speech analysis and recognition.

4. ACOUSTIC DESCRIPTION OF ARTICULATION BY MEANS OF GROUP FORMANTS CHARACTERISTICS

To create a system of parameters of the primary articulation description, let's consider some general characteristics of speech formants that can be found on the spectrum envelope. It happens so that we can extract four types of the spectrum envelope; they reflect the main formant characteristics of articulation (See Fig.1.) In /5/ a system of integral parameters A, F, B reflecting the formant characteristics of the speech spectrum envelope is presented. There is also an evaluation algorithm of the parameters. The main point of the algorithm that in two spectral regions with the adaptive boundary separating them, the formant groups are described by moments from spectrum counts, grouping around the maximum one:

$$1) A = \sum_1^m a_v / m$$

$$2) F = \sum_1^m f_v a_v / \sum_1^m a_v$$

$$3) B = \sum_1^m (f_v - f_{v-1}) a_v / \max_j a_j$$

a_j - counts of the instantaneous power spectrum on frequencies f_j ; $j = 1, 2, \dots$

$v = \alpha z Z_j$; $Z_j = \text{sign}(a_j - h \max a_j)$; $0 < h < 1$; m - the number of counts, above the threshold $h \max a_j$.

Here is the physical meaning of the parameters evaluated by (1)-(3): they express the integral amplitude, frequency and band values of the spectral counts Q_j , representing this formant group. The main qualities of the suggested parameters system and the algorithm of their extraction (separation) are: 1) the possibility of separating of the two first formants even in the case of their mutual (reciprocal) masking /5/; 2) the possibility of reflecting different formant characteristics (See Fig.1b and 1c) without separating the upper formants, that is, of course, the most difficult problem; 3) equal efficiency for reflecting formant characteristics of different, from the point of view of their manner of articulation, sounds.

Fig.2 presents A,F parameter tracks for the words "ə'din" and "sə'se".

5. CONCLUSION

The suggested system of A,F,B parameters of the speech articulation primary description is a good basis for the upper level analysis of the speech recognition model. Firstly, the parameters precisely reflect the speech spectrum formant characteristics. Secondly, they meet the demands of the linear model of parameter approximation /6/, which is a way towards the solution of the speech recognition problems. Thirdly, the suggested model provides the basis for des-

cribing some topologic invariants and, thus, contributes to the solution of the multispeaker recognition problems.

6. REFERENCES

- /1/ КОПЕЦ, Г. (1986) "Formant tracking using hidden Markov models and vector quantization", IEEE Trans. on acoust. speech and signal process. Vol. AISP-34, No 4, 709-729.
- /2/ БУХТИЛОВ Л., ЛОБАНОВ Б. (1988) "Алгоритм оценки формантных частот", Автоматическое распознавание слуховых образов (АРСО-14). Ч.1: Материалы докладов и сообщений. Каунас, 10-11.
- /3/ FANT, G. (1960), "Acoustic theory of speech production", The Hague: Mouton & Co.
- /4/ FLANAGAN, G. (1965) "Speech analysis, synthesis and perception", Springer, Berlin-New-York.
- /5/ ДЕГТЯРЕВ, Н. (1988) "Двухформантная аппроксимация спектров речи", Автоматическое распознавание слуховых образов (АРСО-14). Ч.1: Материалы докладов и сообщений. Каунас, 12-13.
- /6/ ВИНЦЮК, Т. (1982) "О математических моделях речевого сигнала, используемых в распознавании речи", Автоматическое распознавание слуховых образов (АРСО-12): Тезисы докладов и сообщений. Киев, 34-37.

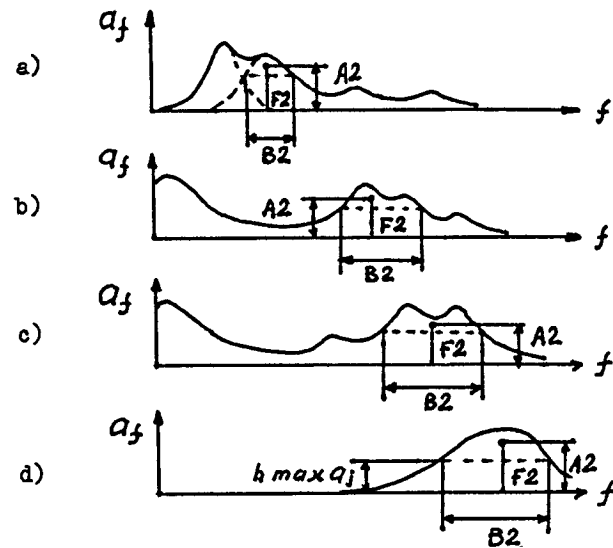


Fig.1. Formant groups and their generalised parameters for a) compact and aspirative; b) diffuse; c) nasal and voiced fricative; d) unvoiced fricative sounds of speech.

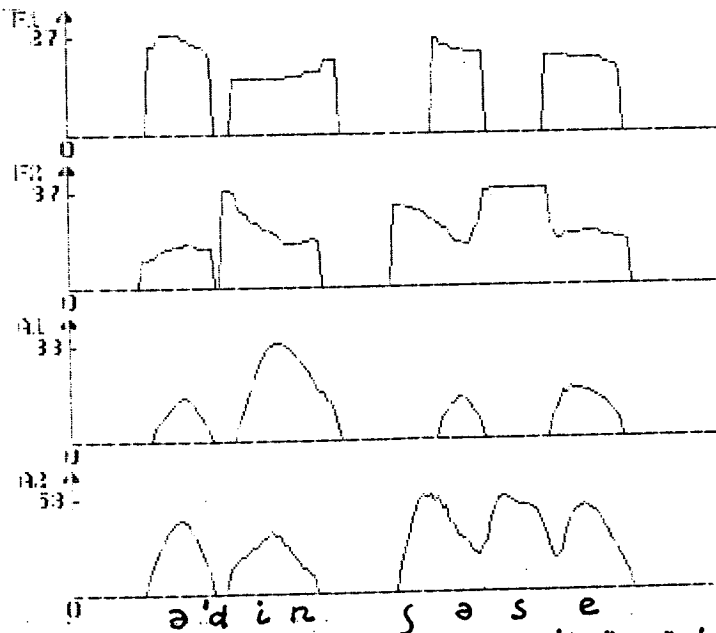


Fig.2. A,F parameter tracks for the words "ə'din" and "sə'se".

ACOUSTIC DESCRIPTION OF SPANISH NASAL CONSONANTS IN CONTINUOUS SPEECH

María Jesús Machuca Ayuso

Departament de Filologia Espanyola
Laboratori de Fonètica-Spain

The aim of this paper is to compare some acoustic cues of the Spanish nasal consonants in two different speaking styles: Continuous speech versus laboratory speech.

The items analyzed in continuous speech have been obtained from an hour recording of speech produced by a male native Spanish speaker. In order to obtain the items in laboratory speech, the same subject was instructed to read the set of analyzed utterances in citation form. Broadband spectrograms were made with a MacSpeech Lab II programme to analyse the consonant duration and nasal formant frequencies. The results suggest that between the analyzed speaking styles, differences in duration are more important than in formant frequencies.

INTRODUCTION

There is an agreement among phoneticians on the relevance of the following features that characterize the nasals as a class (Delattre[1]; Fant[2]; Fujimura [3]): a low first formant at about 300Hz, with higher intensity than the upper formants, the existence of the antiformant and a set formants which have a lower intensity level than those of neighbouring vowels. Furthermore, Fant (1960) states that there is a relationship between nasal and vowel formants. There

is a correspondence of N1 to F1, N3 to F2, and in some instances N4 to F3. This correspondence is a matter of continuity since a formant may be equally dependent on the nasal oral cavities.

Fujimura (1962) distinguishes between variable and invariant formants. According to this author there are some formants that remain relatively unaffected by the context: *The second formant of the [n], for example, is located always around 1000 cps. for all samples of [n] (p.246).*

In our work, we have taken into account the consonant duration and nasal formant frequencies.

PROCEDURE

We used for this experiment an hour recording of speech produced by a male native speaker of Spanish. Natural continuous speech was obtained by asking questions about the subject's work, military service, his village and his family.

"Laboratory speech" refers to utterances read in citation form by a speaker.

In order to obtain the items in laboratory speech, the same subject was instructed to read the set of utterances that appear in continuous speech corpus as naturally as possible at a normal speech rate. The order of these utterances was randomized. The recordings were made in a sound proof room at the phonetics laboratory at

the Autonomous University of Barcelona, using a Revox A/77 tape recorder and a Shure 515 Sb Unidyne microphone.

Formant frequencies were made of nasal consonant in intervocalic context. In some of the cases, the vowel that followed the consonant was stressed and in other cases it was unstressed. The number of items was eight hundred and thirty: three hundred and seventy samples of which correspond to the first context and four hundred and sixty to the second one. Table I shows the number of analyzed cases for each consonant.

Table I: Number of analyzed cases for Spanish nasal consonants in both contexts.

	VCV	VCV
[m]	199	142
[n]	153	249
[ɲ]	24	70

The recorded speech material was digitised at 10 KHz and analysed by means of spectrograms using a MacSpeech Lab II programme. In the selected sequences the duration and the formant frequencies were measured.

RESULTS

The data obtained from the acoustical analysis were subjected to a statistical treatment. The mean values of the given parameters (X) as well as the standard deviation (sd) are shown in the following tables. Table II correspond to laboratory speech values of nasal formants in the context VCV and VCV in laboratory speech and Table III shows the same values in continuous speech.

Table II: Nasal formant values in laboratory speech when the nasal consonant is followed by a stressed or by an unstressed vowel.

VCV	X	sd	VCV	X	sd
[m]	351	33	[m]	348	28
	1071	74		1070	43
	1038	282		783	116
	2194	68		2239	69
[n]	361	29	[n]	370	21
	1032	45		1039	28
	1606	172		1599	183
	2318	51		2323	51
[ɲ]	375	24	[ɲ]	375	20
	1062	29		1064	30
	2001	134		2076	141

Table III: Nasal formant values in continuous speech when a nasal consonant is followed by a stressed or by an unstressed vowel.

VCV	X	sd	VCV	X	sd
[m]	400	66	[m]	414	64
	1060	99		1022	70
	1309	151		1118	417
	2123	96		2177	90
[n]	427	64	[n]	430	54
	1002	66		1042	54
	1562	167		1500	192
	2261	90		2287	77
[ɲ]	468	44	[ɲ]	433	44
	1048	28		1048	40
	1943	130		2009	126
	2334	28		2455	67

If we compare these results, it can be observed that the first formant frequencies in continuous speech are higher than in laboratory speech. In the second formant frequencies there are no significant differences in both speaking styles. In the third formant, the frequencies are higher in laboratory speech, except to the consonant [m]. It can be noted that in laboratory speech N3 of the consonant [m] has values below N2. This finding is due to the influence of the vowels. The fourth formant frequencies have similar results for [m] and [n]

consonants, if we compare both speaking styles. Nevertheless, in laboratory speech, this formant is impossible to be distinguished in the palatal consonant. On the other hand, standard deviation values of the N3 (Tables II and III) show an important dispersion in our data. For this reason, the results of the third nasal formant have been separated depending on vowel contexts. Tables IV and V show these results in continuous speech (CS) and in laboratory speech (LS) taking into account the vowel stress.

Table IV: Values of N3 when the consonant is followed by an unstressed vowel ("pal" and "vel" are palatal and velar contexts and the stress means that the preceding vowel to the consonant is stressed).

VCV	CS	LS
pál [m] vel	1378	834
pál [m] a	1651	—
pál [m] pal	1253	1275
vé [m] pal	1348	1203
vé [m] a	—	797
vé [m] vel	735	727
á [m] vel	—	749
á [m] a	—	—
á [m] pal	—	—
<hr/>		
pál [n] vel	1565	1643
pál [n] a	1616	1695
pál [n] pal	1740	1712
vé [n] pal	1431	1485
vé [n] a	1231	1269
vé [n] vel	—	—
á [n] vel	1299	1340
á [n] a	1290	1388
á [n] pal	1475	1477
<hr/>		
pál [p] vel	1981	2192
pál [p] a	2162	2242
pál [p] pal	—	—
vé [p] pal	—	—
vé [p] a	1998	2212
vé [p] vel	—	—
á [p] vel	1943	2044
á [p] a	2003	2111
á [p] pal	—	—

It can be observed in table IV that laboratory speech results are lower than continuous speech ones for the consonant [m], but in the other nasal consonants they are similar. In some contexts, there are not any data because the number of cases is insufficient for a statistic treatment.

Table V: Values of N3 when the consonant is followed by a stressed vowel ("pal" and "vel" are palatal and velar contexts and the stress means that the following vowel to the consonant is stressed).

VCV	CS	LS
pal [m] vél	—	—
pal [m] á	—	—
pal [m] pá [l]	1339	1271
vel [m] pá [l]	—	916
vel [m] á	—	776
vel [m] vél	—	700
a [m] vél	—	745
a [m] á	—	—
a [m] pá [l]	—	1246
<hr/>		
pal [n] vél	1697	1722
pal [n] á	1464	1599
pal [n] pá [l]	1630	1705
vel [n] pá [l]	1327	1436
vel [n] á	—	—
vel [n] vél	—	1222
a [n] vél	1357	1261
a [n] á	1419	1359
a [n] pá [l]	1469	1497
<hr/>		
pal [p] vél	2088	2061
pal [p] á	2094	2118
pal [p] pá [l]	—	1985
vel [p] pá [l]	—	—
vel [p] á	—	—
vel [p] vél	—	—
a [p] vél	—	—
a [p] á	1917	2000
a [p] pá [l]	1926	2013

In table V, the continuous speech results are usually lower than the results in laboratory speech.

On the other hand, very different results are shown in both tables if we compare velar and palatal context. N3 goes down in velar context and it goes up in palatal one.

The duration results are shown in table VI. We have separated the cases depending on the vowel stress and the speaking style.

Table VI: Duration of nasal consonants according to vowel context.

Context	LS	SD	CS	SD
√ [m] V	92	16	55	16
√ [n] V	64	12	41	11
√ [p] V	98	13	64	16
V [m] √	66	11	61	13
V [n] √	48	9	37	12
V [p] √	69	12	59	11

It can be noted in table VI that the duration is much smaller in continuous speech. But when the consonant is followed by an unstressed vowel the duration difference between both styles is greater (47 ms for [m], 23 for [n] and 38 for [p]). In these cases, it can be observed that [p] > [m] > [n] in all samples.

Nevertheless, when the nasal consonant is followed by a stressed vowel [m] > [p] > [n] in continuous speech and [p] > [m] > [n] in laboratory speech.

CONCLUSIONS

The results presented in the tables above show the relevance of the duration when we compare continuous speech versus laboratory speech. The difference is more significant if we take into account the vowel stress.

The frequency results show differences in the first formant: in continuous speech it is higher than in laboratory speech. But we don't observe important differences in N2, N3 and N4. Only, the third formant frequency in [m] consonant goes down

in laboratory speech with relation to continuous speech results.

On the other hand, in both styles, the frequency of N1, N2 and N4 formants is relatively constant while the N3 frequency varies considerably with the context, having a high frequency before the vowels [i] and [e] and a low frequency before the vowels [o] and [u].

This fact could be explained following Fant's assumptions about the dependence between formants and cavities: N3 could be dependent on the oral cavities. Future research would be necessary in order to decide if this dependence on the oral and nasal cavity is related to the speaker.

REFERENCES

- [1] DELATTRE, P. (1958), "Les indices acoustiques de la parole", *Phonetica*, 2, 1-2:108-118; 3-4: 226-251.
- [2] FANT, G. (1960), "Acoustic theory of speech production", The Hague: Mouton & Co.
- [3] FUJIMURA, O. (1962), "Analysis of nasal consonants", *Journal Acoust. Soc. Amer.*, 34,1865-1875.

DISAMBIGUATING SENTENCES USING PROSODY

Patti Price, SRI International, Menlo Park, CA USA

Mari Ostendorf, Boston University, Boston, MA USA

Stefanie Shattuck-Hufnagel, MIT, Cambridge, MA USA

ABSTRACT

Prosodic information can provide cues to syntactic structure to help select among competing hypotheses, and thus to disambiguate otherwise ambiguous sentences. We show that some, but not all, syntactic structures can be disambiguated via prosody. The phonological evidence relates the disambiguation primarily to boundary phenomena. Phonetic analyses indicate the importance of both absolute and relative measures. Finally, we describe initial experiments involving the automatic use of this information in parsing.

1. INTRODUCTION

The syntax of spoken utterances is frequently ambiguous, yet communication generally succeeds. This success may arise from a variety of sources; we address here the role of prosody. A clear understanding of the mapping between prosodic and syntactic structure would reveal significant aspects of the cognitive processes of speech production and perception. In addition, it would yield more natural sounding speech synthesis. Further, prosody should be particularly helpful in spoken-language understanding, where lexical and structural ambiguities of written forms are compounded by difficulties in finding word boundaries and in identifying words reliably in automatic speech recognition. Here, we study the mapping between prosody and syntax by minimizing the contribution of other possible cues to the resolution of ambiguity.

With few exceptions (e.g., [7]), previous studies have focussed either on relating phonological aspects of prosody to syntax (e.g., [5], [12], [2], [9]), or on relating phonetic/acoustic evidence to syntax and perceived differences (e.g., [15], [3], [16], [6], [8], [4], [17]). A few studies, e.g., [13], have considered the mapping from phonology to acoustics. The more phonetic/acoustic studies typically used a small number of minimal pairs of utterances in order to facilitate the acoustic measurements and to control parameters more precisely. In contrast, the more phonological studies have focussed either on 'illustrative examples' or on text to which prosodic markers have been assigned on the basis of the syntax of the sentence. These studies have typically ignored the

fact that there are several possible prosodic choices for a given syntactic structure. The focus in recent theoretical linguistics on human competence for language production, has resulted in neglect of actual language production and neglect of an area required for speech understanding (by human or by machine): the mapping from acoustics to meaning. Clearly, speech communication involves both production and perception, and it involves performance as well as competence.

The work presented in this paper extends previous work, including the important contribution of [10], in several ways: (1) we investigate the ability of listeners to disambiguate sentences for different types of syntactic structures, using several instances of each type; (2) we consider both production and perception; (3) to increase reliability without assessing a large pool of subjects, we used professional FM radio announcers; (4) we have investigated the possible use of prominence associated with pitch accents, in addition to prosodic phrase boundary cues; (5) to compare durational structures across the various sentences used, and to facilitate generalization beyond the specific sentences used, we present results in terms of relative, rather than absolute, durational patterns; and (6) we consider the automatic use of prosodic information in parsing.

2. CORPUS

We used 35 sentence pairs, ambiguous in that members of each pair contained the same string of phones, and could be associated with contrasting syntactic bracketings. Sentences represented 5 instances each of 7 types of structural ambiguity: (1) parenthetical clauses vs. non-parenthetical subordinate clauses, (2) appositions vs. attached noun or prepositional phrases, (3) main clauses linked by coordinating conjunctions vs. a main clause and a subordinate clause, (4) tag questions vs. attached noun phrases, (5) far vs. near attachment of final phrase, (6) left vs. right attachment of middle phrase, and (7) particles vs. prepositions.

Each pair of ambiguous sentences was preceded by a disambiguating context. For structural categories 1-4, sentence A of the pair involved a larger syntactic break than sentence B. For the attachment ambiguities 5-7, sentence A of the pair had the larger syntactic break later in the sentence than did sentence B. When sentences were recorded by the 4 FM newscasters, con-

trasting members of a pair did not occur in the same session. Speakers were not told there were target sentences, and recording sessions were separated by a few days to several weeks. Our goal was to create segmentally identical but syntactically different sentence pairs.

3. PERCEPTUAL EXPERIMENT

The target sentences were presented in isolation. The 35 sentence pairs produced by each speaker were presented to listeners in two sessions; only one member of each pair was heard in each session (analogous to the strategy used for recording the sentences). The answer sheet included both disambiguating contexts followed by the target sentence. Listeners marked the context they thought best matched what they heard. Subjects were all native speakers of American English, naive with respect to the purpose of the experiment. The number of listeners who heard both sessions ranged from 12 to 17 for the different speakers.

In scoring, we assume speakers produced the intended version, and a *correct* response identifies that version. Accuracy is the percentage of correct listener responses. Table 1 summarizes accuracy for the different structural types. Averages are over the 4 speaker averages, so as not to more heavily weight the utterances that were heard by more listeners.

Type	A	B	Overall
1. Parenthetical or not	77	96*	86
2. Apposition or not	92*	91*	92
3. M-M vs. M-S	88*	54	71
4. Tags or not	95*	81	88
5. Far/near attachment	78	63	71
6. Left/right attachment	94*	95*	95
7. Particle/Preposition	82*	81*	82
Average	87	80	84

Table 1. Perceptual results, averaged over 4 speakers. Version A/B figures are based on 285 observations of each class. An asterisk marks A and B responses with high listener accuracy, where high accuracy is when (average accuracy minus Standard Deviation) > 50%.

Table 1 shows that subjects could reliably disambiguate many, but not all of the ambiguities. On average, subjects did well above chance in assigning sentences to appropriate contexts. Main-subordinate (3B) sentences and near attachments (5B) were close to the chance level; parentheticals (1A), far attachments (5A) and non-tags (4B) were recognized at levels greater than chance but not reliably; all other sentence types were reliably disambiguated.

4. PHONOLOGICAL ANALYSIS

The perceptual experiments show that speakers can encode prosodic cues to structural ambiguities in ways that listeners can use reliably. This section attempts to find a phonological answer to the question: How do they do it? To approach this question, we labeled discrete phenomena that could mark structural contrasts phonologically. We then analyzed the relationship between these labels and patterns in the perceptual study.

We used 7 levels to represent perceptual groupings (or, degrees of separation) between words. These levels appeared adequate for our corpus and also reflected the levels of prosodic constituents described in the literature. We used numbers to express the degree of decoupling between each pair of words as follows: 0 - boundary within a clitic group, 1 - normal word boundary, 2 - boundary marking a grouping of words generally having only one prominence, 3 - intermediate phrase boundary, 4 - intonational phrase boundary, 5 - boundary marking a grouping of intonational phrases, and 6 - sentence boundary.

Break indices of 4, 5, and 6 are *major* prosodic boundaries; constituents defined by these boundaries are marked by a boundary tone and are often referred to as 'intonation phrases'. Boundary tones were labeled using 2 types of falls (final fall and non-final fall), and 2 types of rises (continuation rise and question rise). Prominent syllables were labeled using P1 for major phrasal prominence; P0 for a lesser prominence; and C for contrastive stress (which occurred on fewer than 1% of the total words). The prosodic cues were labeled perceptually by 3 listeners using multiple passes. Correlation across labelers was 0.96.

In general, we found prosodic boundary cues associated with almost all reliably identified sentences. A break index of 4 or 5 was often, but not always, a reliable cue, and was most often observed at embedded or conjoined clause boundaries (often marked by commas in the text). A difference in the relative size of prosodic break indices, or in the location of the *largest* break, was frequently the only disambiguating cue for smaller syntactic constituents (i.e., where fewer brackets would coincide). By and large, larger break indices tended to mean that syntactic attachment was higher rather than lower. Prominence seemed to play a supporting role, and was the sole cue in only a few sentences. Details of these results analyzed by structural types appear in [14].

The main exception to this picture was the main-main (A) vs. main-subordinate (B) sentences. The A versions were typically well-identified, whereas the B versions tended to be close to the chance level. This could be the result of a syntactic response bias. The difference is interesting since the bracketings differ for the 2 versions of the sentence, and yet they are apparently not well separated perceptually. The prosodic transcriptions suggest a rea-

son: both versions have a major prosodic boundary in the same location.

5. PHONETIC ANALYSIS

We have presented perceptual evidence that naive listeners can reliably use prosody to separate structurally ambiguous sentences, and phonological evidence that suggests how listeners might use prosody to select among syntactic hypotheses. In this section we consider phonetic evidence that might be responsible for the prosodic disambiguation. We examine duration and intonation, although we acknowledge that other cues, such as the application or non-application of phonological rules, contribute to the perception of prosodic boundaries. We tried to minimize such effects by asking speakers to reread sentences in which overt segmental cues were produced.

Segment duration was determined automatically using an HMM-based speech recognition system, the SRI Decipher system [18]. Each phone duration was normalized according to speaker- and phone-dependent means as described in [11]. We observed longer normalized durations for phones preceding major phrase boundaries and for phones bearing major prominences compared to other contexts. We measured average normalized duration in the rhyme of word-final syllables and found that higher break indices are generally associated with greater normalized duration. Pauses are also associated with major prosodic boundaries, occurring at 48/212 (23%) boundaries marked with 4 and 17/25 (67%) boundaries marked with 5. No pauses were found after a 0, 1, or 2, and only one pause occurred after a 3.

Analysis of normalized duration of vowel nuclei revealed: (1) major prominences (P1, C) tend to be longer than unmarked or minor (P0) prominences, although the effect is small before major prosodic breaks (where duration is already lengthened); (2) word-final syllables tend to be longer than non-word-final syllables; (3) syllables are longer in words before major breaks than in words before smaller breaks, especially for word-final syllables; and (4) the effects seem to be somewhat independent: the longest vowels are those with a major prominence, in word-final position, before a major break.

Intonational cues included boundary tones, pitch range changes and pitch accents. Boundary tones are involved for break indices 4 - 6. Sentence-final boundary tones are typically either final falls or question rises; level 5 boundary tones were usually labeled non-final falls; and level 4 boundary tones were most often continuation rises, but occasionally non-final falls. Another intonational cue was a drop in pitch baseline and range in a parenthetical phrase, relative to the context. This pitch range change was not always apparent for appositives. Though intonation is an important cue, duration and pauses alone provide enough information to automatically label

break indices with a high correlation (greater than 0.86) to hand-labeled break indices [11].

6. AUTOMATING DISAMBIGUATION

We have shown that listeners can pay attention to prosodic information, and we have shown phonological and phonetic evidence bearing on how this might be done. The next step is to be explicit enough about the use of the phonetic evidence that it could be used automatically to select the appropriate parse. In our initial attempt, since there was a good correlation between normalized rhyme duration and the hand-labeled break indices, we used a 7-state Gaussian HMM to convert automatically estimated duration values to break indices [11], and passed to the parser a break index between every pair of words. This procedure required modification of the existing grammar to handle the new break index category and to allow for empty nodes and their interaction with the break indices. The grammar before and after these changes yields the same number of parses for a given sentence.

In order to make use of the prosodic information an additional important change is required: how does the grammar use this information? This is a vast area of research. In this initial endeavor, we focussed on prepositional phrases, and made very conservative changes. We changed the rule $N \rightarrow N \text{ link PP}$ so that the value of the link (break-index) must be less than 3 for the rule to apply. We made essentially the same change to $VP \rightarrow V \text{ link PP}$, except that the value of the link must be less than 2.

After setting these two parameters we parsed each of the sentences in the 14 sentences in our corpus containing prepositional phrase attachment ambiguities or particle-preposition ambiguities, and compared the number of parses to the number of parses obtained without benefit of prosodic information. For half of the sentences, i.e., for one member of each of the sentence pairs, the number of parses remained the same. For the other member of the pairs, the number of parses was reduced, on average to half the previous number. Thus, the incorporation of the prosodic information resulted in a net reduction of about 25% in the number of parses, without ruling out any correct parses. In many cases the use of prosodic information allowed the parser to identify a unique parse. More details on these procedures and results appear in [1].

7. DISCUSSION

We have confirmed that, for a variety of syntactic classes, but not all, naive listeners can reliably separate meanings on the basis of differences in prosodic information. We have further shown phonological and phonetic evidence bearing on how they might do this: by tendency to associate relatively larger prosodic breaks with larger syntactic breaks. Though evidence relating to boundary phenomena appeared to be most important, there

were some structures for which phrasal prominence either was the only cue or played a supporting role in distinguishing between the 2 versions. Further, we have presented initial evidence showing how extracted phonetic information (normalized duration) can be automatically extracted and communicated to a parser to reduce ambiguity.

Our results have both theoretical and empirical implications. In speech generation applications, the relation between syntax and prosody is important since different prosodic markers will affect the interpretation of a sentence. Prosodic cues are particularly important in computer speech understanding applications, where the semantic rules available to the system are limited relative to the capabilities of human listeners. In addition, in these applications, prosodic cues can be used prior to semantic analysis, to reduce the number of syntactically acceptable parses by eliminating those inconsistent with the prosody [1].

The results reported here provide evidence for systematic relationships between prosody and syntax that should be explored further in several ways. First, a larger number of syntactic structures must be examined in order to make the prosody/syntax relationship more explicit. Second, we note that some sentences were successfully disambiguated with cues that were not represented in our labeling scheme. Since prominences were not differentiated as to type of pitch accent, a more detailed classification of intonation in such contexts could yield more information. Finally, for computer speech understanding applications, it will be important to investigate the extension of these results to spontaneous speech by non-professional speakers, where hesitation phenomena and speech errors will affect the prosodic structure.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IRI-8805680 in coordination with DARPA/NSF funding under NSF Grant No. IRI-8905249. The government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the government funding agencies. We thank A. Levitt and L. Larkey for help in generating the ambiguous sentences; the WBUR announcers who recorded the sentences; the subjects who participated in the experiments; N. Veilleux and C. Wightman for many hours of prosodic labeling. We thank J. Bear for providing syntactic bracketings and for grammar modifications and parsing results, and we thank C. Wightman for the automatic labeling of break indices.

8. REFERENCES

[1] Bear, J. and Price, P. (1990). "Prosody, syntax and parsing," *Proc. ACL*.

[2] Bing, J. (1984). "A discourse domain identified by intonation," pp. 11-19 in *Intonation, Accent and Rhythm: Studies in Discourse Phonology*, New York, de Gruyter.

[3] Cooper, W. and Sorensen, J. (1977). "Fundamental frequency contours at syntactic boundaries," *J. Acoust. Soc. Am.* 62:3, 683-692.

[4] Duez, D. (1985). "Perception of silent pauses in continuous speech," *Language and Speech* 28:4, 377-389.

[5] Gee, J. P. and Grosjean, F. (1983). "Performance structures: A psycholinguistic and linguistic appraisal," *Cognitive Psychology* 15, 411-458.

[6] Garro, L. and Parker, F. (1982). "Some suprasegmental characteristics of relative clauses in English," *J. Phonetics* 10, 149-161.

[7] Geers, A. (1978). "Intonation contour and syntactic structure as predictors of apparent segmentation," *J. Exp. Psych: Hum. Perc. and Perf.* 4:3, 273-283.

[8] Kutik, E., Cooper, W., and Boyce, S. (1983). "Declination of fundamental frequency in speakers' production of parenthetical and main clauses," *J. Ac. Soc. Am.* 73:5, 1731-1738.

[9] Ladd, D. R. (1986). "Intonational phrasing: the case for recursive prosodic structure," *Phonology Yearbook* 3, 311-340.

[10] Lehiste, I. (1973). "Phonetic disambiguation of syntactic ambiguity," *Glossa* 7:2, 107-121.

[11] Ostendorf, M., Price, P., Bear, J. and Wightman, C. (1990). "The use of relative duration in syntactic disambiguation," *Proceedings of the 3rd DARPA Workshop on Speech and Natural Language*.

[12] Nespor, M. and Vogel, I. (1983) "Prosodic structure above the word," *Prosody: Models and Measurements*, Cutler and Ladd, eds., Springer-Verlag, pp. 123-140.

[13] Pierrehumbert, J. (1981). "Synthesizing intonation," *J. Ac. Soc. Am.* 70:4, 985-995.

[14] Price, P., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. (1991). "The use of prosody in syntactic disambiguation," *Proc. 4th Darpa Workshop on Speech and Natural Language*, ed. P. Price, Morgan Kaufman. A longer version of this paper has been submitted for journal publication.

[15] Scholes, R. (1971). "On the spoken disambiguation of superficially ambiguous sentences," *Language and Speech* 14, 1-11.

[16] Thorsen, N. (1980). "A study of the perception of sentence intonation -- Evidence from Danish," *J. Ac. Soc. Am.* 67:3, 1014-1030.

[17] Thorsen, N. (1985). "Intonation and text in standard Danish," *J. Ac. Soc. Am.* 77:3, 1205-1216.

[18] Weintraub, M. et al. (1989). "Linguistic constraints in hidden Markov model based speech recognition," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 699-702.

SIGNALISATION PROSODIQUE DE LA STRUCTURE INFORMATIONNELLE DANS LE DISCOURS RADIO-PHONIQUE EN FINNOIS ET EN FRANÇAIS

Veijo V. Vihanta

Université de Tampere, Finlande

ABSTRACT

This paper studies some aspects of the prosodic signalling of the information structure in Finnish and French based on radio news reports. The investigation is carried out on two levels, perceptual and acoustic. In Finnish the focalisation is a kind of intensified accentuation process whereas in French it is an autonomous process not depending on accentuation.

1. INTRODUCTION

Le rôle de la prosodie dans la signalisation de la structure informationnelle du message parlé a fait l'objet d'études diverses à partir de l'École de Prague pour fournir une meilleure explication de l'interaction des faits prosodiques et syntaxiques dans l'organisation des énoncés. Dans un texte récent (Vihanta 1990), j'ai examiné l'organisation du message dans les informations radiophoniques finnoises et françaises du point de vue de la structuration externe, c'est-à-dire les moyens utilisés pour diviser le message ainsi que les relations entre les unités résultantes, paragraphes, phrases, propositions, syntagmes et mots. Je continuerai sur ce thème ici en examinant certains aspects de la structure informationnelle interne, notamment sa signalisation par les moyens prosodiques sous forme de focalisation. Il s'agit de deux langues typologiquement très différentes du point de vue de l'organisation et de la structuration du message, que ce soit aux niveaux syntaxique, morphologique ou phonétique. En ce qui concerne les moyens prosodiques, les deux langues semblent recourir à des procédés assez divergents. En finnois, langue à accent fixe sur la première syllabe du mot et à dynamisme prosodique général descendant, on a la

possibilité d'accentuer, ou de focaliser, n'importe quel mot selon l'importance qui lui est accordée dans l'énoncé. En cela, le finnois se rapproche des langues germaniques comme l'anglais.

Par contre, on a traditionnellement nié l'existence d'un procédé uniquement prosodique de ce genre en français en prétendant que ce que l'anglais réalise par l'accent de phrase, le français le réalise par des moyens syntaxiques. On a pourtant vu, ces dernières années, une réinterprétation intéressante de Rossi sur ce point de la prosodie française. Le phénomène connu sous le nom d'"accent d'insistance", terme utilisé pour désigner les différents procédés de "mise en relief", considérés comme expressifs, émotionnels, intellectuels, rhétoriques etc., c'est-à-dire des procédés ne faisant pas partie du code linguistique proprement dit, vient d'obtenir un rôle de première importance dans la hiérarchisation de l'information. Rossi l'a rebaptisé "accent énonciatif" (AE) (1985 et 1987). La façon d'interpréter le système accentuel du français est en effet cruciale pour rendre compte du rôle de la prosodie dans l'organisation de l'information des énoncés français.

2. CORPUS ET MÉTHODES

Le corpus sur lequel est basé l'analyse présentée ici est formé de deux bulletins d'informations, d'une durée de 8'20" tous les deux. Dans ce type de discours il s'agit presque uniquement de la lecture d'un texte rédigé à l'avance, dans lequel la structuration syntaxique a déjà été effectuée au préalable, et qui est formé principalement sur le modèle de la langue écrite. La seule liberté, toute relative, qui reste au locuteur, au moment de la production du message, est celle d'utiliser les moyens prosodiques, dans le cadre syntaxique

donné, pour hiérarchiser l'information. Une combinaison d'analyse perceptuelle et d'analyse acoustique est adoptée pour découvrir et analyser les proéminences perceptuelles dans le message. Les syllabes ou mots proéminents sont d'abord repérés et classifiés à l'oreille, après de nombreuses écoutes attentives. Ce n'est qu'après une décision perceptuelle qu'ils ont été examinés au niveau de la réalisation des paramètres prosodiques, ou indices acoustiques, tels que l'utilisation de pauses et de glottalisations, de la Fo et de l'intensité, ainsi que de la durée des segments phoniques. L'analyse est effectuée au moyen de l'ISA (Intelligent Speech Analyzer), un système numérique de traitement de signal de parole développé par Raimo Toivonen.

De toutes les proéminences perçues, ne seront traités ici que celles dont la fonction est liée à la structure informationnelle; le terme retenu pour désigner ce phénomène est la focalisation, le terme proéminence étant utilisé ici uniquement comme catégorie perceptuelle indépendamment de sa fonction. De plus, ont été repérés tous les accents initiaux, dans le corpus français. Par accent initial j'entends l'accentuation perçue, sur la syllabe non finale, normalement sur la première, quelquefois sur la deuxième syllabe d'un mot, mais qui elle, est inférieure à la proéminence, sauf dans le cas de la coïncidence des deux.

3. ANALYSE PERCEPTUELLE

Dans le bulletin d'information finnois le nombre total des proéminences est de 135. La grande majorité, 123 en tout, a comme fonction principale ou unique d'organiser l'information, c'est à dire de focaliser un élément important à l'intérieur du thème ou du thème. 9 peuvent être qualifiées de démarcatives et 3 de continuatives. Le ton neutre et objectif exclut pratiquement la fonction expressive.

Dans le bulletin d'information français, le nombre total des proéminences perceptuelles est de 102. De ces proéminences, 43 coïncident avec l'accent initial, 46 sont réalisées sur un mot d'une seule syllabe, les 13 restantes étant réalisées sur la dernière syllabe d'un mot à deux ou trois syllabes. Le nombre des accents initiaux sans proéminence est de 69, ce qui donne en tout 112 accents initiaux.

Parmi les proéminences, 85 ont comme fonction principale ou unique d'organiser l'information. Le reste des proéminences ont des fonctions démarcatives, continuatives ou expressives. L'expressivité peut se superposer à toutes les fonctions.

4. ANALYSE ACOUSTIQUE

À défaut de présenter une analyse exhaustive des différents paramètres acoustiques utilisés pour signaler l'organisation des énoncés, je devrais me contenter d'illustrer certaines tendances centrales à l'aide d'exemples typiques. Les figures 1a et 1b présentent une phrase entière, initiale d'un sujet, des informations finnoises *Unkarin parlamentti hyväksyi tänään maalle tiukan säästöbudjetin, joka avaa tien kansainvälisen valuttarahaston miljoonien dollarien lainojen virtaamiselle Unkariin* ('Le parlement de la Hongrie a approuvé aujourd'hui pour le pays un strict budget d'épargne qui ouvrira la voie à l'afflux des prêts de millions de dollars du fonds monétaire international'). Les focalisations sont marquées par une double barre.

La première focalisation est en réalité le résultat de l'accumulation de deux fonctions: premièrement, le signal du début de paragraphe, c'est à dire un nouveau sujet, et donc démarcatif vis à vis du paragraphe précédent, et deuxièmement la présentation du thème. Elle est réalisée par une hauteur initiale très élevée de la Fo et par une intensité forte (cf. Iivonen 1990). C'est un cas très typique pour les informations, où il est important de marquer le changement de sujet d'une part et de focaliser l'attention de l'auditeur sur le sujet suivant d'autre part, puisque le passage se fait d'une manière brusque et sans progression thématique. Les focalisations sur *tiukan* et *säästö-* n'en font en réalité qu'une seule; c'est la partie la plus importante du thème de la première proposition. Dans la proposition subordonnée, qui forme le thème vis à vis du thème présenté dans la première proposition, l'apport d'information principal est signalé par la focalisation de *ien* et *miljoonien*. Ces focalisations sont réalisées par une montée rapide et importante de la Fo et une montée simultanée de l'intensité. Elles sont normalement réalisées sur la première syllabe, sauf dans quelques rares exceptions dues à la structure syllabique du mot en question.

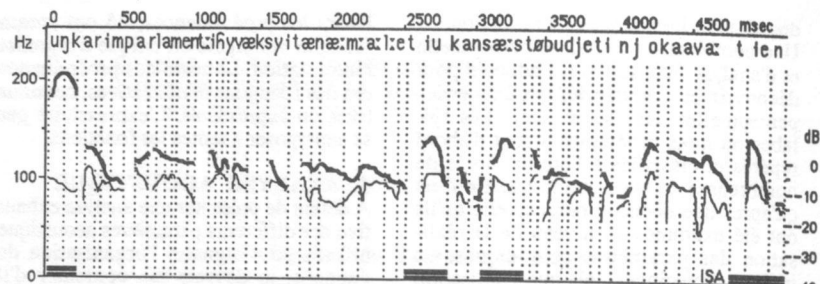


Figure 1a. Fo (courbe épaisse) et Ao (courbe mince) d'un extrait des informations finlandaises *Unkarin parlamentti hyväksyi tänään maalle tiukan säästöbudjetin, joka avaa tien*

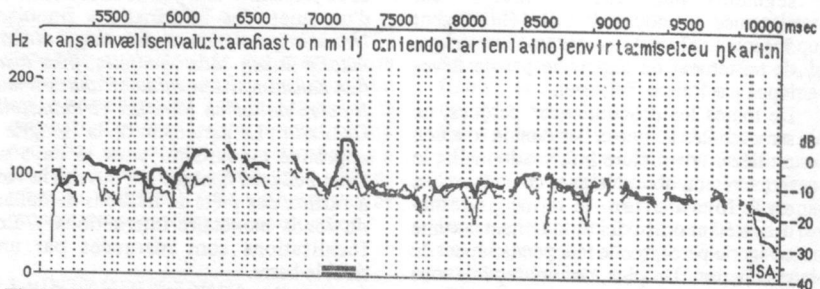


Figure 1b. (suite) *kansainvälisen valuuttarahaston miljoonien dollarien lainojen virtaamiselle Unkariin.*

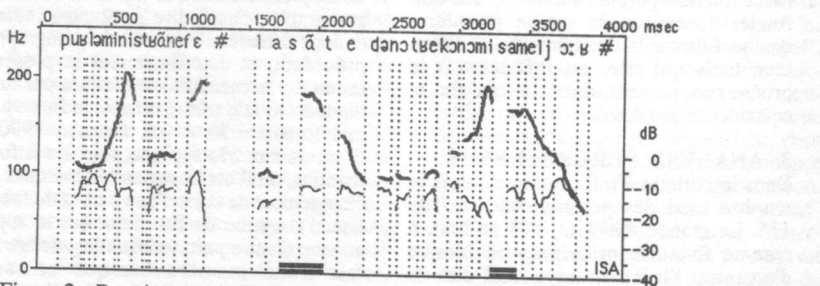


Figure 2a. Premier extrait des informations françaises *Pour le ministre, en effet, la santé de notre économie s'améliore.*

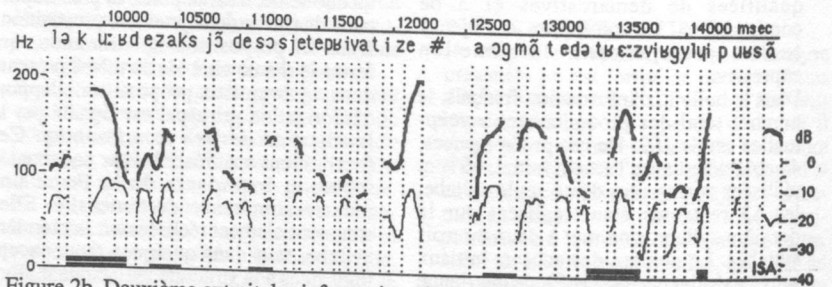


Figure 2b. Deuxième extrait des informations françaises *...le cours des actions des sociétés privatisées a augmenté de treize virgule huit pour cent...*

Outre les focalisations, les énoncés finnois dans ce genre de discours présentent très peu de variation mélodique ou dynamique. La plupart des mots sémantiquement pleins sont pourtant signalés par un accent de mot sur leur première syllabe en utilisant les mêmes paramètres que pour la focalisation, avec toutefois un dosage beaucoup plus modeste. La place de ces montées peut être retardée, principalement pour des raisons rythmiques dues notamment à des mots composés extra-long.

Les figures 2a et 2b présentent deux extraits à l'intérieur d'un sujet des informations françaises. Dans le premier extrait *Pour le ministre, en effet, la santé de notre économie s'améliore*, la première focalisation, *santé*, est réalisée sur le thème et la deuxième, *s'améliore*, sur le rhème de la phrase. Dans les deux cas, il s'agit de l'accent initial renforcé qui se manifeste par une augmentation importante et abrupte de la Fo et de l'intensité, la première également par la durée de tous les segments du mot en question. On pourrait interpréter ce dernier paramètre aussi comme ralentissement du tempo, mais la réalisation reste la même, la fonction aussi; il sert à présenter le thème, où plus exactement à le représenter, car il a déjà été évoqué dans le même paragraphe. C'est également une partie du thème, le mot *cours*, qui porte la première focalisation dans le deuxième extrait. La deuxième focalisation, en deux parties, se fait sur les chiffres *treize* et *huit* qui représentent l'apport principal d'information dans la partie rhématique. Quant aux trois accents initiaux, notés par des barres simples dans la figure 2b, ils ne diffèrent des focalisations que par le degré d'augmentation de la Fo et de l'intensité. Par contre, l'accent initial et la focalisation sont tous les deux réalisés d'une manière très différente comparé à l'accent final. Ces différences ont bien sûr déjà fait l'objet de plusieurs descriptions détaillées sur lesquelles je ne peux pas m'attarder ici (cf. p. ex Rossi 1985 et 1987). Ajoutons que souvent la focalisation peut être renforcée, en français, par une pause qui la précède; elle peut même être isolée du reste de l'énoncé par une pause des deux côtés. La pause fait donc partie des paramètres de la focalisation au même titre que les changements de la Fo ou de l'intensité.

5. CONCLUSION

En finnois, la focalisation se fait en appuyant sur la syllabe normalement accentuée et en utilisant les mêmes paramètres acoustiques que pour l'accent, mais avec un dosage plus fort. En français, la focalisation est, dans le cas de mot de plusieurs syllabes, réalisée le plus souvent sur une syllabe autre que celle qui porte l'accent normal et en utilisant des paramètres acoustiques différents.

Les paramètres utilisés pour la focalisation semblent être les mêmes en finnois et en français. Par contre, l'accent normal, dont le domaine déjà est différent dans les deux langues, est réalisé différemment au niveau des paramètres.

Mon interprétation des facteurs prosodiques jouant un rôle dans la structuration des énoncés français diffère de celui présentée par Rossi (1985 et 1987) sur un certain nombre de points. L'accent énonciatif de Rossi devrait correspondre à ce que j'ai appelé focalisation, c'est à dire la prééminence perceptuelle à fonction informative. L'ictus mélodique de Rossi semble être à peu près identique à mon accent initial sans prééminence perceptuelle. Rossi considère pourtant l'ictus mélodique comme le résultat d'une contrainte physiologique et rythmique (1985: 139). Pour moi, l'interprétation de l'accent initial est basé sur sa fonction linguistique et énonciative qui semble être de signaler l'identité et l'importance des mots sémantiquement pleins, dans les cas où il n'y a pas de raison informative de les focaliser dans le cadre de l'énoncé.

6. RÉFÉRENCES

- [1] IIVONEN, A. (1990), "Style and grammar in the control of text prosody", *Nordic Prosody V*, (éd. par K. Wiik & I. Raimo), 174-182, University of Turku.
- [2] ROSSI, M. (1985), "L'intonation et l'organisation de l'énoncé", *Phonetica*, 42, 135-153.
- [3] ROSSI, M. (1987), "Peut-on prédire l'organisation prosodique du langage spontané?", *Études de linguistique appliquée*, 66, 20-48.
- [4] VIHANTA, V. (1990 à paraître), "L'organisation du message dans les informations radiophoniques finnoises et françaises", *Actes du 4e Colloque franco-finlandais de linguistique contrastive*, Paris 1990.

L'organisation rythmique du discours d'une
personne-bilingue (russe - turc)

L.V. Ignatkina, L.P. Shcherbakova

Leningrad State University, USSR

ABSTRACT

The report deals with the rhythmic structure of a Russian word and the pausation of the speech continuum in the speech of the Turkic languages bearer.

En règle générale, le mot russe est composé d'une suite déterminée de syllabes accentuées et inaccentuées. On reconnaît dans le mot la syllabe accentuée parce qu'elle est prononcée plus nettement avec une augmentation de la durée et de l'intensité, si l'on compare avec une syllabe inaccentuée. La syllabe inaccentuée n'est pas articulée nettement et par conséquent on constate, une réduction des voyelles dans les syllabes inaccentuées.

La place de la syllabe accentuée dans le mot et la présence de voyelles inaccentuées et qui sont réduites en fonction de leur place par rapport à la voyelle accentuée sont des facteurs essentiels, contribuant à l'intégrité du mot russe.

Pour un travail donné, nous entendons, par organisation rythmique du discours russe, d'un côté une détermination correcte de la place de l'accent du mot et une prononciation correspondante à la norme des syllabes accentuées et non

accentuées dans le mot (entre autres correspondant à la norme phonétique, et d'un autre côté, une division objective du flux du discours en groupes sémantico-intonatifs minimaux).

Les descriptions des langues turques qui existent dans la littérature ne font pas apparaître nettement (et de façon identique) la prosodie et y compris les caractéristiques rythmiques du discours par les gens parlant de ces langues là.

Ainsi par exemple, dans la langue kazakh il existe trois points de vue essentiels en ce qui concerne l'accent de mot.

Certains linguistes considèrent que la présence de l'accent est indiscutable et affirment qu'il tombe sur la dernière syllabe. En outre on a observé qu'il y a un accent principal et un accent secondaire, ce dernier déterminant la première syllabe. A.A. Djounisbekov suppose que l'harmonie vocalique est le facteur principal, contribuant à la formation du mot-kazakh. A.M. Cherbak affirme que dans le mot il y a autant d'accents que de syllabes et que l'un d'entre eux est plus fort que les autres mais cela n'a pas une grande importance, ce qui mène à l'absence de fonction sémant-

tico-différentielle de l'accent. Sur le plan articulatoire, les syllabes inaccentuées sont prononcées aussi distinctement que les accentuées.

Dans le discours, en langue kazakh, les mots se regroupent en unités significatives, faisant apparaître des mots qui faisaient une altération, des mots composés, des coordinations syntagmatiques de lexèmes et de mots-outils, de déterminants avec les déterminés, de compléments circonstanciels avec le prédicat et autres. De tels groupes de mots ont un accent unique que l'on appelle accent rythmique. La durée est l'un des composants principaux de l'accent rythmique. A.A. Djounisbekov affirme que le changement de ton n'est pas grave, quoi qu'il change dans une certaine mesure en même temps que la durée. En ce qui concerne la division rythmique le plus important, c'est la pause. Nous notons un point de vue analogue en ce qui concerne l'accent en langue kirghize dans les recherches de A. Orousbaieva "L'accentuation en langue kirghize".

Ainsi, il y a des différences considérables en ce qui concerne l'organisation rythmique du mot et du discours en langue russe et turque.

Une étude a été consacrée à l'examen de la réalisation des syllabes accentuées dans le mot, au rôle de la pause lors de la lecture de textes russes par des gens parlant le turc (le kazakh ou le kirghize).

40 informateurs dont 20 kirghizes et 20 kazakhs natifs de Frounzé et Alma-Ata ont été enregistrés. On a ensuite com-

paré l'enregistrement du discours de référence normatif de l'informateur avec la transcription phonétique idéale de ce même texte. Cette analyse phonétique a été effectuée au L.Ph.E. (à l'université de Léninegrad).

L'analyse acoustique du discours en russe des gens parlant le turc a montré que la place de l'accent dans le mot était le plus souvent déterminé par les informateurs.

Les résultats de l'analyse acoustique montrent que lorsqu'on remarque une accentuation non correcte du mot ceci ne dépend pas du nombre de syllabes, ni de la structure rythmique du mot: dans les mots de 2 syllabes la détermination incorrecte de la place de l'accent est d'environ 50%, autant de cas en ce qui concerne les mots de 3 syllabes et les structures plus compliquées. On avait supposé que la faute d'accent se produirait le plus souvent sur la syllabe finale parce qu'en langues turques l'accent tombe sur la dernière syllabe. Mais il s'est avéré, que l'accent non normatif tombe à peu près toujours avec la même régularité à la fois sur les premières et les dernières syllabes du mot. Dans les structures du type / — — ou — — / il y a une erreur d'accents le plus souvent sur la deuxième syllabe.

Les principales erreurs des kazakh et des kirghizes lors de la prononciation des syllabes non accentuées peuvent être ramenées à cela:

1. Une accélération outre mesure ou un ralentissement de l'articulation des syllabes après l'accent, ce quirompt l'équilibre entre la durée de la voyelle

accentuées et des voyelles inaccentuées dans le mot.

2. L'absence de réduction des voyelles inaccentuées qui sont prononcées aussi clairement que les voyelles accentuées.

3. La diminution ou l'augmentation du nombre de syllabes dans le mot (y compris dans le mot phonétique).

La diminution ou l'augmentation du nombre de syllabes a été également constaté dans le syntagme. Cela se produit lorsque l'informateur a introduit dans le texte ses propres variantes dans le syntagme, un nouveau mot a été ajouté ou les mots présents dans le texte expérimental non pas été prononcés.

L'organisation rythmique du discours des kazakhs et des kirghizes s'est différenciée par des pauses particulières dans le discours.

Comme nous le savons, la pause est le seul moyen qui permet de découper le discours en groupe ou syntagme tant sur le plan de la sémantique que de l'intonation, étant donné que le syntagme minimal peut être égal un mot (ou mot phonétique). Ce qui signifie que dans ce cas la pause est le signal de la fin d'une structure rythmique déterminée.

Le discours russe des kazakhs et des kirghizes à la différence du discours d'un informateur normatif se caractérise par beaucoup plus de pauses. Ainsi dans un seul et même texte les informateurs peuvent faire jusqu'à 293 pauses, alors que l'informateur de référence russe en fera 180, ce qui signifie qu'il y a 1,6 fois plus les pauses dans le discours d'une per-

sonne parlant le turc, que dans celui d'un informateur russe. La différence entre le nombre maximal de pause dans le discours d'un turc et d'un informateur russe de référence est égal à 113. 66% des pauses peuvent être expliquées par le fait que les kirghizes et les kazakhs lisaient un texte "mot-à-mot", ce qui produit une impression de fragmentation et de monotonie.

17% des pauses des informateurs kazakhs et kirghizes qui ne correspondaient pas celle de l'informateur de référence ont été faites là où la norme russe de prononciation ne le permettait pas, par exemple: après une conjonction ou une particule et après les syntagmes nominaux suivants; entre le déterminant et le déterminé; entre le verbe et le mot qui en suit la rec-tion.

Dans le discours des gens parlant le turc, il y a eu des cas de pauses permises par la norme, mais différentes des variantes de l'informateur de référence.

Une des particularités de la fragmentation du discours des kirghizes et des kazakhs est l'abondance des pauses dues à l'hésitation, qui en règle générale témoigne de la difficulté de l'informateur lors de la lecture du texte. Dans le discours de quelques informateurs, on a noté remarqué jusqu'à 20 pauses dues à l'hésitation alors que l'on n'en remarquait pas dans le discours de l'informateur de référence.

Ainsi, les résultats de l'analyse acoustique témoignent du fait que l'organisation rythmique du discours en russe par des gens parlant le turc se distingue considérablement

de l'organisation rythmique du discours d'un russe natif. Principalement beaucoup de ces écarts peuvent être expliqués par les différences de systèmes (de discours) russe et turc dont il est indispensable de tenir compte dans les conditions d'un bilinguisme russe-turc lors de l'étude de la prononciation du russe.

LES MARQUEURS ACOUSTIQUES DE L'ÉNONCÉ EN FRANÇAIS QUÉBÉCOIS

C. Ouellon, C. Paradis et L. Duchesne

CIRAL, Université Laval

ABSTRACT

We intend to identify some acoustic parameters involved as sentence boundary markers in spontaneous speech. The function of the subsequent pause as a marker is emphasized as well as particular aspects of the energy curve and comparison of F_0 values calculated on the vowel nucleus before and after pauses which appear to be important in the perception of sentences, at least in the idiolects considered in this study.

1. INTRODUCTION

Les travaux de recherche que nous poursuivons dans le cadre du projet PROSO visent à mettre en évidence les facteurs d'ordre acoustique qui caractérisent les énoncés du discours oral spontané en français québécois. Nous faisons l'hypothèse que la variation des facteurs acoustiques favorise la variation des taux de perception de la frontière d'un énoncé.

L'analyse des marqueurs acoustiques nécessite, dans la perspective retenue, que le discours soit au préalable découpé en énoncés. Dans une première étape, nous avons donc procédé à la délimitation d'unités théoriques, en retenant une approche inspirée de Nespors et Vogel [3] selon lesquelles le discours s'organise en un ensemble fini d'unités phonologiques organisées hiérarchiquement. De ces unités, le

syntagme intonatif (SI) et l'énoncé (E) nous intéressent tout particulièrement. Le SI est formé par l'assemblage de groupes phonologiques; il est le domaine d'un contour d'intonation et ses limites coïncident avec les positions où peuvent intervenir des pauses. L'unité de niveau supérieur E est constituée d'un ou plusieurs SI et occupe habituellement la longueur de la chaîne dominée par le plus haut noeu de l'arbre syntaxique [3]. D'autres approches de définition du SI ont été considérées [6].

Comme nous voulions dégager les facteurs d'ordre acoustique susceptibles de favoriser la perception des frontières d'unités en discours spontané, nous avons en second lieu soumis le corpus au jugement d'un groupe d'auditeurs, ce qui devrait permettre de dégager des unités concrètes du discours oral spontané et de les mettre en rapport avec les unités abstraites issues de l'application du modèle prosodique de Nespors et Vogel [3].

2. MÉTHODE D'ANALYSE

2.1 Le corpus

Le corpus d'analyse contient dix extraits tirés de deux entrevues réalisées dans la région de Chicoutimi-Jonquière (Québec-Canada) à l'occasion d'une enquête sociolinguistique [5]. Les deux locuteurs (cinq extraits pour chacun) sont de sexe masculin et

appartiennent à la classe moyenne. Chacun des extraits a une durée moyenne de dix secondes et ne comporte aucune interruption de la part d'un tiers. Le corpus a ensuite été numérisé à partir du logiciel d'analyse CSL de Kay Elemetrics.

2.2 Le test d'audition

La bande sonore soumise au test d'audition présentait trois enregistrements de chaque extrait, les deux premiers ayant été filtrés (résidu LPC) de façon à rendre inintelligible le message, le troisième étant la sortie numérisée originale de l'extrait. Vingt-trois juges, des étudiants de niveau universitaire, ont reçu la consigne d'indiquer par marquage électronique le moment qu'ils jugeaient correspondre à un début d'énoncé. Le protocole était inspiré de celui proposé par Lehiste [2] et Kreiman [1]. Au résultat, le corpus a été découpé en 57 énoncés, avec des taux d'accord sur la perception de frontières variant de 4.3% à 91.3% selon l'énoncé. Une procédure de normalisation a été appliquée pour tenir compte des divergences de temps de réaction des divers juges.

3. LE RAPPORT ENTRE ÉNONCÉS PERCUS ET UNITÉS PHONOLOGIQUES

Malgré le nombre relativement restreint de juges (23) et la taille modeste du corpus (10 extraits de 10 secondes chacun), certaines tendances ressortent de l'analyse. En premier lieu, on constate que les débuts d'énoncés perçus (dorénavant EP) coïncident toujours avec des débuts de SI, mais pas nécessairement avec des E (unités phonologiques de niveau supérieur). En second lieu, on peut noter une certaine hiérarchie entre les SI, en lien avec leur statut par rapport à E. Ainsi, le taux moyen de

perception décroît lorsqu'on passe d'un EP qui correspond à un SI constituant à lui seul un E (51.7%), puis à un EP coïncidant avec un SI qui occupe le premier rang d'un groupe de SI formant un même E (43.1%) et enfin à un EP correspondant à un SI de rang n dans le groupe de SI relevant d'un même E (16.4%). Comme les écarts-types sont relativement élevés, il ne peut s'agir là que de tendances qui sont de nature à appuyer l'existence d'une interaction entre données perceptuelles et modèle phonologique.

4. L'ANALYSE DE FACTEURS ACOUSTIQUES

4.1 Procédure

Les extraits numérisés ont été analysés à l'aide du logiciel CSL. Après avoir délimité les unités phonétiques sur les tracés oscillographiques et spectrographiques et sur les listes numériques, nous avons analysé les paramètres suivants:

- la durée des pauses
- la courbe de F_0
- la courbe d'énergie.

Nous avons tenu compte de la variation des valeurs de F_0 et d'énergie à l'intérieur même d'un EP et entre deux EP. Ce choix s'explique par l'objectif que nous nous sommes fixé, ce qui impliquait qu'il faille porter une attention plus grande aux phénomènes qui surviennent aux frontières de cet énoncé.

4.2 La fréquence

On admet communément que les énoncés se caractérisent, en contexte énonciatif, par la décroissance de la valeur de F_0 en finale [8]. Nos données confirment cette observation et concordent avec celles que nous avons présentées au cours d'une recherche antérieure [4].

On calcule une chute moyenne de 2.1 tons entre le Fo maximal de l'énoncé et le Fo du dernier noyau vocalique. Toutefois, dans la perspective où nous nous plaçons, il ne semble pas que ce facteur suffise à marquer la frontière de l'EP; il ne paraît guère exister de lien entre l'importance de la décroissance des valeurs de Fo et le degré de perception de la frontière d'un EP ($r=.085$), dans les corpus filtrés et non filtré.

D'un autre point de vue, la mesure de la pente de Fo donne, comme attendu, une pente négative (-2 tons/sec.) sans qu'il y ait de relation significative entre la variation de pente et la variation du taux de perception ($r=.06$).

Enfin, l'analyse des relations entre frontières de deux énoncés successifs fait voir une hausse de fréquence moyenne de .05 ton, qui confirme l'hypothèse de la réinitialisation de Fo en début d'un EP dans des énoncés consécutifs. On ne note cependant pas de relation entre les variations de Fo, dans ce contexte, et le taux de perception ($r=.02$).

4.3 La courbe d'énergie

Nous avons déjà signalé [4] que la variation de la courbe d'énergie semblait jouer un rôle dans la perception des frontières d'EP, en ce sens que les frontières d'énoncés à haut taux de perception étaient marquées par une chute d'énergie plus forte; cette chute correspond à la chute d'énergie en dB entre la voyelle affichant la plus haute valeur de l'énoncé et la dernière voyelle. Il existe peu d'études sur la variation d'énergie en discours oral, a fortiori en discours oral spontané. Nous avons quand même voulu vérifier la validité des résultats présentés dans Ouellon 90 [4].

Il semble effectivement exister une relation entre les valeurs de chute

d'énergie dans un énoncé et le taux de perception de la frontière de cet énoncé. Pour chacune des deux versions du corpus (filtrée et non filtrée), le calcul de la régression linéaire permet d'obtenir des données comparables, avec $r=-.3061$ et $r=-.3829$ respectivement. La droite résultante fait voir une chute d'énergie qui va de -5.2 dB à -8.8 dB pour des taux de perception qui vont de 20% à 100%. De telles variations d'énergie paraissent significatives, une variation de 1 à 2 dB par rapport au signal naturel étant auditivement détectable [7].

Comme la finale d'énoncé semble marquée par une chute d'énergie, on peut faire l'hypothèse qu'il y aura augmentation d'énergie entre la finale d'un énoncé et le début de l'énoncé subséquent, ou réinitialisation de la courbe d'énergie. Nous avons donc examiné les valeurs d'énergie de part et d'autre de la frontière d'EP. Au résultat, nous pouvons observer, en premier lieu, un écart positif important entre les valeurs d'énergie mesurées sur la voyelle finale de l'EP puis sur la voyelle initiale de l'EP suivant; cette augmentation est de l'ordre de +2dB pour les frontières d'EP à 20% et de +8dB pour celles d'EP à 100%. Donc, à une forte chute d'énergie en finale correspond une forte augmentation d'énergie en initiale d'EP subséquent.

En second lieu, le calcul de la régression linéaire fait voir une relation ($r=.47$) entre le niveau d'augmentation de l'énergie entre deux EP et le taux de perception des frontières.

On peut donc estimer que la variation de la courbe d'énergie dans un EP et entre deux EP favorise la perception des limites d'EP, du moins dans le style de discours et dans les idiolectes du

corpus analysé.

4.4 La pause

L'importance de la fonction de marquage de la pause a été maintes fois signalée, dans Vaissière 1988 [8] entre autres. Dans Ouellon 1990 [4], nous posions l'hypothèse qu'il y avait également une relation entre durée de la pause et taux de perception de l'énoncé. Notre recherche valide cette hypothèse. En effet, la durée des pauses varie entre 122ms et 980ms dans nos exemples et il existe une forte corrélation ($r=.7264$) entre la durée des pauses et le taux de perception des frontières d'EP dans les corpus filtrés et non filtré. La pertinence de la fonction marquage de la pause est confirmée, mais il semble aussi que l'allongement de la durée des pauses favorise un meilleur taux de perception de la frontière d'EP.

5. CONCLUSION

Nous avons fait ressortir l'importance des facteurs que sont la chute de la courbe d'énergie dans un EP, l'augmentation des valeurs d'énergie au passage d'un EP à un autre, de même que la durée des pauses pour le repérage des frontières d'énoncés. Les facteurs liés aux variations de Fo ne paraissent pas, curieusement, jouer un rôle particulier dans la perspective que nous avons privilégiée d'analyser les marqueurs acoustiques de frontières d'EP en regard du taux de perception des énoncés.

Il serait sans doute intéressant de vérifier si nos observations peuvent s'appliquer à d'autres idiolectes français en contexte de discours oral spontané. D'un autre point de vue, nous pensons qu'il faudra solutionner le problème du découpage en unités du discours spontané et approfondir les

connaissances sur l'interaction entre perception et grammaire.

RÉFÉRENCES BIBLIOGRAPHIQUES

- [1] KREIMAN, J. (1982), "Perception of sentences and paragraph boundaries in natural conversation", *Journal of phonetics*, 10.
- [2] LEHISTE, I ET W.WANG (1976), "Perception of sentence boundaries with and without semantic information", *Phonologica*, Dressler-Pfeiffer.
- [3] NESPOR, M. ET I.VOGEL (1986), "Prosodic Phonology", Dordrecht-Hollande, Foris publications.
- [4] OUELLON, C. (1990), "Les marqueurs acoustiques de l'énoncé en discours québécois spontané", *Actes des XVIIIèmes JEP*, Université de Montréal.
- [5] PARADIS, C. (1985), "An acoustic study of variation and change on the vowel system of Chicoutimi and Jonquière (Québec)", thèse de Ph.D inédite, University of Pennsylvania.
- [6] POIRÉ, F., J.M.SOSA, H.PERREAULT et H.J.CEDERGREN (1990), "Le syntagme intonatif en langage spontané; étude préliminaire", *Revue québécoise de linguistique*, 19-2, Université du Québec à Montréal.
- [7] SORIN, C. (1981), "Functions, role and treatment of intensity in speech", *Journal of phonetics*, 9.
- [8] VAISSIERE, J. (1988), "The use of prosodic parameters in automatic speech recognition", *Recent advances in speech understanding and dialog systems*, NATO ASE, F46, Springer Verlag, Berlin.

L'ACCENT EN FRANÇAIS QUÉBÉCOIS SPONTANÉ: PERCEPTION ET PRODUCTION

Denise Deshaies, Conrad Ouellon, Claude Paradis et Sylvie Brisson

Université Laval, Québec, Canada.

ABSTRACT

The present work seeks to examine the relationship between the perception of stressed syllables in spontaneous Québec French with the acoustic parameters of length, intensity and F_0 , as well as the position of the syllable in the metrical structure of the sentence.

1. INTRODUCTION

Notre projet de recherche vise à décrire le système prosodique du français québécois spontané et à dégager des schémas de variations intra- et inter-individuelles dans une perspective variationniste. Avant cependant d'aborder le problème de la variation sociale au plan prosodique, nous avons été amenés à revoir certaines questions de base, dont celle de la place et des corrélats phonétiques de l'accent tonique. En effet, mis à part quelques recherches [1] [7], la majorité des travaux effectués portent sur des corpus obtenus dans des conditions idéales, valant souvent uniquement pour le français de France. En outre, les modèles proposés ont été peu confrontés au discours oral spontané, lequel contraint souvent les chercheurs à revoir certains postulats proposés à partir de données plus contrôlées. Dans ce

cadre, nous avons mis au point une série de tests de perception dont les résultats ont été mis en rapport avec un certain nombre de paramètres. On trouvera dans [4] une présentation de l'influence de certains facteurs linguistiques sur la perception de l'accent. Dans cet exposé, les résultats d'un test de perception de l'accent seront mis en rapport avec les données de l'analyse acoustique des phrases testées.

2. PROCÉDURE

2.1 Test de perception

Le test de perception dont il sera question ici est constitué de 20 énoncés tirés d'une entrevue sociolinguistique faite à Chicoutimi auprès d'un locuteur de 32 ans appartenant à la classe moyenne [3]. Les énoncés sélectionnés de façon quasi aléatoire ont été extraits de l'entrevue et numérisés. Une fois ceux-ci numérisés et segmentés, le test a été mis au point à l'aide du logiciel *MSL AUDIO*: chacun des vingt énoncés était répété 7 fois et chaque répétition était séparée de la suivante par une pause de 2 secondes; la pause entre chacun des 20 énoncés était de 4 secondes. Les sujets recevaient une feuille-réponse sur laquelle apparaissaient les 20 énoncés écrits en alphabet conventionnel et découpés en syllabes à l'aide de barres obliques. Les sujets devaient accomplir deux tâches: on leur de-

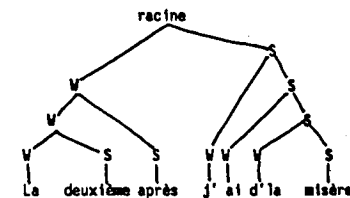
mandait d'abord de souligner les syllabes qui leur semblaient marquées ou mises en relief par rapport aux autres et ensuite d'encercler la syllabe de la phrase qui leur paraissait la plus marquée par rapport à toutes les autres. Ce test a été administré à deux groupes d'étudiants universitaires: le premier groupe était composé de 16 étudiants ayant de bonnes connaissances en phonétique (groupe *averti*), alors que le deuxième groupe comprenait 40 étudiants n'ayant aucune formation en linguistique (groupe *naïf*).

2.2 Analyse linguistique des énoncés

Un certain nombre de facteurs linguistiques susceptibles d'être liés à la perception de l'accent ont été identifiés pour une première analyse sur la base de résultats de recherches antérieures portant sur le français et l'anglais. Sept groupes de facteurs ont été choisis afin d'en vérifier l'effet sur la perception de l'accent, à savoir: 1) la nature du mot, lexical ou grammatical; 2,3,4) la structure syllabique [attaque, noyau, coda]; 5) la position de la syllabe dans le mot; 6) le degré de sonorité du noyau vocalique; et 7) la position de la syllabe dans la structure métrique, facteur que nous décrirons plus en détail, compte tenu de son importance dans les résultats de la première analyse:

- ce groupe de facteurs identifie le degré de proéminence de chaque mot dans la structure métrique de la phrase, laquelle correspond grosso modo à la structure syntaxique de celle-ci. Les relations de proéminence ont été établies sur la base de la règle d'assignation des coups rythmiques, telle que présentée par [2]: cette règle stipule que, pour n'importe quelle catégorie phrastique, l'élément le plus à droite est plus proéminent que l'élément à

gauche. Ainsi, pour l'énoncé 3 du test de perception, nous obtenons la représentation suivante, où S correspond à «strong» et W à «weak»:



La position de chaque syllabe a été codifiée en tenant compte des trois premiers niveaux d'insertion dans l'arbre métrique, à partir du noeud terminal: par exemple, la syllabe *mi* de *misère* a été codifiée SSS, alors que la syllabe *-ième* de *deuxième* a reçu la codification SWW.

3. RÉSULTATS: FACTEURS LINGUISTIQUES ET PERCEPTUELS

Les résultats de cette analyse ont fait ressortir l'importance prépondérante de la position de la syllabe dans la structure métrique, comparativement à tous les autres facteurs. Le programme de régression logistique VarBRul a sélectionné la *structure métrique* comme premier groupe de facteurs lié à la perception de l'accent dans les tests de soulignement et d'encerclement chez les deux groupes de sujets. Les résultats concernant ce groupe de facteurs sont présentés au *Tableau 1*. Les autres groupes de facteurs choisis comme pertinents dans la perception de l'accent variaient quant à leur degré d'importance selon les quatre séries de données et nous n'insisterons pas davantage sur ceux-ci. Mentionnons seulement que, pour la syllabe finale, la probabilité de perception d'un accent allait de .62

Tableau 1: probabilité de perception de l'accent: structure métrique

	A.S.*	A.E.	N.S.	N.E.
Input	.28	.04	.24	.05
degré de signification	.000	.000	.000	.000
STRUCTURE MÉTRIQUE				
rang	1	1	1	1
SSS...	.72	.82	.69	.75
SW...	.60	.73	.58	.64
SSW...	.62	.41	.58	.34
WS...	.35	.18	.37	.26
WW...	.34	.32	.38	.38

* A.S. = groupe averti, soulignement
 A.E. = groupe averti, encerclement
 N.S. = groupe naïf, soulignement
 N.E. = groupe naïf, encerclement

à .68, ce qui démontre le caractère variable de la règle voulant que l'accent tombe obligatoirement sur la syllabe finale [4]. Compte tenu de l'importance de la position de la syllabe dans l'arbre métrique, nous avons inclus ce facteur dans la mise en rapport des données acoustiques et des données perceptuelles en distinguant trois catégories principales: forte (SSS), moyenne (SSW, SW) et faible (WS, WW).

4. PROCÉDURE D'ANALYSE

4.1 Analyse acoustique

Les valeurs de durée, d'intensité et de F_0 , ont été extraites à l'aide du logiciel CSL de Kay Elemetrics. Pour la durée, nous avons tenu compte de la durée totale de la syllabe, alors que les valeurs d'intensité et de F_0 , ont été prises sur la tenue de la voyelle (moyenne d'intensité et de F_0 de la tenue vocalique).

4.2. Analyse statistique

Sur la base des résultats obtenus dans la première série d'analyses statistiques, nous avons d'abord regroupé ensemble les deux caté-

gories d'étudiants et nous avons considéré seulement les données issues du soulignement des syllabes. De plus, en vue de comparer chaque syllabe avec les autres du même énoncé, les mesures acoustiques ont été ramenées en pourcentage de la manière suivante: la durée de la syllabe a été divisée par la durée moyenne des syllabes de l'énoncé et multipliée par cent; pour l'intensité et F_0 , la valeur de la voyelle d'une syllabe a été divisée par la valeur la plus haute de l'énoncé et multipliée par cent. Ces résultats en pourcentage ont pu ainsi être mis en relation avec le pourcentage de perception du trait accent des syllabes à l'aide d'un programme de régression multiple. Dans le modèle de régression multiple, nous avons vérifié l'effet de chacun des facteurs pris isolément ainsi que les combinaisons de tous ces facteurs sur la variable dépendante (perception de l'accent).

5. RÉSULTATS

L'analyse de régression multiple appliquée aux données acoustiques et aux données de perception a retenu le modèle reproduit dans le Tableau 2. Il est d'abord à noter que le très haut seuil atteint par le coefficient R^2 , soit 0.9148, indique que la quasi totalité des données est expliquée par le modèle. Le premier facteur retenu concerne le rôle combiné de la hauteur (F_0), de l'intensité (E) et de la durée (D) dans la perception de l'accentuation, quelle que soit la position de la syllabe dans la structure métrique: plus l'énergie est forte, plus la fréquence est élevée et plus la durée de la syllabe est longue, plus cette syllabe aura de chances d'être perçue comme accentuée. Ce résultat semble donc confirmer le rôle prépondérant, mais non indépendant, de ces trois facteurs

Tableau 2: Effet de la structure métrique, de la durée, de l'intensité et de F_0 , sur la perception de l'accentuation.

Step	Variable	Number	Partial In	Model R**2	R**2	C(p)	F	Prob>F
1	$F_0^*E^*D$	1	0.8025	0.8025	273.6359	861.6074	0.0001	
2	E^*f	2	0.0574	0.8599	134.8443	86.3968	0.0001	
3	E	3	0.0280	0.8879	68.0894	52.4866	0.0001	
4	E^*D^*b	4	0.0203	0.9082	20.3693	46.1088	0.0001	
5	$E^*F_0^*f$	5	0.0046	0.9128	11.1363	10.9111	0.0011	
6	D^*b	6	0.0020	0.9148	8.1343	4.9509	0.0272	

*No other variable met the 0.05 level

acoustiques dans la perception de l'accentuation. Outre ce résultat, le modèle fait ressortir clairement le rôle de l'intensité dans la perception de l'accent; en effet, pour une syllabe en structure forte (E^*f) et pour tous les types de structures (E), soit forte, moyenne et faible, plus l'intensité est forte et plus la syllabe a de chances d'être perçue comme accentuée; l'intensité liée à la durée pour les syllabes en structure faible (E^*D^*b) et l'énergie liée à F_0 pour les syllabes en structure forte ($E^*F_0^*f$) contribuent également à la perception de l'accent. L'énergie est donc le facteur acoustique qui ressort comme le facteur dominant lié à l'accentuation, bien que la durée et F_0 , y jouent également un rôle. De plus, la prise en compte de la position de la syllabe dans la structure métrique s'est avérée pertinente pour mieux comprendre le rôle variable, mais fondamental, des paramètres acoustiques dans la perception de l'accent.

6. DISCUSSION

Si certains auteurs ont pu penser qu'aucun paramètre acoustique n'était pertinent dans la détermination de l'accent [6], les résultats présentés dans cet exposé démontrent leur

importance surtout en ce qui a trait aux syllabes qui n'occupent pas la position «canonique» normalement associée à l'accent tonique. L'oral spontané offre donc un cadre idéal de vérification d'un certain nombre de propositions avancées à partir de données plus contrôlées.

7. RÉFÉRENCES

- [1] Cedergren, H. J., Perreault, H., Poiré, F. & P. Rousseau (1990), «L'accentuation québécoise: une approche tonale», *Revue québécoise de linguistique*, vol.19, no.2, 25-38.
- [2] Liberman, M. & A. Prince (1977), «On Stress and Linguistic Rhythm», *Linguistic Inquiry*, 8, 249-336.
- [3] Paradis, C. (1985), *An Acoustic Study of Variation and Change in the Vowel System of Chicoutimi and Jonquière (Québec)*, Thèse de doctorat, University of Pennsylvania.
- [4] Paradis, C. et D. Deshaies (sous presse), «Rules of Stress Assignment in Québec French: Evidence from Perceptual Data», *Language variation and change*.
- [5] Rossi, M. (1970), «Sur la hiérarchie des paramètres de l'accent», *Proceedings of the Sixth International Congress of Phonetic Sciences*, Prague: Académie des Sciences, 779-786.
- [6] Verluuyten, S.P. (1982), *Recherches sur la prosodie et la métrique du français*, Thèse de doctorat, Universitaire Instelling Antwerpen, Belgique.
- [7] Warren, R. & L. Santerre (1979), «Les paramètres acoustiques de l'accent en français montréalais», in I. Foa & P. Léon (dirs.), *L'accent en français contemporain*, *Studia Phonetica* 15, Ottawa: Didier, 53-63.

MODELLING OF RUSSIAN INTONATION: A "CONTOUR INTERACTION" BASED ALGORITHM

E. Meister

Institute of Cybernetics, Tallinn, Estonia

ABSTRACT

The paper describes the algorithm of modelling of Russian intonation based on the theory of "Contour Interaction". As the basic units the algorithm uses four intonation patterns of syntagma and one pattern of word.

1. INTRODUCTION

There are two opposite theories to characterize the structure of intonation contours: the theory of Tonal Sequence (TS) and the theory of Contour Interaction (CI) /1/. The theory of TS sees intonation contours as being composed of an inventory of abstract tonal elements. According to the theory of CI an intonation contour can be viewed as the composite result of a set of hierarchial patterns. The classical approach to the Russian prosodic system / 2, 3/ is just based on the theory of CI. The basic category of prosodic system in Russian is "syntagma". It is defined as "the phonetic whole expressing one unit of meaning" /3/. The minimal unit of intonation is intonation contour of syntagma which could still be divided into the three functional parts: precentre, centre and postcentre. A special role in a syntagma is played by a

centre because the changes of the pitch in the centre are the most important feature in distinguishing different intonation types. The inventory of basic intonation types according to /2/ includes seven different intonation patterns of syntagma.

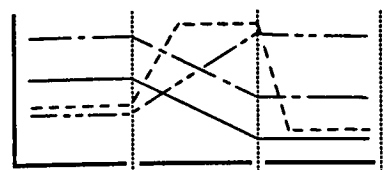
The other essential category of Russian prosodic system is a word stress. In the algorithm introduced in this paper both categories are used as two interacting levels: intonation pattern of syntagma as a higher level and word stress as a lower level. Definitely, the number of interacting levels should be greater (for example the levels of sentence and phoneme should be added) but in the first stage the realization of this simple algorithm has fulfilled two main goals of this work: first, the theoretical goal - to test the validity of intonation patterns of syntagma described in /2,3/ and to find suggestions for further work, second, the practical goal - to get the real programm of intonation modelling for the Russian text-to-speech synthesizer.

2. THE ALGORITHM

2.1. Basic Patterns

In the algorithm four intonation patterns of syntagma

described in /2/ have been realized: the declarative, the interrogative, the non-terminal and the exclamatory patterns.



PRE-CENTRE CENTRE POST-CENTRE

Fig.1. The intonation patterns of a syntagma:

- a declaration,
- an interrogation,
- a nonterminal,
- an exclamation.

To model the word accents one shape of pattern is implemented which is characterized by pitch level before accented phoneme, by rising pitch during the accented phoneme, by falling pitch during the next phoneme and by level during the rest of a word. The peak value of a word contour is approximately 10 % higher of the value of pitch level at the beginning of a word.



Fig.2. The word pattern.

2.2. Input Text

It is assumed that the input text is manually supplied with the marks of word-stress and main-stress because in Russian it is not possible to find correctly

the location of the word-stresses without semantic parsing. The mark of main-stress (") is used once per syntagma and it is located on the most important word of a syntagma (on a centre of a syntagma). All other words are marked with a mark of word-stress ('). In order to distinguish between different intonation patterns the following punctuation marks are used at the end of syntagma: [.] - a declaration, [,] - a nonterminal, [?] - an interrogation, [!] - an exclamation.

These punctuation marks and also conjunctives are used as cues in dividing the input text into syntagmas.

2.3. Contour Generation

The generation of an intonation contour is the third step of the whole algorithm of speech synthesis. It is preceded by the grapheme-to-phoneme transformation and by the speech timing model. The minimal unit processed by the intonation algorithm is a syntagma. The input specification for the intonation algorithm contains the phoneme durations generated by the speech timing model and the string of stress- and punctuation marks. The algorithm works in three steps:

- determination of the intonation pattern of a syntagma,
 - determination of the overall contour of a syntagma according to the durations of the precentre, centre and postcentre,
 - superimposing of word accents into the overall contour.
- The intonation contours are generated within the range from 80 up to 200 Hz.

3. TESTING

In order to estimate the validity of the intonation patterns used in the algorithm two methods were used:

- auditory estimation,
- comparison of the fundamental frequency contours derived from natural and from synthetic speech.

3.1. Auditory estimation

A set of short sentences consisting of one and two syntagmas with four types of intonation patterns were synthesized using the expert system /4/. The listeners were asked to recognize the type of intonation and location of the main stress. In most cases the type of intonation and location of the main stress were distinguished correctly. But in several cases the exclamatory contours were recog-

nized as the declaration. It is due to the similarity of the contour shapes of the declaration and the exclamation.

3.2. Comparison of contours

In order to compare the fundamental frequency contours of natural and synthetic speech the sentence "МАМА МЫЛА МЕНЯ МЫЛОМ" ("The mother washed me with soap") was synthesized and also pronounced by the native speaker as the declaration and as the question with location of the main stress on different words. The fundamental frequency contours were extracted by peak-picker /5/. The visual comparison of natural and synthetic contours exhibits the similarity of the contours (Fig.3, Fig.4).

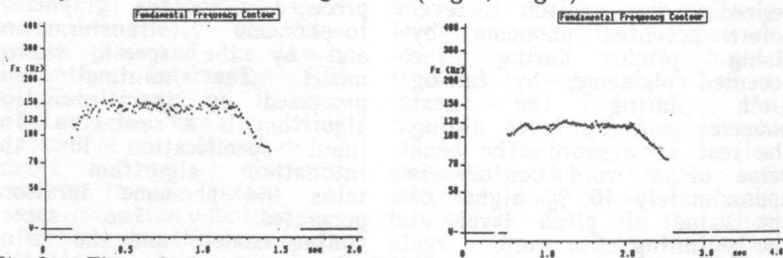


Fig.3. The declarative intonation contours of the sentence: МАМА МЫЛА МЕНЯ МЫЛОМ. left - natural, right - synthetic.

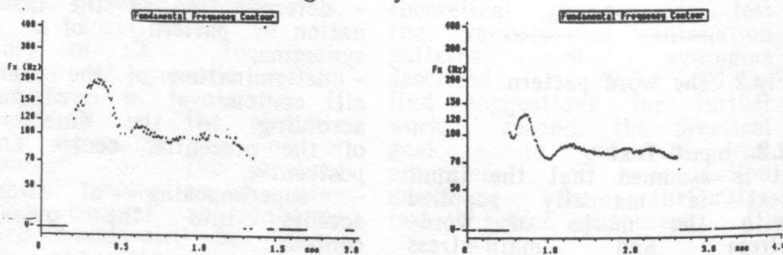


Fig.4. The interrogative intonation contours of the sentence: МАМА МЫЛА МЕНЯ МЫЛОМ? left - natural, right - synthetic.

4. DISCUSSION

The experiments with short sentences showed that the patterns of declarative, nonterminal and interrogative contours described in /2/ are valid for the use in synthesis algorithms. In order to express the exclamation with synthetic voice it is not enough to model the intonation contour correctly. The problems occur with longer syntagmas (more than 5 words) where the intonation contour sounds monotonous. On the one hand this may be caused by the fact that the current algorithm is based on the interaction of two hierarchical levels only: the level of syntagma and the level of word. Introducing into the algorithm the level of sentence and segmental perturbations the quality of synthesized speech will certainly improve. On the other hand the problem of monotonous intonation of long syntagmas can be overcome by manually dividing the long syntagmas into smaller ones (by inserting into the input text additional marks of punctuation and main-stress) although this is not always correct nor theoretically motivated. According to the experiments the optimal length of a syntagma is 3-4 words.

5. SUMMARY

The results obtained in this work on intonation modelling can be formulated as follows:

- some of the intonation patterns described in /2,3/ are valid to practical use in synthesis systems,
- in order to express the exclamation it is not enough to model only the intonation

contour correctly, the other prosodic features should be controlled adequately.

- In order to better modelling of intonation:
- the levels associated with sentence and phonemes should be introduced,
 - the rules dividing the input text into syntagmas with optimal length should be applied,
 - the number of intonation patterns of syntagma should be increased.

REFERENCES

1. K.Silverman (1987) The structure and processing of fundamental frequency contours, Ph.D. thesis, University of Cambridge.
2. Е.А. Брызгунова (1969) Звуки и интонация русской речи. Москва.
3. N.D.Svetozarova (1975) The Inner Structure of Intonation Contours in Russian, Auditory analysis and perception of speech. London, New York, San Francisco, Academic Press, p.499-510.
4. A. Ott, I. Siil (1987) The synthesis-by-rule development system with expert capabilities, Proceedings XIth ICPhS, vol.3, p.278-281.
5. D.M.Howard (1989) Peak-picking fundamental period estimation for hearing prostheses, JASA. 86(3), p.902-910.

INTERACTION BETWEEN PAUSES AND SECONDARY WORD STRESSES IN THE RHYTHMIC SYSTEM OF AN ARCHAIC BULGARIAN DIALECT

P. Vodenicharov

Blagoevgrad University

ABSTRACT

The present paper deals with the interaction between both word and phrase accents and syntactic boundaries in an dialect with very strong enclitic connective tendency. Secondary word stresses and connective phrase stresses in the dialect of Nevrocop are concedered as a rhythmic stresses, results of this interaction. An attemp has made stressing and pausing to be concedered in terms of the theory of metrical grid.

1 INTRODUCTION

In the Selkirk's theory of metrical grid silent syllables have important syntactic function, they delimit words, phrases and other syntactic units in the speech[3] Silent syllables are supposed to coincide with pauses and other delimitative contrasts in the speech. Syntactic units are considered to be able to change the rhythmic units. Final sentence lenghtening ,according to Lehiste, is due to the superimposition of the syntactic structure upon the rhythmic structure of the phrase [2] Our observations on the dialect speech suggest that both word and phrase boundaries often are not well differentiated prosodically because of

existence of very strong enclitic connective tendency in the speech. Pauses have more rhythmic function than syntactic one. Secondary word stresses delimit often rhythmic units then words. As though rhythmic units influence the syntactic units.

The dialect has the following word accent rule. In the basic word forms the accent could not stand on the fourth or further syllables from the end of the word when the final syllable is opened or on the third or further syllables from the end when the final syllable is closed. The basic accent does not shift backwards when some morphemes or clitics subjoin the word but a second or even third stress occur according to a trouchee metrical scheme:

e.g. /'kutʃe/ 'a dog' + /ta/ > /'kutʃ eta/ 'dogs' + /ta/ > /'kutʃ eta/ 'the dogs' + /ni/ > /'kutʃ eta ni/ 'our dogs' + /sa/ > /kutʃ eta ni sa/ 'our dogs are'.

Enclitic joining is often stronger than words boundaries. Even some prepositions and shortened adverbs can subjoin words. In basic word forms accent has both phonological and delimitative function. In derivatives and rhythmic units primary stress has only

phonological function while secondary stress has a delimitative one.

One experiment has been carried out to reveal the role of both primary and secondary stresses in rhythmic organisation of phrase[1]. Experimental data suggest that the dialect has stress-timed rhythm. In terms of the theory of metrical grid [2] the primary stresses coincide with the beats of basic metrical level. Secondary stresses function on an intermediate level which can be supposed to exist between the semi-beats and the basic beats level

1 Experiment 1

In order to reveal the prominence relation between the primary and the secondary stresses we analysed acoustically 25 group of words, each consisting of one basic word and two derivatives, pronounced by two native speakers. Analysis shows that in all cases the words have most prominent primary stress. The primary stressed vowels are longer (87%), on a higher pitch and intensity level (78% and 71% respectively), and have a rising pitch (81%). Secondary stressed vowels are longer than the non-stressed ones (84%) they have rising pitch (72%), and mark second, less prominent, i and f, peak (54% and 69% respectively).

Subjoining of morphemes to the words changes its

prominence patterns in such way:

e.g. /'kutʃe/ + /ta/ > /'kutʃ eta/ + /ta/ > /'kutʃ eta/

The figures refer to the prominence level of the syllables in relation to its f_0 , i and duration values. The prominence contrast between stressed and non-stressed syllables increases subjoining new syllables to the words. In this way increases the rhythmic prominence of the syllables sequences in generated units. Secondary stressing can be considered as a results of the function of enclitic mechanism which integrates rhythmically words in a phrase. This mechanism is often stronger than the word bounding.

The question of rhythmic organisation of syntactic units larger than sentence is very interesting sinse the reality of sentence in spontaneous dialect speech is under discussion. Interesting questions arise. Does integrative mechanism similar to the secondary word stressing functions on a higher speech level? Do the phrase stresses and delimitative contrasts influence the word stresses? We carried out the following experiment to throw any light on these questions.

2. EXPERIMENT 2

2.1. We analysed acoustically one dialect text, a story told by one old illiterate, to reveal the prominence relation between primary and secondary word stresses in connective speach.

The analysis shows that there are some secondary stressed words which do not fall into the considered pattern. With these words the secondary stresses are more prominent than the primary ones; e.g. /'stǎnuvǎf /, /'utjilj'teto/.

The primary stresses are like shifting backwards. Very often the post-stress vowels are elided; e.g. /'kutj'tata/ , /'stǎnuvǎf/. In these cases lowering of f_0 and pitch fall is observed on the primary stressed vowel and lengthening of the secondary stressed vowel. The primary stresses are like deleted by the secondary ones.

We traced the phonetic context of these words. They occur mostly in the middle of the phrases and are regularly followed by pauses. It is interesting that these words have a pitch pattern different from the one typical for mid sentence clause breaks. The pitch falls after peak on the secondary stressed syllable. We have to answer to the question which type of pauses change the prominence of the secondary stresses and why. To reveal the function of the pauses in the considered text we carried out the following test.

2.2.A groupe of 50 native speakers (17-19 years old pupils) were asked after having listened to the text to note the perceived pauses and their length on the transcription forms, listening to it a second

time. This time the text was listened to in pieces for making easier the marking of the pauses. Each piece was recorded together with its proceeding one. We transcribed the text without using capital letters and punctuation marks. After having noted the pauses listeners were asked to put down the punctuation marks. In Bulgarian orthography commas indicate phrase boundaries in sentences. Some of the results we shall discuss.

-The text consisting of 52 predicatives was divided into 5 to 25 sentences. Only 3 full stops are noted by 45 listeners.

-The number of noted pauses is quite great. It varies from 31 to 108. 38 listeners note short pauses which does not exist in reality. The position and the number of these pauses vary. Most of these pauses coincide with commas or full stops.

-The noted punctuation marks are less than the noted pauses. Most of the punctuation marks do not coincide with pauses.

The comparison of the test data with the data of the acoustic analysis shows the following.

-21 of 57 objective pauses, are preceded by pitch typical for mid sentence clause breaks. 17 of them are preceded by secondary stressed words with most prominent second stress. 28 of objective pauses do not coincide with punctuation marks. These pauses will sign as Pr.

-15 of the objective pauses and almost all non-existing in reality pauses are preceded by steep i and f_0 fall, typical for phrase final. They usually follow the f and i peaks of the phrase accent. The final position is typical for it. These pauses will sign as Ps.

The syntactic analysis of the text shows that the phrases are connected usually without conjunction or with the compound conjunction /i/ and. The string of compound connected phrases follow the time sequences of the predicative actions. The syntactic relation between such connected phrases often are complex but they are not manifested lexically. This is may be the reason the listeners to run into difficulties dividing the text into sentences. The other reason is may be that the phrase boundaries are not well differentiated prosodically. 38 of 52 phrases are not limited by objective pauses. Pr pauses interrupt the phrases. They occur periodically, in most cases after the verbs or some adverbs indicating the time sequence of phrase actions. These pauses have more rhythmic function than syntactic one. The periodical alternation phonation: pauses (objective and only subjective) integrate rhythmically the phrases in the text. When Pr pause follow secondary stressed word it increases the prominence of the second stress depressing usually the prominence of the

primary one by causing the elision of post-stressed vowel. In terms of the theory of metrical grid Pr pause can be considered as a group of syllent syllables. It make secondary stress more prominent than some primary ones subjoining secondary stressed word. The rhythmic role of this second stress is change. It probably function not only on a basic beat level with or instead of primary word stress but on a higher metrical level. The final syllable of non-secondary stressed words, preceding Pr pauses, is a start of an upward glide of pitch. In these cases Pr pause can be considered as point of expected pitch change. The words can be supposed to get a silent secondary stress. The secondary stresses preceded Pr pauses have a connective function. They probably function on an intermediate metrical level which can be supposed to exist between the main word stresses and the main phrase stresses levels.

- 1] Karlova, R., Vodenicharov, P., Nic olova, V., Hristoska I., (1990), "Psiholingvisticchno izsledvane na rechevia ritam na edin balgarski govor", *Ezik i literatura*, 1, 12-16.
- [2] Lehiste, I. (1973), "Rhythmic units and syntactic units in production and perception", *IASA*, 54, 5.
- [3] Selkirk, E. (1984), "Phonology and Syntax: The relation between sound and structure", MIT Press.

ANALYSE NUMERIQUE DES TONS DU VIETNAMIEN.

Ngoc QUACH TUAN

Institut de la Communication parlée de Grenoble, France (1982-1986)

Institut National Polytechnique de Hanoi, Vietnam (1987-...)

ABSTRACT:

The ton plays a very important role in the Vietnamese language and manifests under form of 6 accents. So, for example, from the vowel a, there are the pronuciations marked as follow: a, à, á, â, ã, ă. We have realized the digital analysis for theses tons: the analysis by the synthesis and by the inverse filtering. Theses analysis gives the mesures of tons's parameters in the the most exact manner.

1. TONS.

Les faits prosodiques de variation de hauteur en Vietnamien sont utilisés non seulement comme intonation mais aussi comme agent différenciatif de la syllable, ayant la même fonction distinctive qu'une voyelle ou une consonne. Dans cette fonction, on parle de "ton". Les tons jouent un rôle très important dans la langue vietnamienne et se manifestent dans l'écriture sous forme des 6 accents. Ainsi, par exemple, à partir de la voyelle 'a', il y a des pronuciations dérivées marquées comme suit: a, à, á, â, ã, ă.

Dans la réalisation des tons, la valeur absolue de la fréquence F0 varie avec le sexe ou l'âge du locuteur, mais

l'écart, la valeur relative entre plusieurs tons successifs restent les mêmes d'un individu à un autre.

De nombreuses études sur les tons vietnamiens ont été faites par plusieurs auteurs du monde au point de vue phonétique et par des moyens électroniques analogiques; (ainsi ces moyens ne donnent pas les résultats exactes). Parmi eux, c'est le problème de mesure et de mis "en image" des paramètres (les formants, les bandes passants et F0), c'est à dire de rendre visible de 6 tons. L'implantation logicielle sur un système de traitement du signal numérique à l'Institut de la Communication parlée de Grenoble, France [1] nous a permis de développer des mesures beaucoup plus performantes. Pour avoir de "bonnes images" des paramètres des tons, la méthode de "l'analyse par la synthèse" avec le synthétiseur de formants numériques fonctionnant en temps réel a été choisi [1]. Ce synthétiseur a 5 résonateurs connectés en parallèle. Nous pouvons donc éditer leur paramètres pour que les spectres du son naturel et du son synthétique soient assez identiques. D'autre part, la source vocale du synthétiseur a la forme assez proche de l'onde glottique de l'homme [1, 2, 4] donc la forme du signal synthétique est aussi assez proche celle du signal naturel.

Un filtre inverse composant de 4 filtres anti-résonateurs (F1, F2, F3 et F4) avec les fréquences et les bandes passantes correspondant aux paramètres du synthétiseur ont été appliqué au signal naturel. F5 (en haute fréquence) n'est pas nécessaire. A la sortie, nous pouvons obtenir le signal glottique dérivé qui est caractérisé par un retour à zero de manière rapide (figure 2). Ce retour brusque joue un rôle important non seulement dans la source et dans la perception mais aussi dans la

de la fréquence fondamentale F0 d'une manière très exacte. Rappelons que la mesure des paramètres du signal glottique (naturel ou non dérivé) devient difficile car le signal est proche de zéro aux moments significatifs acoustiques.

La figure 3 (voir les 2 derniers pages) se compose de signaux (en haut), des formants et des bandes passantes (au milieu), et de F0 (en bas) en fonction du temps, du résultat d'analyse de 6 tons pour la voyelle 'i', sur l'écran de l'ordinateur. On fait les

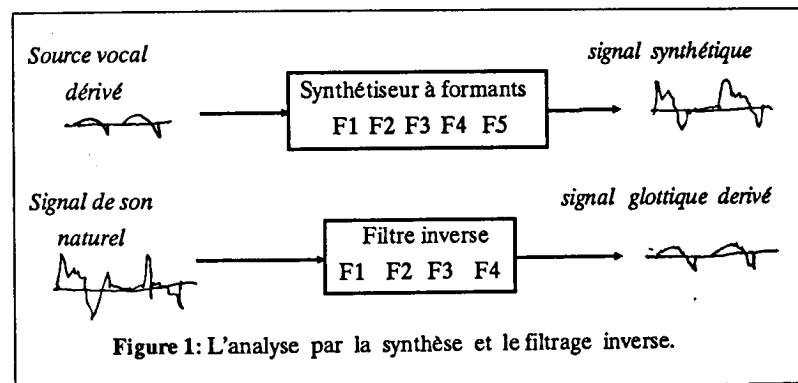


Figure 1: L'analyse par la synthèse et le filtrage inverse.

mesure

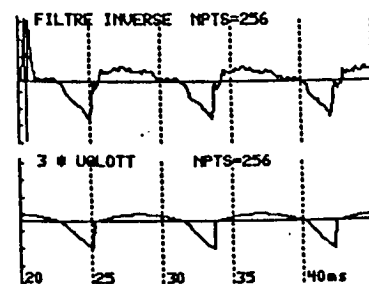


Figure 2: Le signal glottique dérivé du son naturel (en haut) et celui synthétique (en bas).

remarques suivantes:

+ les formants et les bandes passantes des tons sont assez identiques, cela veut dire aussi qu'ils ne decident pas de l'effet des tons.

+ le changement (relatif) de F0 est tout à fait différent de tons. On peut dire que F0 joue le rôle décisif pour les tons.

+ l'allure de F0:

- le ton 0 ou le ton neutre (i):
F0 change très peu.
- le ton bas-descendant (i):
F0 decend
- le ton haut montant (i):
F0 monte

- le ton descendant - montant (i):
F0 descendant - montant
- le ton (i):
F0 de ce ton est le plus complexe.
- le ton bas glottal (i):
au début, F0 descend légèrement et ensuite, elle tombe rapidement.

2. TESTS PERCEPTIFS.

En fonctionnement en temps réel, nous avons réalisé des tests perceptifs pour les tons.

Test 1:

Un signal de ton neutre a été synthétisé avec F0=constant. Ensuite nous l'écoutons et trouvons qu'il est tout à fait marqué 'synthétique'. Cela veut dire que pour le ton neutre (ton 0), malgré des changements très légers (environ 4-5 Hz), le changement de F0 joue un rôle important pour garder l'effet naturel.

Test 2:

Nous trouvons que le ton i a deux périodes de F0: F0 descendant et F0 montant. Une question se pose: Est-ce qu'on obtient le ton i en juxtaposant le signal de ton i et le signal de ton i. La réponse est positive. Cela montre l'effet de masquage entre les tons: il faut avoir un temps nécessaire pour séparer deux tons. Si- non, avec deux tons différents, on obtient le troisième ton.

Test 3:

En décalant la fréquence fondamentale F0 d'un constant pour chaque ton, nous écoutons le même ton mais la perception du sexe se

change. Nous trouvons que dans la réalisation des tons, la valeur absolue de la fréquence F0 varie comme le sexe ou l'âge du locuteur et c'est l'écart, la valeur relative qui décide le ton.

3. BIBLIOGRAPHIE

- (1) QUACH TUAN Ngoc (1986)
Réalisation d'un synthétiseur à formants numérique en temps réel. Caractérisation de la source d'excitation et des transitions d'amplitude.
Thèse de Docteur,
à l'I.N.P de Grenoble, France.
- (2) QUACH TUAN N. et B. GUERIN (1986)
Voice Excitation Sources for Digital Formant Synthesizers.
Bulletin du laboratoire de la Communication parlée.
Vol 1A. pp 67-89.
- (3) LINDQVIST J. (1970)
The voice source studied by means of inverse filtering.
STL-QPSR 1/1970, pp 3-9.
- (4) FANT. G (1979)
Vocal source analysis - A progress report.
STL-QPSR 3-4/1979, pp 31-53.
- (5) DOAN THIEN T. (1980)
La phonétique du vietnamien.
Edition Dai hoc. Hanoi - Vietnam.
- (6) QUACH TUAN N. (1988)
Les tons du Vietnamien.
Dai hoc Bách khoa Hà nội.

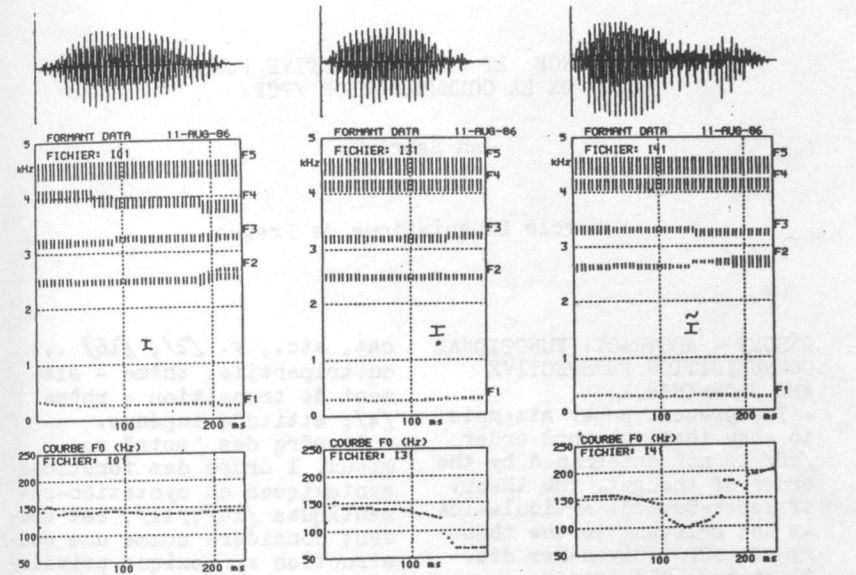


Figure 3: L'image des signaux, des formants, des bandes-passantes et de F0 pour les 6 tons de la voyelle i: i, i, i.

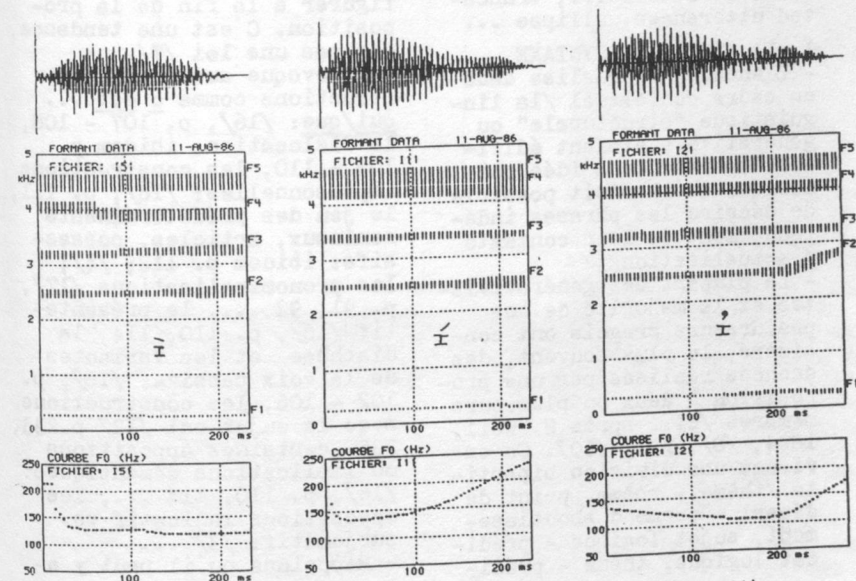


Figure 3 (suite): L'image des signaux, des formants, des bandes-passantes et de F0 pour les 6 tons de la voyelle i: i, i, i.

Jan Šabrůla

Cercle Linguistique de Prague

RÉSUMÉ - ABSTRACT: FUNCTIONAL
COMMUNICATION PERSPECTIVE
AND PROEMINENCE

- The present paper attempts to show that the word order /WO/ is not determined by the order of thought. The theory of Topic-Comment Articulation is not relevant to the theory of FCP /Given-New distinction/. The syntax should not be excluded from examination, but first and foremost attention should be paid to PROEMINENCE, to various types of "stress", emphasis, rhythm, pause ..., characteristic tonal structure ..., truncated utterances, ellipse ...

1. LA PFC ET LA SYNTAXE

- L'énoncé se réalise dans un cadre contextuel /la linguistique "structurale" ou générative s'étaient édifiées à partir de l'idée chimérique qu'il était possible de décrire les phrases indépendamment de leur contexte d'actualisation/.

- La plupart des générativistes et la majorité de nos précurseurs pragois ont considéré, le plus souvent, des énoncés réalisés par une proposition à deux ou plusieurs membres /cf., après H. Weil, 1844, /8/, /9/, /10/. On envisage une division bipartite /thème - rhème, point de départ - terme d'aboutissement, sujet logique - prédicat logique, thème - prédi-

cat, etc., v. /2/, /16/ ... ou tripartite; thème - élément de transition - rhème /4/; attitude ingénue.

- L'ordre des "mots" ou, mieux, l'ordre des fonctions syntaxiques ou syntaxico-sémantiques /16/, /11/, est souvent considéré comme une construction syntaxique privilégiée susceptible de signaler le thème et le rhème de l'énoncé. Ainsi en fr., p. ex., les mots représentant le sujet grammatical, placés en tête de la proposition, expriment souvent le thème, alors que le rhème devrait figurer à la fin de la proposition. C'est une tendance, non pas une loi /7/.

- On évoque aussi les constructions comme c'est ... qui/que: /16/, p. 107 - 108, la dislocation: ibidem p. 108 - 110, les constructions impersonnelles: /16/, p. 111, le jeu des prédéterminants nominaux, articles, possessifs: ibidem p. 112; /4/; les pronominalisations /17/, p. 91, 92 ..., le présentatif /16/, p. 110, 114, la diathèse et les variantes de la voix passive: /16/, p. 102 - 106, les constructions avec le sujet on: /12/ p. 233, 235, certaines oppositions ou implications sémantiques: /16/, p. 110, 114 ..., les oppositions indicatif vs. subjonctif: /3/ ...

- Rappelons qu'il peut y a-

voir des épisémissions entièrement rhématiques /qui répondent à la question "Que se passe-t-il?": /16/, p. 102./

2. LA PROÉMINENCE

2.1. Il est assez étonnant de voir combien les auteurs traitant de la PFC négligent le facteur suprasegmental.

- En réalité, dans l'ordre de l'oral /pourtant cher aux Pragois/, le rhème est marqué par un indice privilégié, la proéminence.

- La PROÉMINENCE /terme de D. Jones, d'H. Rigault .../ est un phénomène complexe: intensité, accent d'insistance, courbe mélodique, modifications articulatoires /quantité et qualité des sons .../ - La proéminence porte en principe sur une unité SIGNIFICATIVE /v. quand même, infra, 2.4/.

- Elle est toujours présente, alors que l'ordre des "mots", la position "gauche" /du thème/ ou "droite" /du rhème: cette terminologie, très en vogue chez les distributionnalistes et, par tant, chez de nombreux auteurs pragois ou français contemporains et autres, n'est pas adéquate pour l'ordre de l'oral où il s'agit d'une succession, présentation de la matière sonore dans le temps, et ne convient pas pour les considérations d'ordre général, étant justifiable pour certains types d'écriture seulement/ - peut TOUJOURS être revalorisé par un indice privilégié, la PROÉMINENCE, toujours pertinente. Dans cette fonction, l'ordre des "mots" peut donc être redondant.

- Les auteurs qui voudraient contredire à l'ordre des mots le privilège absolu, comme facteur quasi unique de la PFC, en viennent jusqu'à affirmer que le fr. ou l'anglais ne soient pas sensi-

bles aux exigences de la PFC /5/.

- Les slavisants exagèrent souvent le caractère "libre" de l'ordre des mots en tchèque. Après H. Paul, F. Trávníček avait pourtant montré que l'ordre des mots en tchèque n'est pas "libre" sans limitations /18/.

2.2. Proéminence et texte versifié

- Dans un texte versifié, l'accent, en contradiction avec les règles de la métrique régulière, peut frapper la syllabe initiale ou plusieurs syllabes successives en fr., si ces syllabes sont le support physique de la proéminence, réclamée par la PFC: VA, SERS, et me laisse en repos /Rac./, v. /15/. - Un seul pied est quelquefois réservé à l'intérieur d'un vers pour les éléments mis en relief, en particulier, pour le rhème: C est ainsi que peut être isolé, p. ex., un substantif en fonction d'apostrophe ou une apposition: v. /15/, p. 20. La proéminence se manifeste ici souvent aussi par la cadence relativement plus lente, avec laquelle on lit la séquence mise en relief dans un vers court.

- Dans un texte versifié, le thème peut être détaché du rhème par un enjambement, donc par une PAUSE. L'opposition entre la syntaxe et la métrique est tranchée ici au profit de la métrique et de la PFC par la pause /14/ 2.3 Le cas de la négation - En français familier, on voit s'affaiblir et même disparaître le segment antéposé et non accentué du signifiant /dénotant/ discontinu de la négation /par lequel la négation s'exprimait en latin: non > ne > ø/. C'est la constitution des groupes rythmiques dans lesquels la dernière syllabe de

chaque groupe est forte, qui explique en partie d'abord la double négation, puis la réduction du signifiant antéposé de la négation: mais, avant tout, c'est le caractère rhématique de son second élément, qui a permis un écrasement phonétique de la partie redondante du signifiant discontinue de la négation dans le langage populaire et familier. V. /13/. 2.4. Le cas du zéro linguistique

- Dans certains cas, le zéro entre dans la relation paradigmatique. Dans une réponse affirmative à la question Tu /ne/ manges /pas/? le rhème est "affirmation". Il suffit de dire Oui ou Si. Dans la réplique possible Je mange, le noème "action de manger" est redondant, il est thématique, et c'est uniquement "affirmation", "exprimée" par le désignant zéro /par opp. à ne - pas/ qui est rhématique. Le verbe Je mange, sert ici de support à l'affirmation impliquée. Le rhème de la réplique est désigné par un zéro dans la structure sonore! Manger a, dans ce type de réplique, une fonction communicative impropre, Je mange est donc dans le cas donné le "substitut" de l'"affirmation", la prééminence est portée par le support physique du substitut du rhème!, dans la mesure où ce substitut permet la réalisation de l'accent.

2.5. Ellipse, phrasillon et la PCF

- Dans le cas limite, la prééminence contribue à conserver dans la structure superficielle l'essentiel /le désignant du rhème/, les éléments facultatifs pouvant être supprimés, à savoir

2.5.1 l'élément rhématique

non lié contextuellement, p. ex. un énoncé exclamatif, une interjection..., constitue à lui seul l'épémion /élément non propositionnel, phrasillon marqué par la prééminence/ - ou, 2.5.2 quand l'élément rhématique est lié contextuellement dans un discours, texte alterné produit collectivement: /6/, p. 197, p. ex. dans une réplique. V. /17/, p. 56, 57.

3. CONCLUSION

3.1. Notre enquête dément donc l'opinion répandue, simpliste et erronée, basée sur le principe de la syntaxe linéaire, sur le rôle primordial de l'agencement syntaxique /ordre des "mots", etc., permutations linéaires des éléments/ dans la signalisation de la PFC.

3.2. L'auteur de ces lignes a été très sensible au propos de M. Peter Blumenthal: /1/, p. 3, selon lequel à Šabršula "revient le mérite d'avoir le premier appliqué les principes des linguistes pragois à la syntaxe du français". Mais une petite rectification s'impose: l'auteur de cette contribution réclame le mérite de s'être opposé à l'application dogmatique et rigide du principe de linéarité, à l'explication de la PFC par des orgies syntaxiques, auxquelles se livrent certains générativistes /leur initiative, dans ce domaine, est assez tardive/ et, dans le cadre d'un structuralisme de l'"expression", très linéaire, certains auteurs de Prague et d'ailleurs. - La phrase n'est pas le niveau privilégié de l'analyse. La fonction l'emporte sur la forme. Les moyens PROSODIQUES désambiguisent les fonctions communicatives des segments linéaires /et du zéro linguistique/.

RÉFÉRENCES

- /1/ BLUMENTHAL, P. /1980/, La syntaxe du message, Tübingen, Niemeyer
- /2/ BONNARD, H. /1976/, "La prédication", G.L.L.F. T. V, Paris, 4556 - 4560
- /3/ BORGESON, L. /1966/, "La fréquence du subjonctif ...", Studia neophilologica, 38
- /4/ FIRBAS, J. /1957/, "On the concept of communicative dynamism ...", Brno, Studies in English 7
- /5/ HOREJŠÍ, V. /1974/, "A propos de la perspective fonctionnelle de la phrase fr.", Olomouc: Acta Univ. Palackianae, Philol.
- /6/ KERBRAT-ORECCHIONI, C. /1990/, Les interactions verbales, T. I, Paris: Colin
- /7/ LEWINSKY, B. /1949/, L'ordre des mots dans Bérimus, Göteborg: Rundquist
- /8/ MATHESIUS, W. /1939/, "O tzv. aktuálním členění větném", Praha: SaS
- /9/ PETR, J. /1987/, et col., Mluvnice češtiny 3, Skladba, Praha: Akademia
- /10/ SGALL, P. et HAJIČOVÁ, E. /1971/, "A Remark on Chomsky's Focus", PBML 14, 3 - 11
- /11/ SRPOVÁ, M. /1990/, "Quelques remarques sur la dynamique énonciative en Tchéque et en Français", Paris: R.E.S., 403 - 416
- /12/ - /17/, Šabršula, J.: /1968/, Slovní druhy současně francouzštiny, Praha: SPN /12/ /1970/, "La description phonétique de la négation ...", Prague: Proceedings of the Sixth Intern. Congr. of Phonetic Sciences 1967, Academia
- /1971/, "La perspective fonctionnelle de l'énoncé dans les vers de Peire Vidal", Montpellier: Actes du VI^e Congr. Intern. de L. et Lit. d'Oc, C.E.O /14/ /1972/, "Intonation, pause et syntaxe dans le langage poétique versifié", Praha: AUC - Phonetica Pragensia /15/ /1973/, "La perspective fonctionnelle de l'énoncé", Praha: AUC-RP VIII, 93 - 124 /16/ /1980/, Substitution, représentation, diaphore, Praha: AUC /17/
- /18/ TRAVNÍČEK, F. /1937/, "Základy československého slovosledu", Praha, SaS, 78 - 86

SYMMETRY AND ASYMMETRY IN MULTI-DIMENSIONAL PROSODIC SYSTEM AS CUES OF TEXTUAL EXPRESSIVENESS

E.A.Nushikyan

Odessa State Univeristy, Odessa, USSR

ABSTRACT

The paper presents the results of experimental phonetic research carried out with the purpose to examine symmetry and asymmetry in the multi-dimensional prosodic structure of expressive texts.

Symmetry is understood here as the general feature of material world reflecting the symmetrical arrangement of the parts of the structure, balanced proportions, correspondence in size, shape, and relative position.

Asymmetry is the opposite of symmetry. However, it is important to understand that the one cannot exist without the other. This is so because in every existing object, during its growth the balance of its parts is violated.

The present paper discusses some new aspects of intonational theory: symmetric and asymmetric features of multi-dimensional prosodic system are explored.

2. SPEECH MATERIAL AND SUBJECTS

Textual prosody was studied here on the statistic data obtained from phonetic experimental investigation of more than 100 expressive and corresponding neutral texts recorded by 20 subjects who were native speakers of English, Russian and Ukrainian. These texts expressed the fourteen most frequently observed positive and negative emotions: joy, sorrow, anger, fear, despair, threat, surprise,

shame, offence, contempt, suspicion, irony, approval, rebuke. The original speech signal was instrumentally analyzed with the help of the Visi-Pitch and IBM speech program, Sena-Graph of the Kay Elemetrics Corporation. For evaluating the average data standard methods of mathematical statistics were applied (t ratio, Student's t and correlation coefficients; calculations were done with the help of IBM program "Lotus"

3. DATA ANALYSIS AND RESULTS

This report generalizes from the results of a long-term investigation carried on by the author. Our previous investigation /1/ proved that information about emotions comes over multiple channels: by lexical cues, grammatical structures and prosodic indicators. These levels of linguistic analysis are closely interconnected. Expressive speech prosody is described as multi-dimensional system characterized in terms of symmetry and asymmetry of its variable components: fundamental frequency, intensity, duration and spectral composition.

The statistical analysis of these main acoustic characteristics shows that a greater symmetry is observed within a temporal framework of the given texts. An act of speech is regularly time-oriented. Speech arrangement in time is related to the specifically regulated nature of acoustic signals. The regular symmetric feature of the temporal structure of expressive texts can be observed in the equality of the mean

syllabic duration of opening and final phrases (see table 1).

Table 1

The mean syllabic duration of opening and final phrases in expressive texts in English

Emotions expressed in the texts	Mean syllabic duration of opening phrases (ms)	Mean syllabic duration of final phrases (ms)
joy	230	225
sorrow	240	225
anger	159	151
fear	221	228
despair	203	193
threat	273	264
surprise	230	225
shame	198	188
offence	185	200
contempt	221	239
suspicion	230	240
irony	254	242
approval	236	241
rebuke	246	256

This regularity is broken in highly emotional texts. For example, the mean syllabic duration of an opening phrase, expressing high degree of despair is 357 ms while in all the other phrases it varies from 150 to 240 ms.

Symmetry of the temporal structure of the text can also be found in the proportion between total text duration and pauses. Table 2 presents the volume of pauses in % in the texts expressing the above-mentioned 14 emotions and the corresponding neutral ones.

Table 2
The volume of pauses (%) in expressive and corresponding neutral texts

Emotions expressed in the texts	Volume of pauses (%)	
	emotion- texts	neutral texts
joy	42	40
sorrow	42	40
anger	48	33
fear	45	36
despair	46	41
threat	41	27
surprise	41	39
shame	38	14
offence	49	41
contempt	44	14
suspicion	47	28
irony	45	30
approval	41	32
rebuke	43	33

The figures in the table show that the regular symmetric feature of expressive texts can be also observed in the equality of the volume of pauses. The corresponding neutral texts do not reveal such symmetry.

The obtained data suggest that there is a principle of symmetric compensation in speech prosody: when the degree of symmetry decreases on one structural level it increases on another. Spectrographic measurements of formant frequencies support this principle. A shift of F_2 , F_3 , and F_4 into higher regions along with the more complicated structure of their harmonics, a constant increase of the total formant energy of the nuclear vowel occurs at the expense of the decrease of formant energy of unstressed syllables. Spectre-

grams of neutral texts revealed more symmetric regularities: well-defined formant structure during the vowels was observed.

The quantitative analysis of the intensity of expressive text prosody demonstrates this principle too: a decrease of energy in one section of the text is accompanied by an increase in another. These changes of energy in expressive texts are closely connected with the changes in the degree of emotional tension. A gradual increase of the total energy to the end is observed in the texts, expressing active emotions, i.e. anger, threat, irony, suspicion, rebuke. For example, in the emotional text expressing all shades of anger - from irritation to rage - the relative intensity of the utterances is: 1,24; 1,39; 1,83; 2,06; 2,41. The decrease of total energy occurs in the texts expressing passive emotions, i.e. sorrow, offence, shame. It appears probable that the asymmetric distribution of energy in expressive texts have to be often specified for certain changes of emotions in them. In contrast, the unexpressive texts are characterized by symmetric distribution of energy.

The symmetry of the melodic structure of expressive texts is found in the similarity of its shapes. However, in highly emotional texts numerous asymmetrically arranged variants are observed. This is due to the dynamic changes of emotional tension, which in turn lead to changes in pitch movement.

CONCLUSIONS

The textual level of analysis has revealed the multi-dimensional nature of the symmetric prosodic space in which the compensatory distribution of prosodic features is taking place. However, the obtained results seem to demonstrate that symmetry of the multi-dimensional prosodic system is no more than an ideal form of the actual asymmetric acoustic features.

REFERENCES

- 1 Nushikyan E. (1987). The typological analysis of emotional speech prosody. - Proceedings XIth ICPnS, vol. 3. - P. 210-213.
- 2 Shafnarovsky I. (1985). Simmetriya v prirode. - Leningrad: Nedra.
- 3 Shubnikov A., Koptsik V. (1972). Simmetriya v nauke i iskusstve. - M.: Nauka.
- 4 Wigner E. (1970). Symmetries and reflections. - Indiana University Press, Bloomington.

ACCEPTABILITY OF SEVERAL SPEECH PAUSING STRATEGIES IN LOW
QUALITY SPEECH SYNTHESIS; INTERACTION WITH INTELLIGIBILITY

Vincent J. van Heuven & Peter J. Scharpf

Dept. Linguistics/Phonetics Laboratory,
Leyden University, The Netherlands

ABSTRACT

We know from previous studies that inserting speech pauses at the end of coherent word groups, but not in any other positions, improves the intelligibility of low quality speech. The present study examines the effect of several pausing strategies on the acceptability, rather than intelligibility, of low quality speech for listeners who either did or did not know the verbal contents of the message beforehand.

1. INTRODUCTION

Our research started from the assumption that speech pauses may help the listener to decode the incoming message. In earlier reports [5,6] we studied the effects on word recognition in connected speech of four different speech pausing strategies applied to low quality diphone synthesis of Dutch sentences. Melodically and temporally well-formed pauses had been inserted in long sentences at more or less regular intervals. Subjects listened to each sentence twice (in order to reduce memory load), and were asked to fill in all the content words they had recognised on their answer sheets. In the answer sheets all the function words had been printed beforehand, interspersed with underlined blanks, one for each content word to be filled in. About 50% of the blanks were filled in correctly after listening to a sentence once, another 30% was added after hearing the

sentence for the second time. At first sight, the effect of synthesizing pauses in the utterances facilitated word recognition only marginally: when taking the condition without any pauses at all as a base line, percent correctly filled in blanks was raised significantly, but no more than by 4 percent on average, as a result of inserting pauses at prosodically motivated boundaries (I and Phi-domain boundaries, cf. [1,3]). However, when pauses had been inserted before important content words (mimicking a speech pausing strategy of certain experienced broadcasters), no significant improvement was obtained relative our baseline condition. When speech pauses had been inserted by a fixed rule after every sixth word, irrespective of grammatical structure or of the communicative importance of the word, the subjects' performance was significantly poorer than in the baseline condition.

The differences between the conditions were larger (8 percent improvement re. baseline), however, when we considered the effects for monosyllabic words only. The results revealed that the recognition of monosyllabic words, but not of longer words, was facilitated by the insertion of grammatically motivated pauses (3 to 4 percent improvement re. baseline).

This interaction between pausing and word length is predictable from what we know about word recognition in connected speech (cf. [4]). Longer words are

usually recognized before the final sounds belonging to the word's acoustic make-up have reached the listener. For instance, the word elephant can be recognized when only the sound sequence corresponding to eleph has been heard: there is no other word in the English lexicon than elephant (and its derivations) that begins with this sound sequence. The final portion of longer words is lexically redundant. This means that, in connected speech, the listener can predict exactly where a new word will start, thus reducing the number of competing recognition hypotheses. Short, monosyllabic words, however, cannot be recognized with certainty until at least some of the following context has been heard: monosyllables can very often be the first syllable of a longer word (cf. cap - captain; cat - caterpillar, etc.). Therefore the number of competing parses for a sequence of monosyllabic words is typically greater than for sequences of polysyllabic words, with better word recognition performance for the latter type [5]. The difference between monosyllabic and polysyllabic words will increase when the average number of competing parses is raised, as happens in poor quality speech. In such cases, word segmentation ambiguity can be reduced by inserting speech pauses, which always occur at word boundaries. Moreover, if the pauses are inserted at grammatically motivated positions, they contain not only information on word boundary location, but also reveal part of the grammatical structure of the input sentence.

In the present experiment we wished to study the influence of the various pausing strategies on acceptability, rather than on word recognition. Conceivably, frequent interruption of the utterance by conspicuous and time consuming pauses - even if conducive to better intelligibility - may be disruptive and annoying to the

listener.

Furthermore, we reasoned that one may expect different acceptability results when the listener knows the text beforehand, than when the text is new to him. If in the latter case the pauses do indeed help the listener to resolve ambiguous word boundaries and recognise the grammatical structure of the sentence, he will gladly pay the price of having to put up with the time delay. However, when the listener is familiar with the message, pauses are not needed, and will sooner be felt as a nuisance. We therefore predict that frequent pauses, especially when they do not contribute to word recognition, will be negatively valued by the listener. However, in novel utterances, which are difficult to understand, pauses that increase intelligibility will be positively valued.

2. METHOD

Seven Dutch sentences, each 36 words and 68 syllables long, were selected from the stimulus material used in the earlier intelligibility test [6]. These sentences had been concatenated from severely quantized diphones with a resulting speech quality that was equal to that of the Philips MEA8000 formant synthesis chip, and were given appropriate intonation contours. Pauses were 200 ms long, marked by a pitch fall B (cf. [2]), and preceded by a 40% lengthened syllable. This means that sentences with pauses lasted longer (by some 250 ms for each pause) than sentences without any pauses. We took the precaution of creating an extra stimulus condition without pauses with a slower speaking rate so that the overall duration here was equal to that of a sentence with pauses inserted. Suspecting that a melodically marked boundary could be counterproductive in the middle of a coherent phrase, we added a sixth condition in which speech pauses before important content words (as

in condition 4 below) were not accompanied by the boundary marking pitch movement. The six different boundary marking conditions are listed below:

1. No pauses, no adaptation of speaking rate.
2. As condition 1, but with speaking rate slowed down so as to make the duration of the utterance equal to that of versions with pauses.
3. Six pauses inserted at fixed intervals (after every sixth word), disregarding any structural considerations.
4. Six pauses inserted at more or less regular intervals, but always immediately preceding important content words; these pauses did never occur at the end of an intonation domain (I) or of a phonological phrase (Phi).
5. As condition 4, but with pauses marked temporally only (no boundary marking pitch movements were executed).
6. Six pauses interspersed more or less regularly, but always at the end of an I or Phi domain.

The full set of 7 (lexically different sentences) * 6 (pausing conditions) = 42 stimuli, preceded by 6 practice stimuli, were presented to two groups of 60 listeners. The first group had taken part in the intelligibility test described in section 1, immediately prior the present test. Each of these listeners had heard the sentences twice before; also, they had printed versions of the stimuli before them. The stimulus material should therefore be perfectly intelligible to this group of prepared listeners.

The material presented to a second group of 60 unprepared listeners who had never heard the sentences before. In this group each listener heard each of the 7 lexically different sentences only once, with maximal variation of pausing conditions within subjects.

All listeners heard the stimuli over headphones, and rated each

sentence along a 7-point acceptability scale, where 1 stood for 'very unpleasant to listen to' and 7 for 'very pleasant to listen to'.

3. RESULTS

The results are presented in Table I.

Table I: mean acceptability score broken down by type of listener (prepared vs. unprepared) and pausing condition (1 through 6, see text); in parentheses the number of reponses.

PAUSING CONDITION	LISTENER TYPE	
	prep.	unpr.
1. no pauses	4.7 (120)	4.4 (10)
2. as 1, but slowed down	4.6 (120)	4.2 (10)
3. pauses after every 6th word	2.7 (120)	3.5 (10)
4. pauses before imp.cont.words	2.6 (120)	3.3 (10)
5. as 4, but no pitch movement	3.1 (120)	3.7 (10)
6. pauses at word group boundary	4.1 (120)	4.3 (10)

When the listener knows the text beforehand (prepared), the condition with no pauses at all, no matter whether speaking rate is slowed down (cond. 2) or not (cond. 1), is rated most favorably. Pauses at the end of word groups (cond. 6), though still above the middle of the scale, are rated less favorably. Considerably lower ratings are obtained for the three remaining conditions.

When the listener is unfamiliar with the message and intelligibility is therefore poor (unprepared) the results are rather different. The differences between the six pausing conditions are less extreme, although the relative ordering of the six conditions is hardly changed. Crucially however,

the condition with pauses at grammatically motivated locations (cond. 6) is now rated in between the two conditions with no pauses at all. Moreover, condition 6 is rated more favorably in an absolute sense by unprepared listeners than by prepared listeners. Since condition 6 was already rated above the middle of the scale by the prepared listeners, the improvement runs counter to the general tendency of unprepared listeners to regress towards the middle of the rating scale.

A classical ANOVA with listener type and pausing condition as fixed factors shows significance for pausing condition and for the pausing*listener type interaction. Newman-Keuls tests for contrasts ($p < .05$) show that conditions 1 and 2 do not differ from each other with prepared listeners; conditions 1, 2 and 6 do not differ from one another with unprepared listeners, as do conditions 3, 4 and 5.

4. CONCLUSION

Listeners evaluate the presence of speech pauses differently depending on the intelligibility of the stimulus. When they do not need the speech pauses in order to decode the message, all pauses, whether placed appropriately or not, are considered a nuisance. However, when the listener is not familiar with the text, and therefore needs the speech pauses in order to decode the message, one type of pausing is evaluated as positively as not pausing at all.

We now know that in the normal situation when the listener is unfamiliar with the message, e.g., when hearing a news broadcast, pauses inserted at the end of coherent word groups, and only these, help word recognition in continuous speech of low quality. Moreover, listeners do not judge the presence of such pauses unpleasant, even though the input speech is interrupted quite frequently. We therefore generally

recommend pausing at grammatical boundaries (but nowhere else) as a means of improving the intelligibility of low quality synthesis of continuous speech.

NOTE

We thank S.G. Nootboom for comments and discussion. This research was partly supported by the Foundation for Linguistic Research, which is funded by the Netherlands Organisation for Research, NWO, under grant # 300-161-035.

5. REFERENCES

- [1] GEE, J.P., GROSJEAN, F. (1983). Performance structures: a psycholinguistic and linguistic appraisal, *Cogn. Psych.* 15, 411-458.
- [2] HART, J. 'T., COLLIER, R., COHEN, A. (1990), A perceptual study of intonation, an experimental phonetic approach to speech melody, Cambridge: Cambridge UP.
- [3] NESPOR, M., VOGEL, I. Prosodic phonology, Dordrecht: Foris.
- [4] NOOTEBOOM, S.G. (1985). A functional view of prosodic timing in speech, in J.A. Michon, J.L. Jackson (eds.): Time, mind and behavior, Berlin: Springer, 242-252.
- [5] SCHARPFF, P.J. (1987). Effect of context and lexical redundancy on continuous word recognition, *Proc. 11th ICPHSc*, vol. 5, 43-47.
- [6] SCHARPFF, P.J., HEUVEN, V.J. VAN (1988). Effects of pause insertion on the intelligibility of low quality speech, *Proc. 7th FASE/SPEECH-88*, 261-268.

PHONOTACTIC KNOWLEDGE ACQUISITION BY SYLLABLE STRUCTURE MODELLING

Alessandro Falaschi

La Sapienza University - INFOCOM Dpt
Via Eudossiana 18, 00184 Roma - ITALY

ABSTRACT

This contribution illustrates applications for a structured analysis method of phonemic transcriptions. The analysis method is based on a morpho-syllabic automata which models the language word structure, and identifies a set of phonological units on a functional basis. Statistical analysis of the structured transcriptions permits definition of a Stochastic Phonotax, representing the phonotactical constraints strength, and allowing a probability value to be ascribed to phonemic sequences, to be related to their articulatory complexity and the information they carry.

1. INTRODUCTION

Phonemic transcriptions usually account for the representation of phonotactical constraints only from a contextual variations point of view, quite disregarding the timing and microprosodic aspects of the phonetic realization. A proper description of these phenomena should account for their relationships with the syllabic structure and constraints of the language. The results illustrated in the rest of the paper rely on a structured phonemic transcription method [1], [2] which explicitly accounts for the syllabic, stress and inflectional structures of words. Analysis of these transcriptions allows investigation of the phonotactical constraints embedded into the considered structures.

As a first result, the structural analysis defines a set of phonological units given by the original phonemic labels plus a set of three indexes reflecting their functional role in the model. These units will be indicated as Structured Phonemic Units (SPU). The

acoustical correlates of the SPUs can be investigated by automatic analysis of a speech corpus for which the structured transcription is available [3].

Acquisition of the inducted phonotactical constraints, together with their strength, is accomplished by gathering the statistics of the SPU pairs occurrence within a collection of structured transcriptions. The SPUs transition probability matrix define a Markov Source which has been called a Stochastic Phonotax [4], and which is an automaton whose states deal with the SPU phonemic label and are connected by a probability-weighted set of transitions.

The applications of the stochastic phonotax as a powerful phonotactical representation level are various. It can act as an acceptor automata for known phonemic sequences, ascribing them a probability value without any knowledge of the existing words frequency, only on the basis of the frequency of the SPU pairs it contains. In this sense, it can give some insight about the relationships between consonant clusters articulatory complexity and their frequency of occurrence. This relation can be further evidenced if the information carried by each phoneme of a string is evaluated, so that its periodic fluctuations due to the morphosyllabic structure of the language is evidenced.

A short review of the practical applications of the phonotactical knowledge acquired through the model is given at the end of the paper, as for automatic speech recognition and speech synthesis segmental quality evaluation.

2. THE SYLLABIC MODEL

As the steps required to obtain an SPU

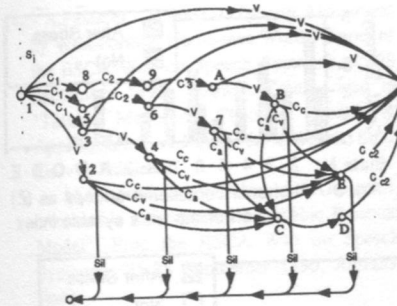


Fig. 1 - Transition Diagram for the intrasyllabic index defining SPU and FAUs

transcription of texts have already been fully exposed earlier [1][4], only a short review is made here.

The SPU transcription process starts from a syllabified phonemic transcription of texts, which is parsed according to the states of a morpho-syllabic automaton, resulting in the labeling of each phoneme with a set of three indexes which indicates their functional role within the considered structure. Fig. 1 shown the transition diagram for the intra-syllabic index, it being improved with respect to previously adopted ones [1].

A vowel falling in state number 2 is the first phoneme in the syllable to which it belongs, and states 4, 7 and B indicate respectively that the syllable begins with one, two, or three consonants. In this way, the nature of the final consonant cluster of closed syllables will depend on the length of the initial consonant cluster. The other two indexes considered are related to the stress and intersyllabic structure, reporting if the syllable follows the lexical word stress (or not), and about the ordinal number of the syllable to which the phoneme belongs.

An SPU transcription has been obtained over a 12543-word long text corpus, covering several areas, such as novels, newspapers, and textbooks. After a statistical analysis of the SPU pairs frequency of occurrence, a Stochastic Phonotax made of 780 states and 3773 transitions has been built, on which the applications described in 4 and 5 are based.

3. PHONETIC CHARACTERIZATION

The first two functional indexes which make the SPU definition, namely the intra-syllabic and the stress ones, naturally identify a set of Functional Allophonic Units (FAU) [5] for the phonemic alphabet considered. The effects of the functional role of a phoneme on its phonetic quality can be investigated by automatic methods. In particular, [3] reports about an acoustic-phonetic decoding system, in which vowels are differentiated, on the basis of their FAU class, as stressed or not, in open or closed syllable, or at word end; the consonants are mainly differentiated as being pre-vocalic or pre-consonantic. Some of the phonetic correlates such as duration and typical spectrum of the FAUs are given in [3].

4. PHONOTACTICAL EXPLOITATION

In the following will be reported some evaluations of the phonotactical knowledge captured by the Stochastic Phonotax. First of all, let us examine Fig. 2, where its conditional entropy [2] is reported, as function of the syllable ordinal number and stress relative position, representing the \log_2 of the average number of outgoing transitions for the equal-indexed phonotactical states. As expected, word endings are much more predictable than word roots, and post-stress syllables bring a quite constant (small) amount of information. A similar plot is given in fig. 3, where the conditional entropy is plotted as a function of the intra-syllabic index. Finally, fig. 4 gives the constraining power of phonemes when they follow the lexical stress or not.

As the Stochastic Phonotax may also serve as an acceptor automaton for unknown words, each phoneme pair of new words is scored with a probability value, so that their product is an estimate of the word probability. The $-\log_2$ of the word probability, once divided by the (phonemic) word length, gives an indication of the average word complexity. In fig. 5 a short list of words is ranked according to these two criteria. Moreover, fig. 6 shows the behaviour of the $-\log_2$ of the SPU conditional probability (e.g. the informative value of the new SPU) for the

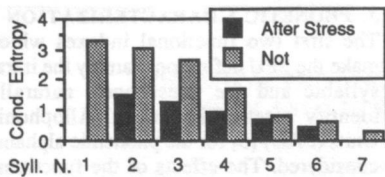


Fig. 2 - Phonotax conditional entropy as a function of the syllable number

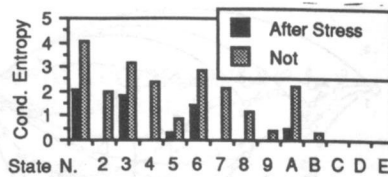


Fig. 3 - Phonotax conditional entropy as a function of the intra syllabic index

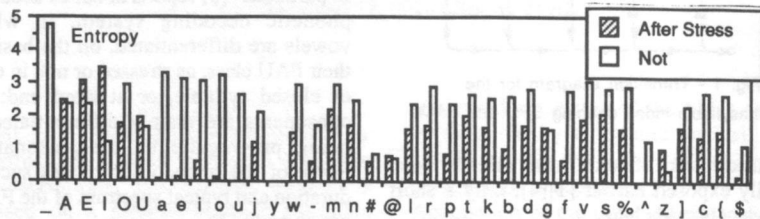


Fig. 4 - Phonotax entropy for the phonemes as a function of the stress relative position

same list of words. As it can clearly seen, the information carried by phonemes varies with a cyclic behaviour, with a ratio given by the stress and the syllabic structures. Moreover, the more informative phonemic events are clearly associated with the more articulatorily complex realizations, characterized by a quite low frequency of occurrence.

Once the Stochastic Phonotax has accepted a whole word, its average amount of information can be evaluated on the basis of the the information carried by its constituent SPU, giving a measure of the average complexity of the word. At the same time, a global probability value for the word can be computed, without using any knowledge about estimated word frequencies, but only on the basis of the occurring SPU pairs frequency. These are the quantities recorded in Fig. 5.

5. APPLICATIONS AND FUTURE

The phonotactical knowledge acquired through the morpho-syllabic model and on which the Stochastic Phonotax definition is based is mainly quantitative, thus allowing correct representation of the strength of these constraints for a language. Some early experiments [6] dealt with recognition of continuous speech, and in that case the Stochastic Phonotax helped in segmenting speech into words, thanks simply to the

morphological knowledge acquired. These experiment continued in [3], in which the microprosodic aspects tied up with the syllabic modelling are exploited.

The capacity of the Stochastic Phonotax to act also as a generator automaton allows its use for the automatic construction of well-formed nonsense words [4], which have been proposed as good material for evaluating the segmental intelligibility of speech produced by synthesizers.

As a final remark, the method is applicable to any language for which a syllabified phonemic transcription is available. For this reason, the author encourages interested researchers in other countries to let him know of their eventual intention of cooperation.

REFERENCES

- [1] - A.Falaschi, "A functional Based Phonetic Units Definition for Statistical Speech Recognizers", Proc. of Int. Conf on Speech Tec., Eurospeech 89, Settembre 1989, Paris, France
- [2] - A.Falaschi, "Phonotactical Constraints Strength Changes as a Function of Inside Syllable Position" Proc. of the 11th Int. Con. on Phonetic Sciences, August 1987, Tallin, Estonia, URSS
- [3] - A.Falaschi, "Phonotactically Driven Speech Recognition", Somewhere in these proceedings

- [4] - A.Falaschi, "Segmental Quality Assessment by Well-Formed Non-Sense Words", Proc. of the ESCA Wsh on Speech Synthesis, 25-28 September 1990, Autrans, France; also in "Talking Machines: Theories, Models & Applications", G.Bailly & C.Benoit Eds., Elsevier Publisher
- [5] - M.Giustiniani, A.Falaschi, P.Pierucci, "Automatic Inference of a Syllabic Prosodic Model", Proc the ESCA Wsh on Speech Synthesis, 25-28 September 1990, Autrans, France

- [6] - A.Falaschi, "Decodifica Acustico-Fonetica del Messaggio Vocale su basi Informativo-Strutturali mediante modelli di Markov nascosti", Tesi di Dottorato di Ricerca in Scienza e Tecnica dell'Informazione e della Comunicazione", Febbraio 1989, INFO-COM Dpt., Roma

Word	(English)	Probability	Word	Av. Information
di	of	4.80e-2	di	1.46
quella	that	2.17e-4	quella	1.74
iedi	feet	2.60e-5	interiore	2.03
allora	then	5.61e-6	inconciliabilmente	2.40
interiore	interior	7.55e-7	allora	2.49
sacra	holy	7.72e-8	iedi	2.54
timbri	stamps	7.50e-9	malpadroneggiabili	3.06
sollevio	relief	3.58e-10	voltandole	3.18
voltandole	turning over st.	2.97e-11	finestrino	3.39
finestrino	(car) window	5.74e-12	sollevio	3.49
orecchini	ear-rings	1.87e-13	timbri	3.86
inconciliabilmente	irreconciliably	1.85e-14	sacra	3.94
malpadroneggiabili	hard to masterize (?)	2.65e-17	orecchini	4.70

Fig. 5 - Probability and average Information for some Italian words

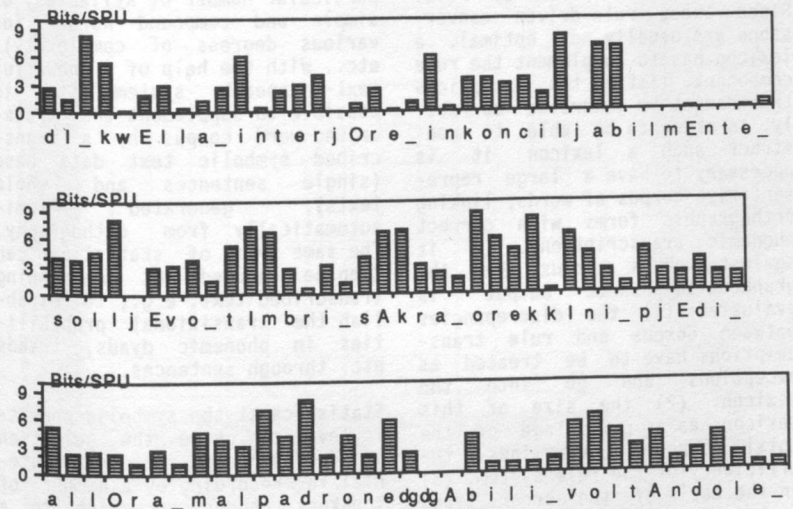


Fig. 6 - Information Flow for the SPUs as given by the Stochastic Phonotax

PHONETIC DATA BASES FOR GERMAN

K. J. Kohler

Institut für Phonetik und digitale Sprachverarbeitung
Kiel, Germany

ABSTRACT

The principles of automatically generating a large transcribed word corpus for German in a TTS environment, its extension to transcribed texts and the subsequent recording of an acoustic data base to be segmented and labelled are discussed.

1. SYMBOLIC, ACOUSTIC AND LABELLED ACOUSTIC SPEECH DATA BASES

In text-to-speech systems (e.g., RULSYS at KTH/Stockholm [1]), a grapheme-to-phoneme module transforms orthographic text into phonemic transcription by rule. Since these rule-driven conversions are usually not optimal, a lexicon has to supplement the rule component, listing the exceptions that cannot be generated correctly. In order to be able to construct such a lexicon it is necessary to have a large representative corpus of words, linking orthographic forms with correct phonemic transcriptions. It is against such a corpus that the grapheme-to-phoneme output is evaluated: (1) the discrepancies between corpus and rule transcriptions have to be treated as exceptions and go into the lexicon, (2) the size of this lexicon, as a percentage of the total corpus, determines the efficiency of the rule system, (3) on the basis of the error types contained in the lexicon, an attempt can be made to improve the grapheme-to-phoneme rules and to reduce the exceptions lexicon, in

a series of cycles, until an optimal balance is struck between lexicon size and number as well as complexity of the grapheme-to-phoneme rules, i.e. between the demands on storage and computing time, respectively.

Such a transcribed word corpus constitutes a symbolic speech data base, which - apart from being useful in TTS lexicon construction - can also be the basis for a great variety of phone statistics: frequency of phonemes, of phoneme sequences and clusters, of syllable types, of words containing a particular number of syllables, of simple and compound words (of various degrees of complexity), etc.. With the help of a powerful text-to-speech system it is possible to supplement the transcribed word corpus by a transcribed symbolic text data base (single sentences and whole texts), generated semi-automatically from orthography. The same type of statistics can then be carried out on running transcribed text, e.g., to establish the transitional probabilities in phonemic dyads, triads etc. through sentences.

Statistics at the symbolic phonetic level can guide the selection of word, sentence and text material for recording by a number of speakers, chosen according to a set of criteria (sex, age, social background, dialect etc.), to set up a representative acoustic speech data base for various

purposes (basic research, improvement of the phonetic rules in a TTS system, training speech recognisers). To be optimally useful the acoustic data base needs processing: (a) segmentation, (b) alignment of transcription symbols with the demarcated signal sections, (c) various signal analyses (e.g. FFT) whenever necessary (labelled acoustic speech data base). Then a new type of statistics becomes possible which combines information about the symbolic and signal aspects of speech. We may, for instance, want to collect all the instances of the signal portions corresponding to a particular phoneme in the labelled acoustic speech data base.

2. SYMBOLIC SPEECH DATA BASE FOR GERMAN

2.1. Generating a Corpus of Transcribed Words

At the Kiel Phonetics Institute, a corpus of 23986 orthographic words was compiled from two frequency-ordered word lists. One contained the just over 9000 most frequent words from a computer survey of the newspaper 'Die Welt' carried out at the German Department of Lund University. As a newspaper style lacks many common words, such as 1st and 2nd pers verb forms, and therefore does not provide, e.g., 'bin, bist' ((I) am, (you) are), it was necessary to supplement this corpus to get a more comprehensive and representative coverage. So a second list was compiled from all the words in a wide spread of literary texts (fiction at different levels, various forms of journalism, legal administration texts, user instructions, but not scientific texts), available in ASCII format in the German Department of Kiel University, amounting to appr. 24000 frequency-ordered items. They were edited: spelling errors were corrected and most personal and topographic names as well as

foreign loans excluded, except for very common ones, resulting in just over 21000 words. With the help of RULSYS support programmes, each word list was arranged in lines of five words and the lines consecutively numbered in steps of 5, starting with 0. Then the two lists, which had 6250 words in common, were combined to the total corpus of German words in frequency order. (Rolf Carlson at KTH carried out this amalgamation.)

This combined corpus was automatically transcribed by the grapheme-to-phoneme conversion module within the German TTS in the RULSYS environment, marking lexical stresses (' before the stressed vowel) and word boundaries in compounds (#), as well as affixing a general word class marker (w). The output was then manually corrected with regard to errors in phonemes, stresses, and word boundaries; at the same time function words were marked by + after the segmental string, and the general word class indicator changed to various sub-categorisations in function words. This function word marking and classification is important for a syntactic analysis component within TTS. The corpus of German words thus has two related files, an (orthographic) text file and a (phonemic) transcription file, with identical item arrangements. The following examples illustrate the principle of corpus organization.

CORP.TX

0 DER, DIE, UND, IN, VON.

...
110 SONDERN, IHRER, BUNDESREPUBLIK,
NEU, HIER.

CORP.FO

0 D'E:R+de, D'I:+de, 'UND+bc,
'IN+pp, F'ON+pp.

...
110 Z'ONDERNW, 'I:RER+ns,
B'UNDEZ#REPUBL'IKw, N'EUw,
H'I:Rw.

As can be seen the transcription is very close to conventional

spelling and quite abstract, i.e. morphophonemic rather than phonemic or phonetic, in that it keeps final voiced obstruents in order to preserve the relationships within inflectional paradigms, e.g. 'Haus' and 'Häuser' are both transcribed with Z at this level of abstraction although the former is [haus], the latter [hauzəs] at the phonetic realisation. Similarly, [ə] and [ʊ] are represented as E and ER, because the former can be derived from the latter by a lower-level phonetic rule. By applying these phonetic rules other, more-phonetic corpora can be derived, as, e.g., in CORP.FON

0 D'E:r+de,D'I:de,'UNT+bc,
'IN+pp,F'ON+pp.

...
110 Z'ONDrNw,'I:Rr+ns,
B'UNDEOS#REPUBL'IKw,N'EUw,
H'I:rw.

(Note the use of " for secondary stress and r for [ʀ].) By a simple conversion programme the phonetic notation can also be transformed into SAM-PA [5]:

CORP.SAM
0 d'e6,d'i,'Unt,'In,f'On.

...
110 z'ONDn6n,'ir6,
b'Und#srEpUbl'Ik,n'OY,h'i6.

2.2. Generating an exceptions lexicon

The machine output from the TTS grapheme-to-phoneme module (CORP.MA) and its manual correction (CORP.FO) differ in those items that are not generated correctly by rule. They have to be included in the exceptions lexicon of the TTS system. This lexicon is generated automatically with the help of another RULSYS support programme that compares CORP.MA and CORP.FO (in relation to CORP.TX) and lists all the non-congruous items together with their orthographic versions in a new file, which is brought into alphabetical order by SORT. The following is an excerpt, providing, in each line, the orthographic word, the correct

transcription, the machine transcription and a number referring to the frequency rank of the item in the corpus.

LEX.GE
AM 'AM+pp * 'A:Mw ## 45
AMEISEN 'A:MEIZENw * AM'EIZENw
6452
AMERIKAREISE AM'E:RIKA:#R'EIZEw
* AMERIKAR'EIZEw ## 6454
AMERIKAS AM'E:RIKA:Zw
* 'A:MERIKAZw ## 2655

At present, this comparison yields 4150 errors, i.e. a rule efficiency of 82.7%. Changes in the grapheme-to-phoneme module can be quickly tested in their power of reducing the lexicon size by running the automatic programmes of corpus transcription, comparison and lexicon generation with reference to the transcribed data base CORP.FO, which is thus of fundamental importance for TTS development. But the procedure described so far has a serious flaw in that it includes all the items of an inflection or derivation paradigm in the lexicon, when there is a discrepancy between CORP.FO and CORP.MA, although the listing of a root would be sufficient to cover the exceptions of the whole paradigmatic set. This optimisation can be achieved by generating a corpus with suffix markings and root forms, applying a TTS suffix stripping module to CORP.TX and putting the result in the same format, as in CORP.IN

150 MÜB,MÜB-EN,IHR-EN,FRAG,FRAG-E.

The generation of this inflected corpus is semi-automatic through further RULSYS support programmes and with manual correction of wrong suffix markings. Expanded CORP.TXR and CORP.FOR, containing the added roots, are also generated semi-automatically. The result is a corpus of 29183 items, i.e. 5197 roots have been added. The creation of the exceptions lexicon is then again fully automatic and yields 3754 entries at present,

i.e. a reduction of 9.5% compared with the original base not containing roots. This also means that the rule efficiency has been increased and that the TTS system has been made a great deal more general, allowing the correct generation of many more exceptions than are actually contained in the original data base. All that is necessary is the application of the same suffix stripping module in TTS processing of orthographic text input and subsequent lexicon look-up procedures, followed by a root modification module making phonemic adjustments in morphemic composition.

2.3. Generating corpora of transcribed texts

With the help of the German TTS system, developed and improved on the basis of an extensive word corpus, it is now possible to enlarge the symbolic speech data base for German and incorporate phonetically transcribed texts in addition to isolated words. Because the combined rule and lexicon efficiency in grapheme-to-phoneme conversion is very high, manual correction is minimal, and new transcribed texts can thus be generated from orthography more or less automatically. I have done this, starting with the standard sentences for German speech tests (Berlin (Sotschek) and Marburg, [4]) and two standard texts (The Northwind and the Sun; The Butter Story), illustrated in the following excerpt in adapted SAM-PA transformation (Q = [ʔ]):

001 h'OYt@ QIst S'2n@s
fr'yIINsv"Et6.
002 di z'On@ l'Axt.
003 QAm bl'AU@n h'Im@l ts'i@n di
v'OIk@n.

These transcribed data were then searched by appropriate programmes for all the phoneme dyads in German (including the transition to the first and from the last phoneme of a sentence). Frequencies of occurrence were entered into matrices, and empty cases

that resulted from phonotactic restrictions and low phonotactic probabilities disregarded. For the remaining empty cases further sentences were constructed to cover all the 1308 most likely phoneme dyads in German by at least one instance. This resulted in a corpus of 398 sentences plus the two texts.

3. ACOUSTIC DATA BASE FOR GERMAN

The text materials in 2.3. were DAT recorded by 25 male and 25 female speakers, one in each group reading the whole corpus, the others various subsections of appr. 80 sentence equivalents, on average. The total recorded corpus comprises 4836 sentence equivalents, available in computer-readable form (16 kHz, 16 bit) on cassettes, with headers providing information about speaker, sentences etc. It will be segmented and labelled following the principles in [2,3]. The aim is to progressively automatise the processing.

4. REFERENCES

- [1] CARLSON, R., GRANSTRÖM, B. & HUNNICUTT, S. (1990), "Multi-language text-to-speech development and applications", in *Advances in Speech, Hearing, and Language Processing*, Vol. 1, (W.A. AINSWORTH, ed.), London: JAI Press, 269-296.
- [2] CARLSON, R. & GRANSTRÖM, B. (1985), "Rule controlled data base search.", *STL-QPSR*, 4, 29-42.
- [3] id. (1986), "A search for durational rules in a real-speech data base", *Phonetica*, 43, 140-154.
- [4] SOTSCHKEK, J. (1976), "Methoden zur Messung der Sprachgüte I: Verfahren zur Bestimmung der Satz- und der Wortverständlichkeit", *Der Fernmelde-Ingenieur*, 30(10), 1-31.
- [5] WELLS, J. C. (1987), "Computer-coded phonetic transcription", *JIPA*, 17, 94-114.

ACOUSTICAL DATA BASE AS A TOOL FOR THE RESEARCH OF VOWEL SYSTEMS

Antti Iivonen

Department of Phonetics, Helsinki, Finland

1. ABSTRACT

A formant data base representing ca. 20 languages has been collected. The main purpose is to use this data base for a comparison of language specific vowel qualities and vowel systems, but it also can be used as a research tool to avoid sources of errors due to research methods and materials.

For comparison of vowels, an $F1/F2$ -plot on a Bark-scale has been utilized. This representation can be considered to be an approximation for psycho-acoustical vowel space. The vowels are presented as 1 Bark-sized circles in order to show the auditory distances between them (cf. [3]).

2. APPROXIMATION OF THE PSYCHO-ACOUSTICAL VOWEL SPACE

Several studies have indicated that the simulation of vowel space using the first two formants $F1$ and $F2$ on a Bark scale is a strong approximation of vowel perception. Fig. 1 shows a gliding vowel series [i-e-ε-æ-a-ɔ-o-u], produced as a continuous utterance by the author according to the Finnish articulation base. The glide was analyzed in 20 ms steps, and it forms a trajectory on the $F1/F2$ -plot which corresponds well to the traditional location of vowel qualities on a vowel quadrilateral. FFT spectra (with a 30 ms time window) were used. The first approximation of the glide is presented in Fig. 1.

The power of the $F1/F2$ representation might be explainable on the basis of motor perception theory: motor facts correspond to the perceptual ones in the sense that the listener 'hears' an $F1/F2$ -pattern as a corresponding tongue/lip gesture. A vowel

quality is presented as a freely mobile, 1-Bark-sized circle on the $F1/F2$ -plot. Its position is calculated from its measured Hz-values. A Bark circle can be understood as a point that scans its surrounding space to check if there is per-

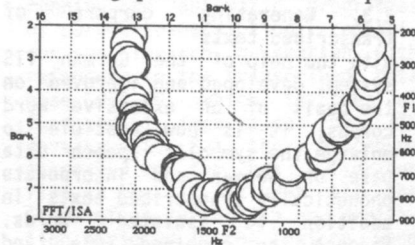


FIG. 1. Continuous vowel series [i-e-ε-æ-a-ɔ-o-u] produced by the author. The circles represent the $F1/F2$ points in 20 ms time intervals.

ceptually distinct psycho-acoustic distance from other vowels. In most cases a circle covers an area smaller than the distribution of the single occurrences of the same phoneme. It can be assumed that if two circles overlap, the listener may have difficulty distinguishing the vowels considered. Fig. 2 shows the East Central Bavarian vowel means measured by Trau Müller [11]. The rounded, front vowels represent earlier lateral sounds in the dialect. According to Trau Müller's auditory judgement, it is difficult to distinguish the vowels of the pairs [e, ε], [ø, œ], and [o, ɔ]. This effect has the correspondence in Fig. 2: The circles of these vowels overlap.

Lindblom [6] calculated the first four formant values of 19 "quasi-cardinal vowels" representing psycho-acoustically equal quantization steps. Fig. 3 presents these vowels on an $F1/F2$ -plot. It can be seen that the vowels are mainly

equidistant on the plot concerning each formant separately, but the distances are generally greater for $F2$ than for the $F1$.

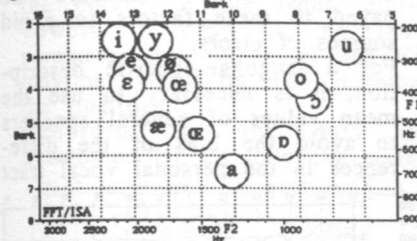


FIG. 2. East Central Bavarian vowels. Data from [11]. Means of three speakers. Note that some mid-vowels overlap which corresponds to auditory confusion.

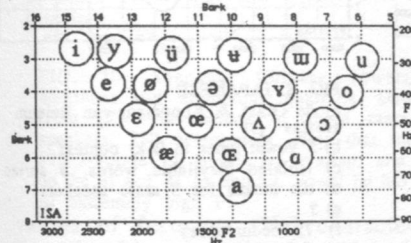


FIG. 3. Theoretical quantizing of a vowel space, formant data from Lindblom [6]. The positioning of 19 quasi-cardinal vowels.

Fig. 3 also illustrates that an empty space remains between the vowel circles. This is understandable if, for example, we consider the total number of the possible

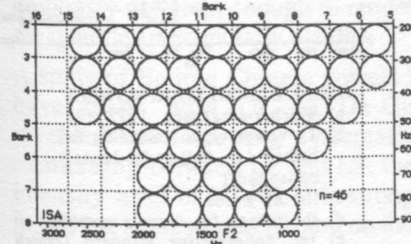


FIG. 4. If the articulatorily possible vowel space is filled with 1 Bark-sized circles, the result is 46 circles. This number seems to be near the amount of universally distinguishable vowels.

phonetic vowel symbols. The newest IPA chart (1989/1990) contains 25 vowel signs, the Stanford Phonological archive has even more: 37. Articulatorily, the number of possible vowels is unlimited.

Hence, the explanation for the maximal number of vowel qualities must lie in the human auditory capacity. When the articulatorily possible $F1/F2$ vowel space is filled with 1 Bark-sized circles, the result is 46 circles (Fig. 4). This number corresponds well to the number of the possible vowel qualities (quoted above) (if height, frontness and rounding are considered; cf. [5]).

A diphthong can be depicted as a $F1/F2$ -glide from beginning to end (excluding the transitions). The glide can be measured in 10 ms intervals (Fig. 5).

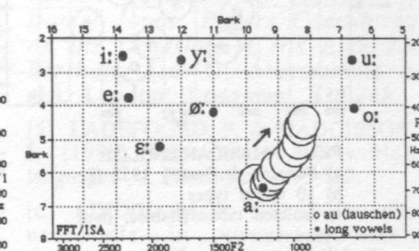


FIG. 5. One possible way to present a diphthong. The glide of a single occurrence has been displayed in 10 ms steps with the speaker's long vowels (=means) as the background. A North German male speaker, [au] in lauschen.

3. LANGUAGE DATA

Most of the formant data included in this study have been collected from the literature. The following are the main features that were included in the data base: a) author(s), b) vowel phonemes (allophones, types) considered, c) utterance type used (isolated words, list of words, the carrier sentence, etc.) plus the consonant context of vowels, d) number and sex of the informants, e) language (dialect, regiolect, sociolect), f) number of occurrences, g) equipment utilized for analysis, h) formant measurement principles, and i) formant values. Figures 6-9 illustrate four language examples. Considering the research features, it can be argued that very few language comparisons can be made with-

out a bias that is a result of the differences in research methods.

4. CRITICAL REMARKS

In some cases, the representation based on F1/F2 proves to be problematic. The areas of concern are (1) the non-phonemic factors influencing the vowel positioning on the F1/F2-plot and (2) the

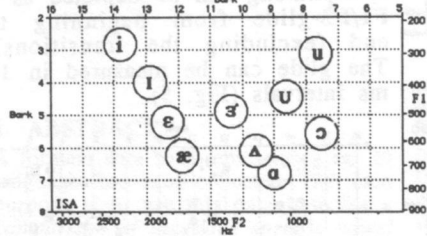


FIG. 6. AMERICAN ENGLISH
a) Peterson & Barney 1952 [8]
b) 10 vowel types
c) isolated monosyllables, [h-d] context
d) the means for 33 male speakers
e) majority General American
f) 666 occurrences
g) spectrograms and sections.

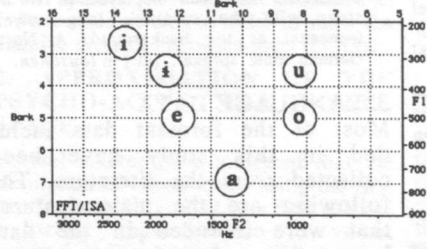


FIG. 7. POLISH
a) Jassem 1964 [4]
b) 6 vowel types
c) 44 "items" in a word list
d) the means for 3 male speakers
e) "educated Polish"
f) 132 (?) occurrences
g) Kay El. Sonograph 661, broad band sonograms, broad and narrow band sections

role of formants in vowel quality characterization.

3.1. Non-phonemic factors

F1/F2 positioning is influenced by vowel reduction (due to the stress degree), vocal tract length, larynx height, allophonic variation, several voice quality

types, pure chance, and formant measurement principles. Special attention must therefore be paid to these factors to avoid sources of errors.

For a language-specific description, it is necessary to use the mean values of several speakers to avoid the bias of the differences in the personal vocal tract

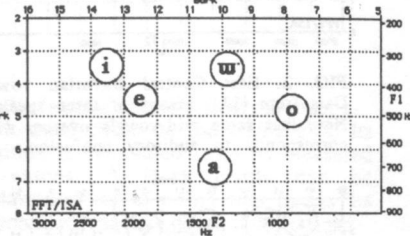


FIG. 8. JAPANESE
a) de Graaf & Koopmans-van Beinum 1982/83 [7]
b) 5 vowel types in [k-k] context
c) isolated bisyllabic words, 5 series
d) the means for 3 male speakers
e) ?
f) 75 occurrences
g) LPC-analysis

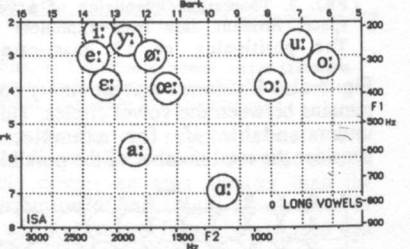


FIG. 9. DANISH LONG VOWELS
a) Fischer-Jørgensen 1972 [2]
b) 11 vowel types, including [a:] before /-r/, [h-l] or [h-dental consonant]
c) list of words
d) the means for 8 male speakers
e) rel. conservative standard Danish
f) 88 occurrences
g) narrow and wide band spectrograms

length. The means are also necessary, because the single occurrences show considerable variation.

3.2. Difficulties in formant approach

In languages like Swedish, Danish (cf. Fig. 9), and Chinese (Fig. 10), the corner of the front, close vowels is crowded, so that parameters other than F1 and F2 may be needed for distinguishing the qualities.

According to the data in Svantesson [9], the following mean

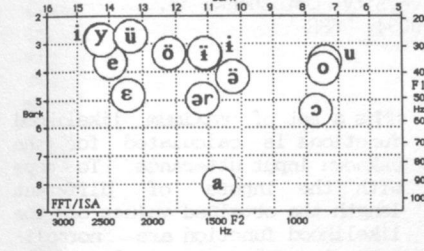


FIG. 10. SHANGHAI CHINESE VOWELS, dominant and allophonic qualities. Data from [9]. The means of three male speakers, 230 occurrences. Note that the F1/F2-points of [i] and [y] overlap, and do the [i] and [ɨ], [o] and [u]. Vowel [a] has an extremely high F1.

values of formants (Hz) characterize the Shanghai Chinese [i] and [y] by three male speakers:

	F1	F2	F3
i	274	2442	3387
y	270	2455	3465
diff.	4	13	78

The differences of the formants are obviously not great enough to render the qualities perceptually distinct.

A study of Beijing Chinese vowels I have made with Dr. Li De-Gu revealed the following systematic relationship between L2 and L3: the intensity level of F2 (=L2) is strong in [y], but suppressed in [i], and on the contrary L3 is very strong in [i] but weak in [y] (cf. also the discussion in [1] and [10]).

REFERENCES

[1] AALTONEN, O. (1985) "The effect of relative amplitude levels of F2 and F3 on the categorization of synthetic vowels", *J. of Phonetics* 13, 1-9.

[2] FISCHER-JØRGENSEN, E. (1972) "Formant frequencies of long and short Danish vowels", *Studies for Einar Haugen presented by friends and colleagues* (ed. E. S. Rirchow et al.), The Hague/Paris: Mouton, 189-213.

[3] IIVONEN, A. (1987) "Regional differences in the realization of Standard German vowels", *Proceedings of the XIth International Congress of Phonetic Sciences, Tallinn*. Vol. 4, 161-164.

[4] JASSEM, W. (1964) "A spectro-graphic study of Polish speech sounds", *In Honour of Daniel Jones: Papers Contributed on the Occasion of His Eightieth Birthday* (ed. D. Abercrombie & al.), London: Longman, 334-348.

[5] LADEFOGED, P. & MADDIESON, I. (1990) "Vowels of the world's languages", *J. Phonetics* 18, 93-122.

[6] LINDBLOM, B. (1986) "Phonetic universals in vowel systems", *Experimental Phonology* (ed. J. J. Ohala & J. J. Jaeger), Orlando etc.: Academic Press, 13-44.

[7] de GRAAF, T. & KOOPMANS-van BEINUM, F. J. (1982/83) "Vowel contrast reduction in Japanese compared to Dutch", *Proc. of the Institute of Phonetic Sciences, Amsterdam*, 27-38.

[8] PETERSON & BARNEY (1952), "Control methods used in a study of the vowels", *J. Amer. Acoust. Soc.* 24, 175-184.

[9] SVANTESSON, J.-O. (1989) "Shanghai vowels", *Working Papers 35* Lund University, Dept. of Linguistics, 191-202.

[10] SCHWARTZ, J. L. & ESCUDIER, P. (1987) "Does the human auditory system include large scale spectral integration?", *The Psychophysics of Speech Perception* (ed. M. E. Schouten), Dordrecht etc.: Nijhoff, 284-292.

[11] TRAUNMÜLLER, H. (1982) "Der Vokalismus im Ostmittelbairischen", *Zeitschrift für Dialektologie und Linguistik*, II. Jahrgang, Heft 3, 289-333.

AN INTERACTIVE SYSTEM FOR LANGUAGE IDENTIFICATION

V.B. KUZNETSOV

Department of Applied and Experimental Linguistics,
Moscow Linguistics University, Ostozhenka 38,
Moscow 119034, USSR

ABSTRACT

This paper describes a prototype language identification system LANIS based on hidden Markov modelling (HMM) of the way in which sounds combine together in a particular language. The results of LANIS performance in closed identification tests comprising four languages (English, French, German, Russian) are presented and discussed. The speech material used in the tests was of two kinds: transcribed texts and oral speech. An attempt is made to assess the effects of the length of the unknown speech sample and the structure of HMM (the number of hidden states and sound classes) on identification score.

1. INTRODUCTION

It is a well established fact that frequency distributions and patterns of occurrence of phonemes vary from language to language. House and Neuburg [1] have shown that this information can form the basis of a powerful language recognition tool even if the phonetic inventory is reduced to a very small number of gross phonetic categories such as stop, fricative, nonvocalic sonorant, vowel, and silence. To model the phonotactics of the broad phonetic categories the authors applied an HMM technique. The devised procedure for automatic language identification goes as follows. During training session reference HMMs are generated by means of a maximization technique for the languages of interest. Then, using the constructed

HMMs a set of maximum likelihood functions is calculated for the unknown input utterance. To cope with the inputs of different length the obtained values of the likelihood function are "normalized", i.e. the natural logarithms of the function values are divided by the number of elements in the utterance. According to the decision rule the unknown utterance is assigned to the language whose HMM has produced the highest score. It follows from the above that the identification takes place within the limits of a closed trial in which it is predetermined that the input utterance is spoken in one of the languages from the fixed set. Due to the fact that the speech material was restricted to transcribed texts and the performance testing of the identification procedure was rather sketchy, the results obtained by House and Neuburg are primarily of methodological significance. The present work is designed to develop and evaluate a prototype language identification system implementing the two main features of the approach suggested by House and Neuburg: broad phonetic classes and HMM technique. Two more issues are addressed: the determination of the optimal structure of the HMM and the minimum length of the unknown utterance sufficient for its reliable recognition.

2. SYSTEM DESCRIPTION

The basic structure of the LANIS

system is shown in Fig. 1. The implementation strategy was strongly influenced by research needs. The system can deal with two types of input: a string of phonetic symbols or the speech signal. It consists of four program modules and a database.

DATA PREPROCESSING MODULE. The function of this module is to extract from the input data set the information needed for the generation of an HMM. Firstly, the frequency of occurrence of phonetic elements is counted and the average duration for each class of speech segment is determined. Secondly, the expert specifies the structure of HMM: there can be up to 5 states and 10 phonetic classes. Thirdly, the initial values of the HMM parameters are either computed using a standard formula or set by the expert. Then, the input data set and the information obtained become available either for HMM construction or language identification.

HMM GENERATION MODULE AND DATABASE The parameters for each of the language models are estimated from the training data set (maximum size - 2000 elements) using the Baum-Welch algorithm. There is a special mode of the algorithm application (chosen by the expert) where the parameter variances are evaluated as well. The resulting HMM with the associated value of the maximum likelihood function are stored in the database. For the present the storage capacity of the database is limited to 5 languages and 10 HMMs. If a language is represented by more than one HMM, then an average HMM can be calculated and stored. For any pair of languages marked by the expert the significantly different HMM parameters can be discovered.

LANGUAGE IDENTIFICATION MODULE. The identification program tests the unknown utterance against the reference HMMs. Then, judging from the computed values of the likelihood function the program decides the language of the test utterance.

The segmentation module is described in a separate section below.

3. EVALUATION OF PERFORMANCE ON TRANSCRIBED TEXTS

To master the identification procedure and to evaluate its efficiency we used transcribed texts in four languages: English, German, French, and Russian. The length of the texts varied from 2.0 to 3.5 thousand phonetic symbols. Two types of transcription were used. One had a phonetic inventory of 6 classes: 1) vowels and sonorants, 2) voiced plosives and affricates, 3) voiced fricatives, 4) voiceless weak fricatives, 5) voiceless strong fricatives, and 6) voiceless plosives and affricates. The other represented the speech stream with 4 categories: they were the same as in the previous inventory, except for the first three classes that had fallen into one category.

For the generation of reference HMMs we used as training data either 1000 element extracts or the whole texts. The size of the test samples was 100 and 300 segments. Before reporting the results of the recognition tests, we want to discuss the data shown in Table 1. For the two types of the test samples Table 1 presents to the left of the sloping line (/) the average frequency of occurrence of the phonetic classes (multiplied by 1000), to the right - the corresponding variation coefficient (the ratio of the standard deviation of the frequency score to the mean, expressed in percent). Each mean was derived from 10 to 12 measurements. It is apparent from this data that the reduction in the size of the test utterance from 300 to 100 elements brings about on average a two-time increase in the dispersion of frequency of occurrence. There were a few cases where phonetic classes with the lowest frequency of occurrence were not present in the samples of 100 elements at all. Not surprisingly, the

results of language identification on these samples are rather poor. Table 2 displays the identification error rate for each of the languages and the averaged score. Several important characteristics of the tests are specified in Table 2: the structure of the reference HMMs; the total number of identifications (N), distributed more or less equally among the languages; the size of the unknown utterance (V) and its origin, i.e. whether it was taken from the test or training data sets. It should be remembered that in the latter case the training set covers the whole text of a particular language. From the analysis of the data presented in Table 2 two major points can be made. First, four phonetic categories are not enough to discriminate reliably between the languages in question. Second, the size of the test set should be about 300 elements. It was noticed that the choice of the initial HMM parameter values can play an important role in language identification. However our limited experience prevents us from making any definite suggestion on how to handle this problem.

4. AUTOMATIC SEGMENTATION AND LABELLING MODULE

The LANIS system incorporates a module performing speaker- and language-independent automatic segmentation and labelling (ASL) of continuous speech. Each extracted segment is identified as one of the following: 1) vocalic segment, 2) strong fricative, 3) weak fricative, 4) unidentified fricative, 5) stop, 6) silence. Segmentation and labelling are carried out on the basis of four parameters: fundamental frequency, zero-crossing rate, intensity and a parameter indicating sharp drops in intensity of the speech signal. The parameters are computed once every 16 ms.

The ASL program is a knowledge-based procedure. The rules implemen-

ted in the program were initially worked out by an experienced phonetician who analyzed the traces of the four parameters. In its present form the ASL module works as a two-pass routine. The first pass makes a provisional identification of each 16 ms frame as a member of our set of phonetic categories. Since this results in too many implausible identifications, the second pass attempts to group these labelled frames into likely phonetic segments, producing as output a string of phonetic symbols each of which has a duration value.

The ASL module is still at an experimental stage and no serious attempt has been made to test the degree of the accuracy obtained. None the less we decided to carry out a pilot experimental identification using the outputs of the ASL module. Speakers of the four languages in question were recorded while reading a piece of prose. 30 second passages of four male speakers were used to construct the reference HMMs. For a given language the training data sets of the other languages served as test samples.

Table 3 lists the normalized values of the maximum likelihood function computed for each speech sample. Negative signs are omitted in the Table. Analysis of the figures in the rows reveals that in all cases the highest value was obtained where a speech sample had been tested against its "native" reference HMM. Evidently, this result does not lead to any general conclusion, and yet it is not felt to be discouraging.

5. REFERENCES

[1] HOUSE A.S. and NEUBURG E.P. (1977), "Toward automatic identification of the language of an utterance: 1: Preliminary methodological considerations", J. Acoust. Soc. Am., 62(3), 709-713.

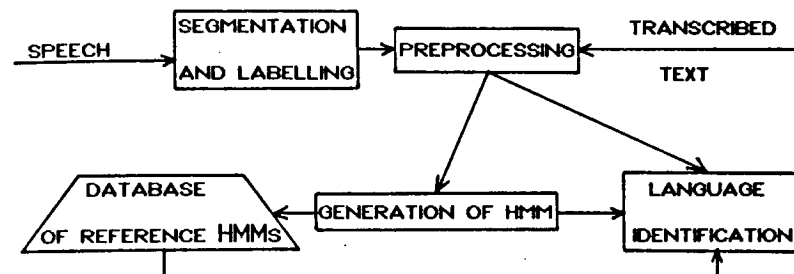


Figure 1. Basic structure of the LANIS system

Table 1. Frequency of occurrence / Variation coefficient

LANGUAGE	PHONETIC CATEGORIES						
	1	2	3	4	5	6	
ENGLISH	100	602/7	73/30	86/32	42/31	75/20	120/30
	300	610/5	78/22	82/33	50/14	68/27	112/8
FRENCH	100	638/7	75/36	62/36	59/38	62/17	104/24
	300	651/3	82/17	52/29	56/24	52/16	107/17
GERMAN	100	645/4	102/20	30/72	35/51	68/30	120/18
	300	659/3	88/16	35/37	39/25	62/23	116/8
RUSSIAN	100	665/6	48/42	68/39	20/78	59/28	140/22
	300	650/2	55/24	58/23	25/30	62/13	150/19

Table 2. Identification error rate (%)

	4 categories 3 states		6 categories 4 states		
	training V=300 N=42	training V=300 N=40	test V=300 N=37	training V=100 N=40	test V=100 N=54
GERMAN	46	10	30	20	29
ENGLISH	60	30	29	20	50
FRENCH	10	40	10	50	47
RUSSIAN	27	0	20	10	27
AVERAGE	36	20	22	25	38

Table 3. Normalized values of maximum likelihood function (negative signs are omitted)

Test sample of speech	Reference HMMs			
	German	English	French	Russian
German	0.594	0.607	0.612	0.673
English	0.554	0.541	0.547	0.608
French	0.538	0.529	0.523	0.547
Russian	0.427	0.433	0.427	0.397

Modélisation explicite de la coarticulation pour le décodage acoustico-phonétique : les triplets phonétiques

Y. Laprie, F. Lonchamp

CRIN / INRIA-Lorraine, Nancy, France
Institut de Phonétique de Nancy, Nancy, France

Abstract

This paper presents a knowledge-based approach of acoustic-phonetic decoding of continuous speech. Knowledge is stored in the form of triplets which represent contextual phone prototypes in terms of acoustic events and acoustic correlates. We describe how vowel centered triplets and plosive centered triplets are matched to the reference triplets. We then show how relaxation techniques may be used to increase the consistency of the global solution. Lastly, we indicate the aspects of the triplet approach which requires further investigation.

1 Introduction

Depuis plusieurs années l'équipe de parole du CRIN consacre une part importante de ses efforts à l'approche à bases de connaissances du décodage acoustico-phonétique. Cela a permis de développer Aphodex [Fohr 89] qui est un système expert en lecture de spectrogrammes. Cette expérience a mis en évidence qu'il est primordial de prendre en compte le contexte pour représenter la connaissance comme pour l'utiliser afin de reconnaître un son inconnu.

Cela nous a conduit à proposer un nouveau grain de connaissance pour l'approche experte en DAP : *le triplet phonétique*. Il s'agit d'un prototype de son en contexte (par exemple le son /t/ en contexte /a/ à gauche et /i/ à droite, noté /_at_i/) structuré en deux niveaux de description, une description acoustique en termes d'événements acoustiques de base (par exemple les formants), une composante experte faisant appel aux combinaisons d'événements acousti-

ques (que nous appelons indices acoustiques) reconnues significatives par l'expert (par exemple la position relative de deux formants dans un contexte phonétique donné). Pour justifier l'utilisation des triplets en DAP il faut montrer les avantages qu'ils apportent pour les points suivants :

- la précision et la pertinence de l'information acoustico-phonétique qu'ils permettent de stocker,
- la facilité de mise en œuvre de cette connaissance et notamment dans le cadre des techniques de l'intelligence artificielle qui peuvent permettre de contrôler le processus de décodage.

2 Pertinence de la représentation de la connaissance acoustico-phonétique

2.1 Représentation d'un triplet

La description acoustique d'un triplet fait appel aux événements acoustiques suivants qui peuvent être accrochés à chacune des frontières du son central d'un triplet :

- les trajectoires formantiques (valeur et pente à la frontière complétée par la valeur au centre du triplet),
- la barre d'explosion décrite par les principales concentrations d'éner-

gie, la durée, l'intensité, les indices de compacité et de diffusion,

- bruit de friction défini par la limite inférieure de bruit et son intensité.

La composante experte regroupe l'ensemble des indices acoustiques dont la pertinence pour la reconnaissance d'un triplet a été reconnue par l'expert. Ces indices sont des combinaisons d'événements acoustiques, par exemple la position relative de deux formants ou encore d'un formant par rapport à une concentration d'énergie de la barre d'explosion.

2.2 Outils destinés à l'étude de l'approche par triplet

Nous avons d'abord développé un éditeur de triplets (intégré à Snorri) qui permet à l'expert de construire les triplets directement sur le spectrogramme.

Cet éditeur permet de manipuler aussi bien les triplets de référence que les instances de triplets (de la phrase à décoder) puisque leur représentation est identique. Par ailleurs, nous avons intégré à cet éditeur le module de segmentation automatique (développé dans le cadre d'Aphodex), le module de suivi automatique des formants [Laprie 90] et celui destiné à la localisation des barres d'explosion. L'utilisateur de l'éditeur peut donc s'aider de ces modules pour construire les triplets de référence qui constituent la base de connaissances.

2.3 Identification de triplets vocaliques et occlusifs

Le système de décodage que nous avons développé commence par segmenter et étiqueter en classes phonétiques (voyelles, fricatives, occlusives, sonantes et autres sons) la phrase grâce au module de segmentation automatique. La description acoustique des instances de triplets est ensuite construite automatiquement en faisant appel au suivi automatique de formants et au détecteur-analyseur de barres d'explosion. Comme les autres détecteurs acoustiques n'ont

pas encore été réalisés notre système se limite à identifier les triplets centrés sur une voyelle ou une occlusive.

Afin de vérifier la pertinence de la description acoustique des triplets, nous avons conçu et testé les modules d'identification pour les voyelles et les occlusives. Ces modules évaluent une distance acoustique grossière entre l'instance à reconnaître et les triplets de référence.

2.3.1 Triplets centrés sur une voyelle

La distance calculée ne fait intervenir, pour l'instant, que les trajectoires formantiques, après que le triplet inconnu et le triplet de référence ont été normalisés en temps. Le calcul consiste à évaluer le coût de déplacement des arcs de trajectoires formantiques du triplet inconnu vers ceux du triplet de référence, et pénalise lourdement les grands écarts de pentes formantiques. Pour chaque instance le module d'identification acoustique donne les triplets de référence les plus proches du triplet inconnu.

Nous avons testé ce module sur 25 voyelles extraites de deux phrases du corpus de parole continue < La bise et le soleil... > pour 10 locuteurs (soit 250 voyelles). Les voyelles de ces deux phrases sont bien réparties dans le triangle vocalique. Comme le nombre des triplets de ce test est limité nous avons seulement pris en compte les trajectoires formantiques du son central pour le calcul de la distance acoustique. Cela pénalise donc les résultats de notre module d'identification acoustique mais donne une idée plus juste des performances réelles de notre approche. En conservant les trois meilleurs candidats pour chaque voyelle nous avons obtenu 91% de bonnes reconnaissances, ce qui est tout à fait encourageant. Après examen des erreurs, il apparaît qu'il s'agit dans la plupart des cas d'erreur de suivi de formants, ce qui souligne l'importance de disposer de détecteurs acoustiques très performants.

2.3.2 Triplets centrés sur une occlusive

En ce qui concerne les occlusives il faut ajouter au terme de distance concernant les trajectoires formantiques un terme destiné à évaluer la distance entre

deux barres d'explosion. Cette distance est calculée en considérant les concentrations d'énergie des barres d'explosion à comparer et consiste à trouver le plus grand recouvrement possible entre les deux bursts. Cette distance est pondérée par les coefficients de diffusion et de capacité. Le module d'identification acoustique correspondant a été testé sur 15 occlusives pour 10 locuteurs du même corpus que pour les voyelles. Les résultats en ne considérant que les deux meilleurs candidats sont moins bons (71%) que pour les voyelles. Pour améliorer ce taux de reconnaissance nous étudions des procédures de comparaison de barres d'explosion plus évoluées qui séparent l'effet de la voyelle de celui de l'occlusive.

3 Techniques d'intelligence artificielle pour un système de décodage à base de triplets

Le choix du triplet phonétique répond aussi aux exigences suivantes : représenter et utiliser les connaissances acoustico-phonétiques d'une manière aussi souple et efficace que possible au cours du décodage. À ce titre l'avantage du triplet est d'être une unité d'information complète qui se prête bien aux techniques de l'intelligence artificielle comme nous allons le voir maintenant.

On peut déduire à partir de deux triplets de la base de connaissances un certain nombre de relations portant sur les positions relatives des formants décrivant ces triplets. Lors du décodage il est donc possible de vérifier que ces relations sont satisfaites entre deux instances de ces mêmes triplets, même s'ils sont très éloignés dans le temps. Pour assurer que le résultat du décodage est cohérent au niveau de la phrase, et donc que les relations précédentes sont satisfaites, nous avons utilisé un algorithme de relaxation flou [Mohr 86] qui élimine les étiquettes incohérentes du treillis phonétique. Avec des paramètres de relaxation appropriés 70% des étiquettes incorrectes sont éliminées. Ces premiers résultats sont encourageants mais montrent qu'il faut améliorer la construction des relations acoustiques

(jouant le rôle de contraintes pour la relaxation) car certaines d'étiquettes correctes sont éliminées (environ 30%).

4 Perspectives

L'utilisation de triplets en décodage acoustico-phonétique soulève un certain nombre de sujets de recherche, parmi lesquels :

- **La modélisation des déformations que peut subir un triplet** Le choix du triplet comme unité phonétique repose sur l'idée que le triplet permet de « capturer » une bonne partie des effets de coarticulation et donc que sa réalisation acoustique varie peu. Malgré tout, il reste que le triplet est soumis à un certain nombre d'influences qui peuvent altérer sa réalisation. Certains effets de coarticulation ne peuvent pas être pris en compte, c'est le cas des séquences V_1CV_2 avec un consonne labiale où la première voyelle peut influencer la seconde. Par ailleurs, nous ignorons encore les effets de l'accentuation et de la position du triplet dans le mot et dans la phrase.
- **Détection des événements acoustiques** C'est sans doute le point qui, à l'heure actuelle, limite le plus les systèmes de décodage à base de connaissances. Le suivi de formant automatique est un problème classique pour lequel de nombreuses solutions ont été proposées. En revanche la détection des barres d'explosion et la détermination de ses caractéristiques est un problème peu souvent abordé et pour lequel il reste donc de nombreux progrès à accomplir. L'état des recherches est encore moins avancé en ce qui concerne l'étude fine des bruits de friction (comment trouver, par exemple, la limite inférieure de bruit en tenant compte des formants de bruit ?) et ce point nécessite donc encore de substantiels efforts.

- **combinaison des résultats partiels** Lors du décodage il faut combiner un certain nombre de résultats de natures différentes : segmentation, événements et indices acoustiques, déductions phonétiques, déductions lexicales... Traditionnellement on attache à chacun de ces résultats un score indiquant la confiance qu'on lui accorde. Les scores sont ensuite judicieusement combinés pour évaluer le niveau de confiance d'une déduction obtenue à partir de résultats partiels. Cependant ce mécanisme de contrôle ne permet pas de connaître l'ensemble des faits de base et des déductions partielles qui ont contribué à une déduction. Il est donc possible qu'il existe des incohérences entre plusieurs de ces faits et déductions : (i) un ou plusieurs événements acoustiques ont été interprétés de manière contradictoire (ii) plusieurs faits ou déductions partielles apparaissant à des niveaux d'analyse différents (segmentation, formants, indices acoustiques...) peuvent être incompatibles à la lumière de nos connaissances des phénomènes de la parole. Le raisonnement hypothétique [de Kleer 86] qui permet d'associer à une déduction l'ensemble des faits qui la sous-tendent et d'assurer qu'ils sont compatibles entre eux, représente une voie de recherches intéressante pour la parole. Dans ce cadre, il faut aborder deux points clés : d'une part, comment mettre en évidence les incohérences parmi un ensemble d'événements acoustiques et de déductions partielles ; d'autre part, comment traiter le cas où une incohérence apparaît?

De nombreux efforts restent donc à fournir dans le cadre d'une approche à base de triplets du décodage acoustico-phonétique, tant au niveau de la détection fine des événements acoustiques, qu'au niveau des techniques mises en œuvre pour utiliser le plus efficacement possible les connaissances acoustiques et phonétiques.

Références

- [de Kleer 86] J. de Kleer. An assumption-based TMS. *Artificial Intelligence*, (28):127-162, 1986.
- [Fohr 89] D. Fohr, N. Carbonell, and J.P. Haton. Phonetic decoding of continuous speech with the aphodex expert system. In *Proceedings of European Conference on Speech Technology*, pages 609-612, Paris, France, September, 1989.
- [Laprie 90] Y. Laprie. Optimum spectral peak track interpretation in terms of formants. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 1261-1264, Kobe, Japan, November, 1990.
- [Mohr 86] R. Mohr and C. Henderson. Arc and path consistency revisited. *Artificial Intelligence*, (28):225-233, 1986.

WORKSTATION FOR SPEECH ANALYSIS

S. Andreyev and V. Chuchupal

Computer Centre of the USSR Academy of Sciences, Moscow

ABSTRACT

SLIRE-3 (Speech and Language Interactive Research Environment) is the workstation software for IBM PC compatible computers fitted for research in such areas as speech recognition and synthesis, phonetics, criminology and so on. The system was developed in Computer Centre of the USSR Academy of Science during the last 3 years. The main features of this system is its openness and the possibility for phonetists and linguists to organize data bases on it.

1. INTRODUCTION

Several years of work in the area of speech analysis gave us an understanding of the features of the workstation which we need for our investigations [1,2]:

- visualization of the main speech signal forms; the multy window system with windows easily modified for specific applications;
- realization of the main speech signal analysis algorithms;
- system openness; new algorithms can be easily added to the system;
- possibility to organize data bases;
- realization of statistical analysis procedures;
- system friendliness and comfortable interface.

We realized our project on IBM PC compatible computer. We don't think that it is a very suitable for speech analysis computer and we know that there'll be some requirements which we'll not have a chance

to fulfil on such a computer, but today it is the most popular computer in phonetic and linguistic laboratories of USSR.

The most valuable requirement which, we think, can't be realized on IBM PC compatible computers is the real time performance. The most prolonged procedure in speech analysis is spectra calculation and everything based on it, for example, sonogram calculation. For spectra calculation we use different FFT algorithms: Vinograd algorithm (slightly modified by ourselves) and Walch transformation. Certainly, they can't be used for real-time analysis, but they are suitable for ordinary speech analysis with data base.

Though our system can be used in different areas it is oriented on phonetic and linguistic analysis, especially in the part connected with data bases.

The most part of speech analysis is impossible without the data base, because:

- a lot of files (signals) are used;
- users may need the segmentation of those signals;
- different projects may use different segmentation of the same files;
- a lot of researchers use the statistical analysis, which can be done only on stored data;
- users may need an instrument for quick search through a lot of accessible data.

2. INTERFACE

Specialists in speech analysis rarely are good computer specialists that's why the workstation which is difficult to use will not be used at all.

The same result will be with the workstation easy to use but primitive in availability. Our system is menu driven, sends different messages, includes brief and full HELPs. The user will open new possibilities during work sessions and he must not remember a lot of information at the beginning of his work.

3. PROGRAMMABILITY

We understand perfectly that though we tried our best to fulfil the maximum of users requirements nearly every researcher wants to add his own algorithm to the system. The only possibility for researcher to add his own algorithm to the system is to become a programmer for some time and to prepare only his algorithm. The visualization of the results will be done automatically.

4. SIGNAL FORMS

SLIRE-3 can display speech in traditional forms: different waveforms, spectrograms, sonogram and diagrams of parameters (zero-crossing, energy and so on) together with timing and segment marks (from database). Users can easily switch analysis format to choose the most suitable for their task. User can work with the signals of any duration (from very short to very long); different types of analysis filters can be selected for better time and frequency resolution. Any part of any signal can be cut out and deleted or added to another signal and in such a way the new artificial signals can be obtained.

Figures 1, 2 show examples of different SLIRE-3 layouts, composed by different users for their specific

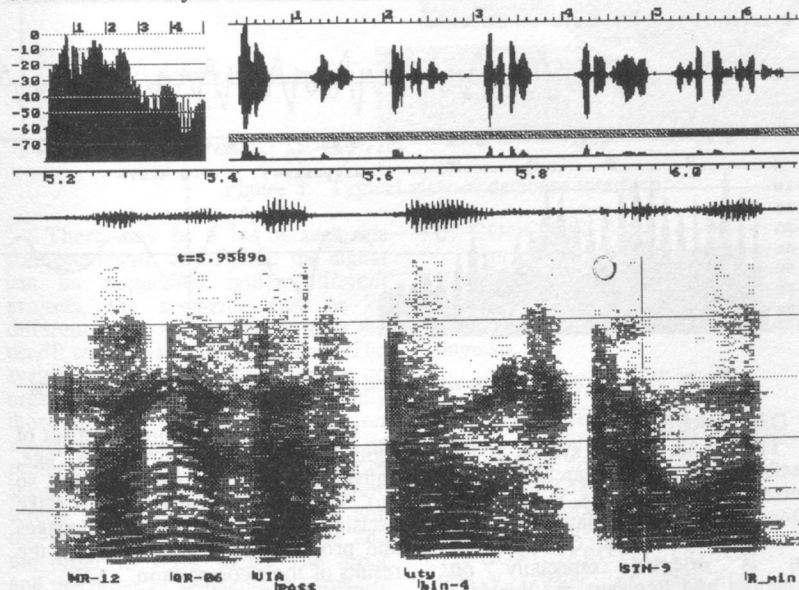


Figure 1: A layout with entire signal and sonogram of the signal part.

research needs. Any part of a signal can be shown in most suitable view from extension to high degree compression and in big amplitude scale diapason. Each waveform type is shown in different window.

The direct control over all display parameters of all windows is very simple and the disposition of the signals on the screen can be changed in zero time. Any window can be shown with (or without) the time scale and the user can choose the

number and type of characters to show in each window. The different windows are not time-synchronised; it allows to compare on the screen different parts of the signal. The synchronization is done at the window change moment.

The system disposes of interface convenient for the user, and all the analysis parameters as well as signal

processing and parameters of visualization (near 160) including colour palette option for every window, can be adjusted by any user. All these parameters are easily changed by user during the session and automatically stored in configuration file after the end and used by system for next sessions. The system is bilingual (all messages are in Russian or English).

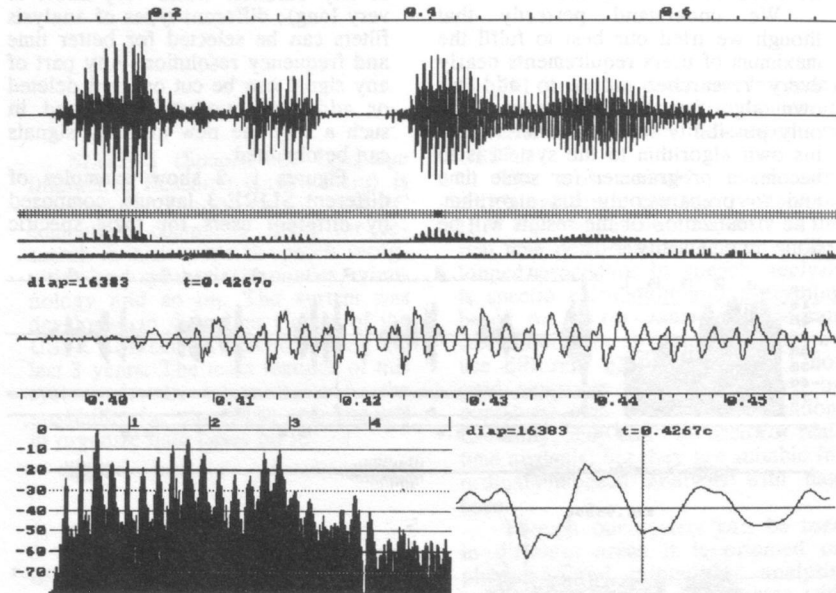


Figure 2: A layout of entire signal and different views of different parts.

5. DATABASE

The kernel of our system is the possibility for any researcher to organize speech data bases, to store and change data in those data bases and to analyze these data. Our workstation is oriented especially on phonetists and linguists; the interface of the system was organized in such a way that such specialists can easily work with it.

The users of our system can create new data bases and change any information which was included earlier. Data bases will contain information (objects) of 4 different types: DICTORS, PROJECTS, SIGNALS and SEGMENTS.

Object DICTOR contains a lot of signs (for example: birthplace, age, native language, education and so on). This information allows to make decisions about different languages and pronunciations and to obtain the results of their comparison.

PROJECTS allow to organize the different segmentations of the same signals by several researchers under the different rules and algorithms and those results will not interfere with one another during the statistical analysis.

Each object of SIGNAL type corresponds to the real signal (not file itself but only its special descriptor). It includes some information about

signal (file's name and type, pronounced text, phonetic and linguistic notation and auxiliary information). The signals may be of two different types: initial (obtained by A/D) or artificial (obtained as the combination of parts of different initial signals). This sign is very important not to distort the results of statistical analysis. Every initial signal has a reference to the object DICTOR of this signal.

Object SEGMENT corresponds to every segment marked on the signal. All the statistical analysis is

based on the segment's parameters (time, duration, type and an information to visualize with the mark of this segment). Each user can define his own types of segments or use common definitions for a group of users. There can be different segment types for sentences, words, phonemes, letters, sounds, types of sounds and so on. If the user wants to visualize the segmentation then each segment will be marked and auxiliary information connected with this mark will be seen too.

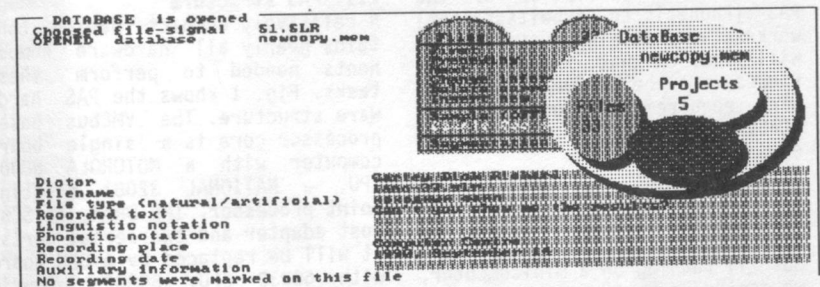


Figure 3: Typical view of database interface.

There may be a lot of segments connected with any signal; the signal can be segmented under different projects, the segments may be of different types and not all of them the user'll want to see simultaneously. The system has the simple interface to organize the visualization of only those groups of segments the user wants to see and in such a way he prefer to see them. Different types of segments can be shown on different screen levels one under another and the user only chooses the types and the order to show them. The order and the types of segments to show are choosed for each signal form independently so the user can organize it for example in such a way as to see words and accents under the full signal view, marks of vowels under sonogram and so on.

6. CONCLUSION

An interactive research environment was designed for speech signals analysis. It includes the possibility to organize data bases for phonetists and linguists.

7. REFERENCES

- [1] ANDREYEV, S., CHUCHUPAL, V. (1990), "SLIRE-2 - an interactive system for speech signal analysis on IBM PC", Computer Centre of USSR Academy of Sciences, Moscow (in Russian)
- [2] GONCHAROV, S., CHUCHUPAL, V. (1987), "An interactive software for speech signal analysis", XI-th ICPHS Proceedings, v.5, 63-67 (in Russian)

WORKSTATION AND SIGNAL PROCESSING SOFTWARE FOR EXPERIMENTAL PHONETICS

M. Scheffers and W. Thon

Institut für Phonetik und
digitale Sprachverarbeitung (*ipds*)
Kiel, Germany

ABSTRACT

We present a description of the PAS (Phonetische Arbeitsstation) workstation, developed at *ipds*. We will focus on what, from our experience, are the hardware and software requirements for a workstation for phonetics research and education.

1. INTRODUCTION

Based on the experience with the speech signal processor software SSP [1] running on a minicomputer, we decided some years ago to implement a network of PAS (Phonetische Arbeitsstationen) workstations with an advanced speech signal processor (ASSP) package optimized for phonetics research and education. While this configuration provides standard functions (signal acquisition, display and analysis) nowadays implemented in commercially available products, it offers additional facilities that cannot easily be obtained:

- Manipulation of signal and parameter files.
- Computer-controlled listening experiments and measuring of reaction times.
- Variability in processing parameters for different signals and for educational purposes.
- Support for novice users.

Though relying upon hardware and software standards (VMEbus, 'C' programming language) for easy upgrading, the software system as a whole is not easily portable to different workstations.

2. HARDWARE

2.1. PAS structure

A relatively compact device contains nearly all hardware components needed to perform these tasks. Fig. 1 shows the PAS hardware structure. The VMEbus based processor core is a single board computer with a MOTOROLA 68000 CPU, a NATIONAL 32081 floating point processor, 1MB DRAM, an SCSI host adapter and two serial ports. It will be replaced by a CPU board with 68030 CPU, 68882 numeric coprocessor, 4MB DRAM and the same peripheral controllers. A separate VMEbus memory board provides addi-

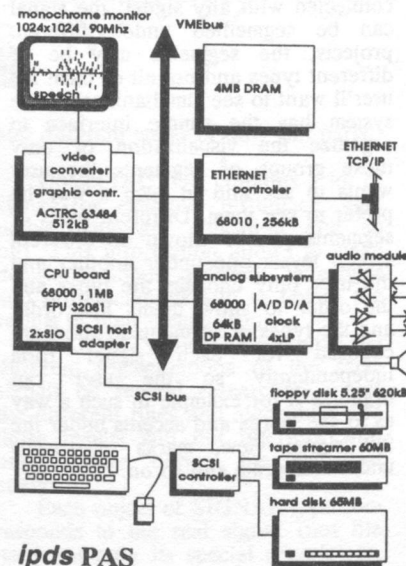


Fig. 1 PAS hardware structure

tional 4MB DRAM. The SCSI bus controls a 5.25" floppy disk drive, a 60MB QIC02 cartridge tape streamer and a 65MB hard disk. One serial port is used for the keyboard, the second one is available for a printer. A mouse or a trackball may be connected to the keyboard.

A graphic controller with an ACRTC HITACHI 63484 and 512kB video frame buffer provides graphic and alphanumeric display. An *ipds* designed high resolution converter generates a monochrome video signal with 1024 x 1024 pixels and 90MHz video dot clock, which is displayed on a 15" monitor.

The analog subsystem is a VMEbus board made up of two parts: the industrially available part contains a separate 68000 CPU with 64kB of 'dual ported' RAM, which is part of the memory map of the main CPU. The *ipds* designed part contains two 12 bit A/D-converters, two 12 bit D/A-converters, four low pass filters with software controlled cut-off frequencies and a software controlled sample clock generator common to all converters. The subsystem processor firmware developed at *ipds* allows for setting of sample rate and low pass cut-off frequency and for continuous conversion and data transfer from or to the hard disk for two channels with a sample rate of 20kHz max., each. These may be two input channels, two output channels or one input and one output channel.

The *ipds* designed audio module provides a user frontend to the analog subsystem. It contains all further analog components like amplifiers, a monitor loudspeaker and an operation front panel with channel switch, amplitude controls, input/output jacks. A digital operation state display is controlled by the analog subsystem. Hard- and software of the PAS analog interface provide features similar to a stereo tape recorder.

2.2. Networking

The six PASs implemented now are connected to the ETHERNET LAN of the *ipds* covering the complete premises of the department. The PAS's ETHERNET controller features a separate 68010 processor and 256kB of buffer memory. The networking protocol TCP/IP implemented as onboard firmware allows for easy transfer of files between the PASs and other nodes of the LAN even with different operating systems. At present, besides the six PASs running OS9/68k as operating system, an APOLLO DN3500 (UNIX SYSTEM V and BSD 4.3) and two AT compatibles (MS-DOS) are connected.

2.3. Supplements and expansions

For listening tests, an *ipds* designed device to acquire reactions and reaction times from up to 16 listeners simultaneously may be connected to each PAS. During a listening test a number of systematically varied speech signals are offered to the listeners through the PAS's analog interface. The subjects are asked to make a decision by pressing one of several buttons. The reaction time device, controlled by the PAS, starts a time measurement with reference to the speech signal at times specified by the experimenter. It monitors the keys of any listener up to a specified maximum time. At the first reaction of a listener, time and decision are stored. The data collected are transferred to the PAS for further processing at the end of the test.

The PAS is designed to provide the basic computing capabilities for phonetic speech signal processing. The widespread networking protocol TCP/IP allows for incorporation of virtually any computer designed for special signal processing tasks into the network, thus making its computing capability available to the users of the PAS.

3. SOFTWARE

3.1. Program structure

The software package ASSP has been developed to give both experienced and inexperienced users easy access to the processing facilities described in the introduction. The package consists of a main program and several sub-programs. Since, in our experience, graphical display and interactive manipulation of signals play a central role in nearly every task, these functions are included in the main program. The subprograms perform functions such as analysis, synthesis, analog I/O, etc.

The main program is characterized by three levels: A graphical, menu-driven user-interface, an interpreter for the problem-oriented command language ELK (Easy Language Kit), and a toolbox with general file-handling and signal processing routines. Menu items can be selected using the mouse or the cursor keys and by single keystrokes. Some frequently used functions such as setting and deleting time markers and acoustic output of (parts of) displayed speech signals are directly accessible via function keys. The dialogue with the user leads to the generation of ELK commands such as 'DISPLAY 3,(0.1,2.3)' (display the selected parameters of all files allocated to window 3 in the time range from 0.1 to 2.3 seconds). The command line interpreter either executes such a command itself, using routines from the toolbox, calls a subprogram to do so, or issues a system command. Macros may be defined e.g., to design a series of manipulations with identical structure. For this the ELK language provides predefined variables, jump labels, and IF and GOTO statements to build loops.

The sub-programs have a standardized user-interface, very similar to the one in the main program.

They can, however, also be called with an argument list. If all arguments are specified, the user-interface is not invoked so that these programs can also run as a background process. It may be clear that the sub-programs can also be run without starting the main program. This implies that new processing may be designed and tested outside the package. The data structure, the toolbox, and the standard user-interface greatly reduce the overhead and permit accessing existing data files and creating new ones that can be handled by the package. The modular structure of the package facilitates inclusion of new functions.

3.2. Data structure

Data files are grouped in so-called AREAs (subdirectories of the user's directory ASSP). Each AREA contains a configuration file with processing parameters such as sampling frequency, LPC order, analysis frame size and shift, etc. This configuration concerns all data files in the AREA, thus relieving the user of repeatedly having to specify or acknowledge them. A user may define and store several configurations optimized for the signals he is working with (e.g. physiological signals rather than speech) and select one of these when creating a new AREA instead of using the default settings. Furthermore, each data file contains a header, specifying the data type and all basic information for handling and display. Thus file handling and display routines in the toolbox could be made very general and small in number. It also means that introduction of a new data type generally requires modification of only one routine, viz. the one for creating a data file.

3.3. Summary of features

The package provides standard LPC analysis (autocorrelation method, reflection coefficients), formant

analysis based on root-solving of the LPC polynomial, and F0-analysis. Intermediate analysis results, such as autocorrelation coefficients, may be stored for educational purposes. As an aid for segmentation/labeling and for loudness manipulation, an energy analysis (short-term rms) can be performed. For manipulation of speech rate during synthesis, a file may be created and filled with the standard frame duration values. Routines for converting LPC parameters (e.g. cepstral coefficients to area functions) are available.

The synthesis program requires definition of an F0 and a filter file. Using the information stored in the file header, the program will automatically convert the filter parameters if necessary and select the appropriate synthesis routine. Optionally, an energy and a frame duration file (see above) may be specified. Synthesis results may be written to a new file or appended to an existing one.

All data files can in principle be displayed. Both time and y-scales may freely be adjusted by the user. Up to 10 graphical windows can be defined to which up to 64 files can be allocated. If more than one signal is displayed in a window, a vertical shift may be defined to ease comparison. Contours may be plotted on a linear or logarithmic y-axis. For multi-parameter files such as formants, a selection can be made on a subset to be displayed. FFT and LPC spectra may be blended in. Per default, the windows have a common time axis, meaning that if the time range in one window is changed, the other windows will automatically be redisplayed with this new range. It is, however, possible to decouple a window from this common axis to provide a zoom or an overview window or to compare signals in different time ranges.

Interactive manipulation on displayed signals include CUT, COPY, APPEND, and INSERT of signals, MULTIPLY, DIVIDE, ADD, and SUBTRACT of constant values, SETTING single values, DRAWING contours, SMOOTHING, and INTERPOLATION. File-to-file manipulations include COPY, INSERT, APPEND, ADD, and SUBTRACT of signals, MAPPING (selective copying to adjust time range), and MASKING (transferring voiced/unvoiced information). Processing such as filtering, up/down sampling, and automatic stylization of contours is also provided. All manipulations are available for all data files even if they make little sense. Manipulations may be performed on all parameters or on a subset.

Whereas for analog input the duration of the recording is limited by the available disk space (typically 15 min. at 16 kHz), this was deemed undesirable for acoustic output. Since a sequence of output calls yields pauses between the stimuli that cannot be controlled very well and disrupt reaction time measurement, a batch list option has been developed. A batch list contains the sequence of segments, pauses, and marking signals that should be output. In the output program, a pre-processor will check the list, open the data files that contain the segments, generate the marking signals, and create a local command list for the analog sub-system. The firmware of this system can process this command list without creating gaps while simultaneously keeping the reaction time system synchronized. This means that there is virtually no limit on the duration of acoustic output.

4. REFERENCES

- [1] BARRY, W. and KOHLER, K. (eds) (1982), "Phonetic data processing at Kiel University", Arbeitsberichte des Instituts für Phonetik, Universität Kiel (AIPUK), 18.

a pas de multiples états, donc pas de risque de désynchronisation entre hôte et interface.

3. interpréteur de commande

S'il n'y a pas d'états, il n'y a pas non plus besoin d'un processus serveur tournant en permanence sur l'hôte. En pratique les requêtes de l'interface sont servies dans le cadre de l'interpréteur de commande habituel de l'hôte ("shell", interpréteur lisp, etc.). Un avantage est que l'interface de commandes de l'hôte demeure accessible à l'utilisateur (à travers une fenêtre de terminal incorporée à MapSignal). Un autre est que le protocole est robuste et facile à déboguer.

4. format des échanges

Les échanges d'information se font à l'aide de caractères ascii affichables (pas de caractères de contrôle, pas de bit de parité). Les requêtes et réponses utilisent un format lisible par l'homme (nombres en ascii), sauf pour les données graphiques.

5. code graphique "min-max"

Un signal est affiché sur l'écran du Mac sous la forme de son *enveloppe*. Celle-ci se matérialise graphiquement par un certain nombre de traits verticaux, à la résolution de l'écran. Ces traits sont définis par les ordonnées de leurs extrémités, qui dépendent simplement des *min* et *max* de groupes de *k* échantillons de signal (*k* est fonction de l'échelle horizontale):

$$a_n = \max_{i \in [kn, k(n+1)]} S_i$$

$$b_n = \min_{i \in [kn, k(n+1)]} S_i$$

Un algorithme spécial permet des échelles fractionnaires. Pour assurer la continuité visuelle (rompue lorsque $a_n < b_{n+1}$ ou $b_n > a_{n+1}$) ces valeurs sont remplacées avant affichage par:

$$A_n = \max\left(a_n, \frac{a_n + b_{n-1}}{2}, \frac{a_n + b_{n+1}}{2}\right)$$

$$B_n = \min\left(b_n, \frac{b_n + a_{n-1}}{2}, \frac{b_n + a_{n+1}}{2}\right)$$

Ces données graphiques sont calculées sur la machine hôte, codées sur 4 octets et envoyées à l'interface utilisateur qui les décode et s'en sert pour tracer l'enveloppe du signal. Le résultat graphique est identique à celui qui serait obtenu si tous les échantillons étaient dessinés, mais le volume de données à transmettre est beaucoup plus faible.

Ainsi, ce ne sont pas les données numériques qui transitent entre l'hôte et l'interface, mais seulement leur *représentation graphique*. Cela permet à MapSignal d'offrir un temps de réponse tout à fait honorable malgré la capacité limitée d'une liaison série. La gestion des fenêtres (réaffichage, etc.) est purement locale et ne nécessite aucun transfert de données.

3.3. Mécanisme d'extension

MapSignal n'offre en propre que des fonctions de visualisation, mais il possède un mécanisme qui permet de d'incorporer de nouvelles fonctions. Un "éditeur de commandes" permet de constituer un menu de commandes. A chacune d'elles correspond une chaîne de caractères qui est envoyée à l'hôte lorsque la commande est choisie dans le menu. La chaîne de caractères peut comporter des *mots-clé* qui sont interprétés au moment de l'envoi, ce qui permet d'inclure dans la chaîne des informations dynamiques telles que les coordonnées du curseur, le nom du fichier signal affiché, etc.. L'utilisateur peut ainsi diriger interactivement l'exécution de programmes disponibles sur l'hôte, en particulier ceux qu'il aura réalisés lui-même. Puisque l'affichage et l'interaction sont pris en charge par MapSignal, ces programmes peuvent être simples et portables.

3.4. Mécanisme d'interrogation de paramètres

Avant l'affichage d'un signal, MapSignal interroge l'hôte pour obtenir des informations sur ce signal (nombre d'échantillons, etc.). L'hôte envoie les informations qu'il possède: celles qui manquent sont remplacées par des valeurs par défaut. L'hôte a aussi le choix de la méthode par laquelle il obtient une information. Par exemple, la taille en échantillons du fichier signal peut provenir d'un "header", du système d'exploitation, ou d'un programme qui mesure explicitement la taille du fichier. MapSignal peut ainsi accommoder des formats de données très divers.

3.5. Moniteur d'acquisition

Une fonction originale mérite d'être signalée. Il s'agit d'une *fenêtre-moniteur d'acquisition de données analogiques*. Associée à un programme d'acquisition continue sur l'hôte, qui utilise un *buffer*

circulaire (buffer contenant en permanence les *n* dernières secondes de données), elle affiche son contenu. Ce contenu est en constant renouvellement au rythme de l'acquisition de nouvelles données, ce qui se reflète par l'image du signal qui "flotte" de droite à gauche à travers l'écran. L'utilisateur peut ainsi "geler" l'acquisition au moment opportun. On évite ainsi les problèmes de "timing" dus au mauvais alignement temporel, ou les distorsions dues au mauvais réglage du niveau.

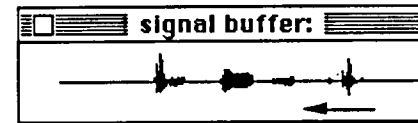


Fig. 3. Moniteur d'acquisition montrant le contenu d'un buffer circulaire.

L'acquisition continue est plus conviviale pour un locuteur que celle, discontinue, qu'imposent les solutions classiques.

3.6. Implémentation

MapSignal est écrit en C sur le Macintosh. Il nécessite deux commandes simples sur l'hôte, écrites elles aussi en C. Le programme est disponible gratuitement à fins de recherche. La distribution comprend le programme Macintosh, les sources des commandes hôte en C, et la documentation.

4. COMPARAISON AVEC X-11

L'idée principale défendue dans cet article est qu'on a intérêt à dissocier interface-utilisateur et traitement (calcul, entrées-sorties) et à les attribuer à des machines différentes. Cette idée sous-tend celle de "frontal" utilisée pour les bases de données ou le calcul scientifique (Mathematica). Elle se retrouve aussi dans X-windows, qui permet à un programme tournant sur une machine d'interagir indifféremment avec toute console du réseau. X-windows est de plus en plus répandu, mais il souffre des inconvénients suivants:

a) On ne peut associer que des machines possédant une version compatible de X-windows. Une mise à niveau partielle peut tout désorganiser, par ailleurs certaines machines de calcul n'offrent pas X. On perd ainsi dans la pratique une part

de la liberté de choix qu'est censée apporter la dissociation des fonctions.

b) X-windows est gourmand en ressources (mémoire, disque).

c) X-windows ne permet n'offre pas un environnement logiciel aussi complet que ceux disponibles sur les micro-ordinateurs (cette situation change avec l'apparition des émulateurs X pouvant s'intégrer dans l'environnement logiciel d'un micro).

En comparaison, le modèle de communication sur lequel est fondé MapSignal pose peu de conditions quant à la machine hôte ou la nature de la liaison. Il faut cependant reconnaître que les contraintes de protocole que sont la relation maître-esclave stricte et l'absence d'états impliquent une communication moins riche que celle offerte par X-windows.

CONCLUSION

Le problème de la portabilité des logiciels de manipulation du signal peut trouver un élément de réponse dans la dissociation des fonctions de traitement du signal et d'interface utilisateur entre machines distinctes. La communication entre machines que nécessite cette solution peut se faire de façon simple et robuste dans le contexte de l'interpréteur de commande (shell) de la machine hôte. Conçu selon ces principes, le logiciel MapSignal permet la visualisation et la manipulation de données de signal se trouvant sur des machines très diverses. Sa conception fait appel à des idées originales qui peuvent être utiles dans d'autres contextes.

REMERCIEMENTS

Une partie de ce travail à été effectué lorsque l'auteur était à *ATR Interpreting Telephony Research Laboratories* dans le cadre d'une bourse du programme STP des Communautés Européennes. Il tient à remercier ATR de son hospitalité.

BIBLIOGRAPHIE

- [1] de Cheveigné, A., M. Abe and S. Doshita (1985), "The human interface of a speech workstation", *Studia Phonologica. XIX*, 18-26.
- [2] de Cheveigné, A. (1989), "The MapSignal remote speech editor", *Proc. Acoust. Soc. Japan Autumn meeting*, 383-384.
- [3] de Cheveigné, A. (1990), "The MapSignal remote speech editor", ATR Technical report TR-1-0137.

SPEECH SYNTHESIS COMPUTER FOR THE BLIND

G.LOSIK

INSTITUTE OF TECHNICAL CYBERNETICS, MINSK, USSR

ABSTRACT

The invention of speech synthesizers have made revolution in the computerization of the blind's life. But these people can use special computers only, for serial computers are rarely equipped with synthesizers whereas displays are obligatory. Speech synthesizer that does not speak Russian cannot be reconstructed to be used by the Russian speaking blind. We have made a computer for the blind with the speech synthesizer "PhonemophoneWS", that is speaking Russian. It is being produced in small series.

1. Its hardware represents a printed circuit board for IBM PC. Its software - a 64 kbyte control file and a 10 kb. lexical dictionary file. Synthesizer's driver can be loaded automatically or from keyboard. The text for the sound-track in Russian can be taken from the file, from the user's programme, from display, from keyboard.

2. Synthesizer is able to carry out the following functions:
- synthesize oral speech on arbitrary text in Russian;
- oral speech can be synthesized by a female voice or by one out of two male ones;

- modulate the phrase's intonation of the type of question, exclamation, narration, enumeration;
- stress the words automatically;
- provide for about 99% understanding of words in speech;
- provide for the sound-track of numbers from 0 to one milliard;
- provide for the sound-track of the signs like * # @ \$ & * () [] = / \ " - ' [] _ , articulate in full voice abridgements in the text (like etc., i.e., no, tu,), abbreviations (like USSR, IBM, USA, UK, ICPHS).

3. The work of a blind with a computer in most cases implies the text revision. That is why the speech synthesizer is additionally equipped with the form of the programme CONVEYOR of multistage transformation of the Russian orthographic text into orthoepic text. The whole conveyor is called READER. It consists of the following successive blocks:
- the block of text's segmentation into sentences;
- the block of words' stressing according to a special 100000 words Russian language dictionary of stresses;
- the block of substitution of abbreviations for full-toned words (for example: etc. etcetera) according to the 1000 words Russian language dictio-

nary of abbreviations;
- the block of substitution of orthographic abbreviations for orthoepic words;
- the block of substitution of signs of the type * # @ \$ for the words corresponding to their names;
- the block of substitution of numbers from 0 to one milliard for words allowing to pronounce these numbers;
- the block used to delete the so-called "text rubbish", for example: unnecessary blanks, dot doubles, exclamation and question marks' doubles;
- the block to provide sound-tracks of English language key words according to 200 key words dictionary (BASIC).

4. When working with the text for the blind, the modes of asking again about what had been said and about detailed elaboration of phrases' sounding are quite common. That is why the speech synthesizer is equipped with MONITOR, which controls a group of keys located to the right on an ordinary IBM keyboard. Monitor's keys realize the following modes:
- return to the listening to a previous phrase (word);
- return to the beginning of the previous paragraph;
- jump forward to every other phrase;
- jump to the beginning of the next paragraph;
- a temporary stop of the listening to the text;
- transition to the detailed listening to the text by words or letters;
- substitution of one diction voice for another male voice or female one;
- change of synthesized loudness and tempo speech.
5. The synthesizer connected to the computer via CONVEYOR & MONITOR allowed to produce for a blind the

following special programmes meeting different requirements these people were examined of:
- "SOUNDING KEYBOARD" provides for the sound-track of all keyboard's alphanumeric and service keys. Capital letters are rendered into a sound-track by one voice, ordinary letters - by another, Latin ones - by the third. A key can be heard after it had been pressed or after the end of the word;
- "READING OF THE TEXT FILES" is carried out after a blind man's answer to the question concerning what file he wants to listen to using monitor;
- "THE CLOCK" reports time, date and day when called;
- "CALCULATOR", when called, allows a blind to perform computations with six-digit numbers using subtraction, division and multiplication;
- "NOTR-BOOK" helps a blind to record on the disk from the keyboard short text notes for future reproduction of them and listening to them in own interests;
- "TYPEWRITER" provides a blind with the possibility to type quality and long text in upper and lower case letters using the format common for the sighted;
- "ENGLISH-RUSSIAN DICTIONARY" allows to learn oral translation of the English word entered from the keyboard;
- "PSYCHOLOGICAL TESTS" are entered from the keyboard via synthesizer in the form of oral questions the one being tested answers via keyboard. With the help of the key computer analyses the answers and individually estimates the one being tested on different psychological scales.

All above mentioned special programmes are written in C language in MS DOS for the computers IBM PC AT/XT.

THE AUTOMATIC RUSSIAN TEXT TRANSCRIBER

E.B.Ovcharenko, J.V.Ipatov, S.B.Stepanova

Leningrad State University, USSR

ABSTRACT

This report gives a linguistic description of the program that automatically translates Russian orthographic text into transcription signs. This transcriber provides linguistic material for computer interface and allow complex automatization for linguistic research, i.e. dictionary-making, text analyses, scientific apparatus compiling and besides to serve as a reliable base for different educating programs.

1. INTRODUCTION

The automatic transcriber (AT) is hoped an indispensable component of the Phonetic fund. Any language material (a dictionary or a text) may be transcribed with the help of the AT sufficiently providing the necessary phonetic details.

Traditionally transcription is known as a system of signs and rules of using them for recording speech and its' sound structure. The main aim of a linguist when working out a system of automatic transcription is to determine and algorithmize the rules of correspondence between 33 Russian letters and speech sounds.

2. THEORY OF TRANSCRIPTION

Several automatic transcribers had been worked out by this time. All their differences can be reduced to 4 general items.

1) The choice of phonologic "ideology". There are several tendencies in modern soviet phonologic science which it their approach to the definition of the main minimum sound language unit - phoneme. Thus, one considers a phoneme to be a representative of a morpheme which is constant (eg, the final consonant of a word 'ропод' is /d/ as it is in 'ропода'); and other (Leningrad or Shcherba's phonological school) - to be an independent unit (eg, in the same word 'ропод' the final phoneme is /t/).

2) The form of presenting material. It may be detailed or simplified transcription (from general symbols C and V for consonants and vowels or phonemic transcription to the detailed indication of sound qualities).

3) Difference in choosing the object of transcription, i.e. that one may transcribe separate words, sentences or texts.

4) The choice of the pronunciation variant. For Russian there are two different standard variants - Moscow and Leningrad).

3. PHONEME TRANSCRIBER

The authors of the AT - members of the Department of Phonetics of Leningrad state University - tried to develop and improve all these aspects.

There is a phoneme block of the AT, responsible for forming the phoneme record of speech material according to Shcherba's phonologic theory.

In phoneme transcription the following phenomena of the Russian language system are considered.

1) In non-stressed Russian vocalism practically there are no /o/ and /e/ vowels. Some frequent words coming from foreign languages, where this vowels still remain in non-stressed syllables, are included into a list of exceptions (eg, какао, радио, анданте etc.).

2) Voiced consonants are replaces with voiceless before a pause. (завод - /zavo"t/, мороз - /maro"s/, гроздь - /gro"s't'/ and so on.)

3) Regressive consonant assimilation in feature of voice and its' absence. (трубка - /tru"пка/, сделать - /z'd'e"lat'/ and so on.)

4) Consonant assimilation in feature of softness and hardness. (шесть - /se"s't'/, пенсионер - /p'in's'ian'e"r/ and so on.)

5) Consonant assimilation in the place of forming (сшить - /ššy"t'/, сжать - /žža"t'/, заказчик - /zakka"š:ik/, городской - /garracko"j/ and so on.

6) "Non-pronounced consonants" (честный - /č'e"snnyj/, поздно - /po"zna/, счастливый - /š'č:isl'i"vnyj/ and so on.)

7) "Double consonants".

Formalisation of rules of pronouncing long or short consonants in place of two similar letters is rather complicated and in this case pronunciation depends on the "double" letter position inside the word, on the neighbour letters and on morpheme borders. Special analyses of all Russian words containing double letter combinations helped to find the rules of their transcription (длинный - /dl'i"nnnyj/, but: сделанный - /z'd'e"lannyj/, грамм - /gra"m/, ввод - /vvo"t/ and so on.)

The phoneme transcription is only one of the ways how to present the speech material with the help of the AT.

4. PHONETIC TRANSCRIBER

Any text can be simultaneously presented as a sequence of allo-phones - concrete realizations of phonemes determined by simple rules of correspondence and assimilation of sounds and reduction of unaccented syllables.

Examples: - [z'd'e"lkʌ], изво"зчик - [izvo"š:ik], голова - [gɔ'lvʌ].

There are hundreds of such elements. And at last, the description of speech flood as a sequence of phonetic elements giving quite detailed description of very subtle but important for perception of speech realization differences, eg, modification of vowels after labial consonants /b/, /p/, /v/, /f/: ба"л - [b'a"l], ва"за - [v'a"zʌ]; or after nasal consonants: ма"ма - [mā"mʌ], не"с - [n,ōs], different degrees of reduction of non-stressed vowels depending on their position to the stressed syllable,

different symbols for vowels after voiced and voiceless consonants, after soft and hard consonants. While transcribing consonants the following phenomena are considered: labial character of consonants before [o] and [u] (стык - [s^v t^v u^k], кот - [k^o o^t]), appearance of faucal explosion in combinations дн, тн, бм, пм (дно - [d^{no}"], обман - [o^bmaⁿ"], etc), lateral explosion in combinations тл, дл (подлый - [p^o d^l y^j], etc), devoicing of sonorants in some positions (eg театр - [t['] i a['] t^r], надсмотрщик - [n^o d^s m^o t^r s['] i k^k] etc). In AT desinged for speech synthesis are considered changes of consonants at the end of words: affrication or aspiration of consonants /p/, /p'/, /t/, /t'/, /k/, /k'/ and nasals' implosiveness /m/, /m'/, /n/, /n'/ and so on.

The set of phonetic elements of this unit corresponds to the set of acoustic elements, which is sufficient for the automatic synthesis of distinctive and naturally sounding Russian speech. That's why the number of these elements is rather large - approximately 700 elements. Still, when presented in a graphic form, all the three types of transcription look compact and usual for phonetists and speech scientists. There are two systems (Roman and Cyrillic alphabets) of graphic representation of phonemes and those of the International Phonetic Alphabet.

The described AT allows to transcribe both separately taken words and sentences, in this case it realizes rules of words' bounds, eg, voicing of voiceless consonants having

no pair: мех животного - [m['] e["] z^{vo} tⁿ y^v ^]; connection of nominative words and relying prepositions and particles into one phonetic word, eg, без огня - [b['] i z^Λ gⁿ ' a["]], не знаю - [n['] i zⁿ a["] u] and so on.

5. ORTHOEPIC STANDARD

The suggested transcriber is oriented to the modern literary Russian pronunciation standard. Two variants of orthoepic standard (Moscow and Leningrad) are generally acknowledged. But nowadays there is a definite tendency to eliminating of differences between variants, "to formation of some common pronunciation standard embracing features of both Moscow and Leningrad variants" (Л.А. Вербицкая, "Русская орфоэпия", 1976, p.115). The AT considers all recent orthoepic researches.

6. STATISTICAL PROCESSING

Important advantage of the AT is its' ability to get information on statistical processing of the text, on distribution of letters, phonemes and allophones both in digital and graphic form. The program is written in algorithm language C for personal computers based on 8086-compatible processor.

7. SUMMARY

The AT described in the report worked out on the Shcherba's phonological principles allows to get both phoneme and phonetic text transcription of two degrees of detalization. The AT takes into account sound modification inside a phonetic word and some phenomena taking place at the nominative words connections. The transcription is

based on modern Russian language pronunciation standard. The program allows to make statistic processing of the text.

COUPLAGE ENTRE LE MODELE A DEUX MASSES ET UN MODELE ANALOGUE DU CONDUIT VOCAL A REFLEXION: THEORIE ET IMPLANTATION

Loan TRINH VAN, Bernard GUERIN, Eric CASTELLI

Institut de la Communication Parlée INPG/ENSERG - Université Stendhal
UA n° 368 46, av Félix Viallet 38031 Grenoble cedex FRANCE

ABSTRACT

In this paper we describe the effects of source-tract coupling. The model of the vocal tract used is the acoustic tube model proposed by Kelly and Lochbaum. The vocal tract is excited by a source based on a two-mass model of the vocal folds. In our case, the source is simulated as an elementary tube of the vocal tract. The influence of the source-tract coupling on the glottal flow is compared in three cases : (1) the above mentioned source-tract coupling; (2) the coupling with a model of the input impedance of the vocal tract ; (3) the no coupling case.

1. INTRODUCTION

La source vocale utilisée dans notre synthétiseur est le modèle à deux masses proposé par Ishizaka et Flanagan (1972) et basé sur des équations mécaniques et aérodynamiques. La simulation du conduit vocal, quant à elle, repose sur le calcul de propagation des ondes acoustiques dans une série de N tubes élémentaires (Kelly et Lochbaum, 1962). Il faut donc établir la relation entre les paramètres de la source et les ondes acoustiques qui se propagent dans le conduit vocal. Cette relation permet aussi d'évaluer l'interaction entre source et conduit vocal.

2. THEORIE ET IMPLANTATION

Dans notre simulation, le couplage source-conduit vocal est basé sur le modèle proposé dans [5]. En effet, la source vocale elle-même peut se représenter comme une source de débit d'impédance Z_g . On utilise l'indice $k+1$ pour le premier tube du conduit vocal: l'aire, la pression et le débit à l'entrée de

ce tube seront A_{k+1} , P_{k+1} , U_{k+1} respectivement (fig 1a).

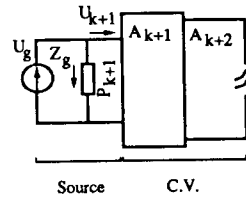


Fig 1a. Modèle de la source vocale chargée par la première section du conduit vocal.

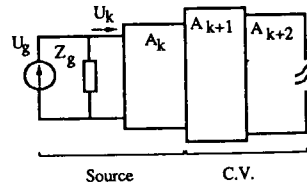


Fig 1b. Source vocale considérée comme un des tubes du conduit vocal.

La figure 1a montre que le débit U_g est la somme du débit qui traverse l'impédance Z_g et de celui qui entre dans le tube $k+1$:

$$(1) U_g = \frac{P_{k+1}}{Z_g} + U_{k+1}$$

Le débit et la pression dans le tube $k+1$ s'écrivent :

$$(2a) U_{k+1}(t) = U_{k+1}^+(t-\tau) - U_{k+1}^-(t+\tau)$$

$$(2b) P_{k+1}(t) =$$

$$\frac{\rho c}{A_{k+1}} (U_{k+1}^+(t-\tau) + U_{k+1}^-(t+\tau))$$

où : ρ est la densité de l'air; c est la célérité du son; τ est le temps de propagation du son dans le tube $k+1$; ($\tau = \ell / c$, les tubes sont de la même longueur ℓ); A_{k+1} est l'aire du tube $k+1$; $U_{k+1}^+(t-\tau)$

et $U_{k+1}^-(t+\tau)$ peuvent être interprétées comme l'onde incidente et l'onde réfléchie dans le tube $k+1$. La source vocale est considérée (fig 1b) comme un tube spécial d'indice k , d'aire A_k et d'impédance Z_k :

$$Z_k = Z_g = \frac{\rho c}{A_k} \text{ avec le coefficient de}$$

$$\text{réflexion : } r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k}$$

Les expressions (1) et (2) nous donnent alors l'équation suivante de l'onde incidente dans le tube $k+1$:

$$(3) U_{k+1}^+(t-\tau) =$$

$$r_k U_{k+1}^-(t+\tau) + U_g \frac{1+r_k}{2}$$

Pour notre implantation, $A_k = A_g$ où A_g est l'aire de la glotte.

Pour valider le couplage ci-dessus, nous avons utilisé comme référence le modèle de l'impédance d'entrée du conduit vocal proposé par Mrayati et Guerin [6]. Dans ce modèle, le conduit vocal est remplacé par deux circuits résonnants aux premier et deuxième formants (fig. 2).

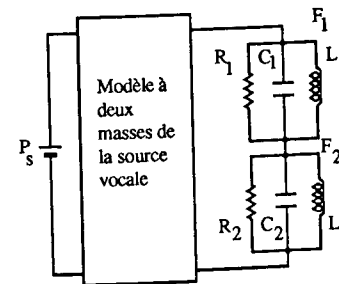


Fig 2. Circuit équivalent à l'impédance d'entrée du conduit vocal couplé au modèle à deux masses de la source vocale.

Dans [6], il a été montré que l'on pouvait se limiter à une représentation de l'impédance d'entrée pour les deux premiers formants.

3. RESULTATS

Pour les voyelles orales françaises étudiées, les paramètres des deux circuits de résonances sont déterminés à partir de la configuration du conduit vocal pour chaque voyelle et des mesures de son impédance d'entrée. Les caractéristiques de l'onde de débit de la source vocale (fig. 3) sont données par les paramètres suivants:

$$+\text{Le quotient d'ouverture : } Q_0 = \frac{T_1 + T_2}{T_0}$$

$$+\text{Le quotient de dissymétrie : } Q_d = \frac{T_1}{T_2}$$

$$+\text{La fréquence fondamentale: } F_0 = \frac{1}{T_0}$$

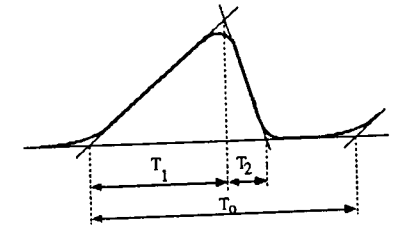


Fig 3. Forme générale de l'onde de débit.

Les paramètres de commande du modèle à deux masses sont la pression subglottique P_s et le coefficient Q représentant la tension et la masse des cordes vocales. Pour nos mesures, P_s varie entre 4 et 16 cmH₂O et $Q = 2$.

Les caractéristiques de l'onde de débit ont été mesurées dans les trois cas (la figure 4 donne les résultats pour 4 voyelles seulement):

- sans couplage;
- avec couplage et le conduit vocal est remplacé par le modèle de l'impédance d'entrée;
- le couplage source-conduit vocal mentionné ci-dessus.

Pour les configurations du conduit vocal correspondantes aux voyelles orales étudiées, les résultats ont montré la concordance des deux cas de couplage. Les évolutions de F_0 , Q_0 et Q_d en fonction de la pression subglottique sont les mêmes dans les cas avec couplage. Il est difficile de donner une loi d'évolution applicable à toutes les voyelles. Pourtant, nous avons remarqué les caractéristiques suivantes:

-pour les voyelles [a], [ɛ], [ə], [o], [ø], [e] et [œ] la fréquence fondamentale augmente légèrement par rapport au cas sans couplage;

-les quotients d'ouverture et de dissymétrie sont modifiés quand il y a l'interaction source-conduit vocal. Quand la source vocale est chargée par le conduit vocal, on constate une déformation de l'onde de débit sous la forme d'une oscillation additionnelle. Cette sur-oscillation a une fréquence proche du premier formant [1]. En conséquence, le quotient de dissymétrie sera fortement modifié et généralement augmenté. En ce qui concerne le quotient d'ouverture, il est diminué dans le cas de couplage source-conduit vocal.

Nous avons comparé la qualité des voyelles orales synthétisées dans les deux cas : sans couplage et avec le couplage décrit ci-dessus. Des tests informels ont montré que le couplage améliore la qualité et le naturel des voyelles orales synthétisées.

3. CONCLUSION

En utilisant le modèle de l'impédance d'entrée du conduit vocal comme référence, nous avons montré que la méthode du couplage source-conduit vocal que nous avons utilisée dans notre

synthétiseur est tout à fait raisonnable. Ce modèle du couplage contribue à améliorer la qualité de sons synthétisés et permet de respecter les effets connus du couplage source-conduit vocal.

REFERENCES

- [1] AL-ANSARI A., GUERIN B. (1981), " Effet du couplage source-conduit vocal sur les caractéristiques de l'onde de débit ", *12èmes Journées d'étude sur la parole*, Montréal (Canada).
- [2] DEGRYSE D. (1981), " Temporal Simulation of Wave Propagation in the Lossy Tract ", *The Fourth F.A.S.E. Symposium*, Avril, 21-24, Venezia (Italy)
- [3] ISHIZAKA K. and FLANAGAN J.L. (1972), " Synthesis of Voiced Sounds from a Two-mass Model of the Vocal Cords ", *B.S.T.J.*, 51, 1233-1268.
- [4] KELLY J.R. and LOCHBAUM C. (1962), " Speech Synthesis ", *Pro. Stockholm - Speech Communication Seminar - R.I.T.* 127-130.
- [5] MARKEL J.D. and A.H. GRAY, JR. (1976), " *Linear Prediction of Speech* ", Springer - Verlag Berlin Heidelberg New York.
- [6] MRAYATI M., GUERIN B. (1976), " Etude de l'impédance d'entrée du conduit vocal, couplage source-conduit vocal ", *Acustica* Vol. 35, No. 5, 330-340.

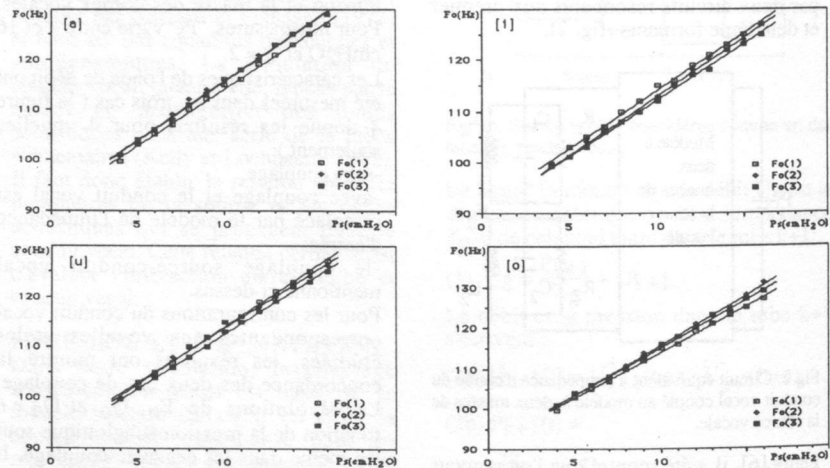


Fig 4a. Relation entre F_0 et P_S pour les voyelles [a], [i], [u] et [o].
 $F_0(1)$: sans couplage, $F_0(2)$: couplage avec circuits résonnants,
 $F_0(3)$: couplage avec le modèle Kelly-Lochbaum.

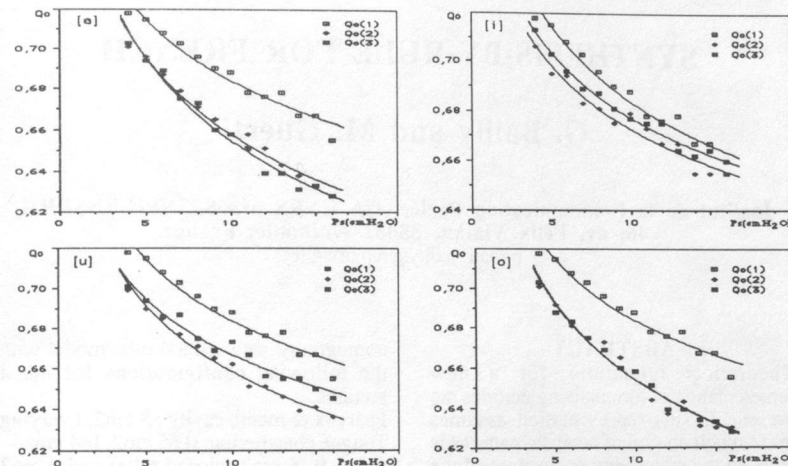


Fig 4b. Variation de Q_0 avec P_S pour les voyelles [a], [i], [u] et [o].
 $Q_0(1)$: sans couplage, $Q_0(2)$: couplage avec circuits résonnants,
 $Q_0(3)$: couplage avec le modèle Kelly-Lochbaum.

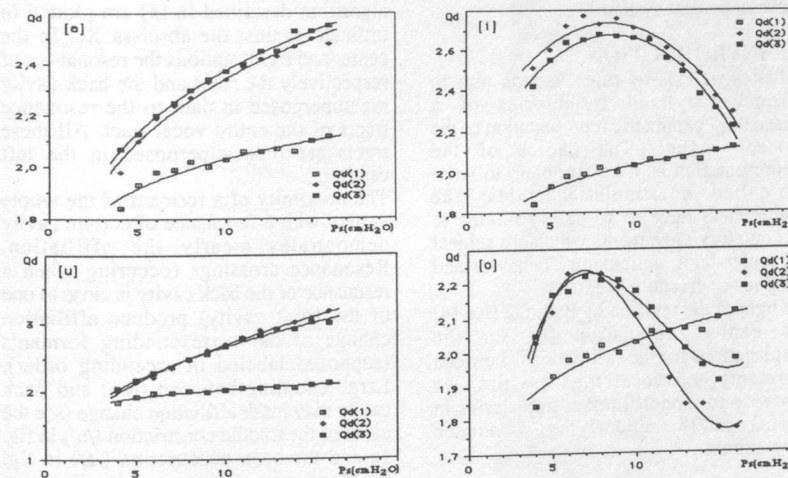


Fig 4c. Evolution de Q_d avec P_S pour les voyelles [a], [i], [u] et [o].
 $Q_d(1)$: sans couplage, $Q_d(2)$: couplage avec circuits résonnants,
 $Q_d(3)$: couplage avec le modèle Kelly-Lochbaum.

SYNTHESIS-BY-RULE FOR FRENCH

G. Bailly and M. Guerti

Institut de la Communication Parlée, UA CNRS n°368. INPG/ENSERG
46, av. Félix Viallet, 38031 Grenoble, France.
e-mail: bailly@icp.imag.fr

ABSTRACT

Theoretical foundations for a new representation of formantic trajectories are presented. This representation assumes that explicit control of acoustic patterns is done on the underlying resonances of the vocal tract. An application of this approach within a synthesis-by-rule system using an extended version of the Klatt synthesizer is described. Current implementation of this application was done using the COMPOST development system.

1. INTRODUCTION

Most synthesis-by-rule systems aim to produce stylized trajectories of a "readable" parametric representation of the speech signal. The choice of the representation is a key-problem to write so-called coarticulation rules: the parameters have to be easily related to articulatory movements which are subject to undershoot, anticipatory behavior and degrees-of-freedom in excess. In light of the revision by Badin & Boe [4] of Fant's explanation [6] for the explanations for the "affiliations" between formants and vocal tract cavities, we propose to model formant trajectories in terms of the underlying resonance trajectories.

2. DEFINITION OF THE VOCAL TRACT RESONANCES

The vocal tract resonances are defined as the union between the individual theoretical resonances of each cavity of the vocal tract: formant trajectories may be deduced from this set of resonances by applying inter-cavity coupling and vocal tract losses.

The figures 1a and 1b sum up the resonance concept thanks revised Fant's

nomograms: we used a 4-tube model with the following configurations for the 4 sections:

Pharynx & mouth cavity : 8 cm², l varying
Tongue constriction: 0.65 cm², l=4 cm

Lips: 0.16 cm² (closed tube) and 4 cm² (open tube), l=1 cm

The constriction center Xc can vary from 2.1 cm to 11.9 cm while keeping vocal tract length constant and equal to 15 cm.

In all captions, the formants detected by an algorithm described in [5] are plotted in ordinate against the abscissa Xc. In the center and right captions the resonances of respectively the front and the back cavity are superposed in dark to the resonance tracts of the entire vocal tract. All these tracts are then superposed in the left caption.

The proximity of a formant of the whole system with a resonance of certain cavity demonstrates clearly the affiliation. Resonance crossings (occurring when a resonance of the back cavity is close to one of the front cavity) produce affiliation change of the corresponding formants (supposed labeled in ascending order). Large coupling between front and back cavity may mask affiliation change (see the cases of the middle constriction (/u/) in fig. 1a and the back constriction (/a/) in fig. 1b. More interesting is the formant convergence between F2-F3 in case of a front constriction (/i/) in fig. 1b: the half wavelength of the back cavity cross the lower resonance of the front cavity (intermediate between a helmholtz resonance and a half wavelength).

We adopt the following notations for designation of resonances : R1, R3 are the two first low resonances of the back cavity and R2, R4 are the first two low resonances of the front cavity.

3. RESONANCE TRIANGLE

A careful examination of vocalic transitions VV uttered by two male speakers for couples of the 10 French oral vowels in light of interpretation of vocalic nomograms in terms of resonances as developed above show an interesting design of the representation space for vowels. Figure 3 presents the R1-R2 plane deduced from the database for one speaker by interpolations on a grid of formant candidates obtained by closed-phase LPC analysis. Figure 2 shows an example of such an interpolation on a /a-i/ vocalic transitions.

A brief comparison between the nomograms of the Fig.1 and the resonance triangle shows the relative adequacy between the theoretical predictions and the natural data: the large dispersion for the F3 values of the back vowels is accounted by the affiliation of this formant with R4 and thus correlated with the lip aperture. On the contrary the F2 of the front open vowels and the F3 of the front closed vowels are affiliated with R3 and thus are not sensible to the lip aperture. It accounts for the non-linearity of the coarticulation characteristics we already published in [7]: /i/ seemed to have the most influence on the successive vowel while bearing no undershoot. In fact target variability for front-open vowels like /i/ and /e/ in French is mostly accounted by F3 (thus R2) and not F2 (thus R3).

4. MODELLING FORMANT TRANSITIONS VIA RESONANCE TRAJECTORIES

We used this resonance representation for modelling formant trajectories: the synthesis-by-rule system we will detail in the following generates resonance trajectories using the COMPOST rule compiler [3]. These resonance trajectories are then converted in formant trajectories using simple equations which capture the essential characteristics of the inter-resonance coupling.

4.1. The COMPOST language

The COMPOST system is a rule-based system for transducing trees: basic atoms on which COMPOST is working are instances of user-defined generic objects. These instances thus inherit of the properties of the class their generic object belongs to: set of features and numerical

values. For example, a declaration of word, syllable and phoneme classes will include for French :

```
class Word
/*Dt,Pp stand for determinant & pers. pronoun*/
object(Noun, Verb, Dt, Pp...) feature(Content)
Noun, Verb are Content; Dt, Pp are -Content;
endclass
class Syllable
object(Syl) feature(acc) Syl are -acc;
endclass
class Phoneme
object(a,i,u,e,y,b,d,g...) feature(voc,nas,liq...)
cue(duration) a,i,u,e,y,b,d,g. have duration=90;
endclass
```

COMPOST then consists of manipulating a complex structure (n-ary tree) whose leaves are instances thanks to an extension of the well-known rewriting rules:
SubTF -> SubTT / SubTL+SubTR;

SubTF is the focus subtree, SubTT is the transformed subtree while SubTL and SubTR are the left and right context subtrees.

The powerful COMPOST subtree matching is labelled with special instructions for local operations :

- memorization of focus instances and/or subtrees in SubTF and their replication in SubTT (thus enabling tree manipulation)
- numerical capabilities: memorization of numerical attributes in SubTF and affectation of complex numerical expression to numerical attributes in SubTT.

For example, a rule for syllabic parsing for French will include:

```
/* create a father node Syllable for any non-liquid consonant followed by a vowel */
regS: [-voc,-liq] -> Syl(#1 / + [voc] ;
```

A COMPOST sketch consists of a set of grammars (each containing a set of ordered rules working on the internal COMPOST tree structure) and external calls. A library of standard routines may be augmented by the user using the COMPOST C-toolkit.

4.2. Modelling trajectories

COMPOST library includes the routine Gentrj which produce frames of parameters according to instructions present in its actual internal tree. It scans for any instance of the class Phoneme and generates the absolute time reference axis according to the actual values of Duration

cue (expressed in ms). The subtree of each Phoneme is then scanned for any instance of the class Target. The first cue of each Target object gives the delay of this target in ms according to the beginning or the end of the father Phoneme (according to the feature \pm Final). The following cues precise the entire set of parameters used by the synthesizer. The targets will be then connected together according to the name of the generic object (splines, straight lines, step functions...).

Like in object-oriented design all cues of a certain generic object are allocated on instantiation. To avoid obligatory synchronization of all parametric trajectories, the target is validated only if the parameter's value differs from a default value specific to each parameter. For example, the following rule generates the two targets (oral and naso-pharyngeal parts) for a French nasal vowel / \tilde{a} /:

```
class Target
  object(X0,X1,X3) feature(Final) cue(t,F1,F2,F3)
endclass
an:  $\tilde{a}$  -> #1(<X3, t=20, F1=515, F2=1350, F3=2100><X0, t=50, F1=530, F2=1000, F3=2200>);
```

The set of F1-F4 targets for each phoneme is then taken for R1-R4 targets as developed in the preceding chapters. Once the entire resonance and residual parametric trajectories have been computed by Gentraj, two other functions of the library are then called to produce sound: Coupling and Klatt. Coupling computes new ordered values for F1-F4 using the following algorithm:

```
- sort F1-F4 in ascending order
- for all Fi do
  for all j>i do
     $d = \sum 200 \cdot \text{sgn}(i-j) \cdot \exp(-|Fi-Fj|/200)$ 
     $Fi = Fi + d$ 
  enddo
enddo
```

4.3. Vocalic target resonances

Examples of resonance target values for the 10 French oral vowels and some consonants are given below:

//	R1	R2	R3	R4	B1	B2	B3	B4
a	620	1240	3330	2500	75	60	130	150
o	460	1000	3000	2600	80	65	130	50
o	370	800	2260	3000	55	65	110	100
u	250	750	2100	3000	60	60	95	110
\tilde{a}	480	1420	3200	2300	70	70	90	95
ϕ	360	1600	2150	3000	60	70	80	75

y	250	1720	2060	3000	75	95	125	80
e	590	1900	3300	2300	60	100	110	120
e	320	2700	2000	3300	55	85	80	100
i	250	3000	2000	3400	55	60	100	100
j	250	2500	2100	3200	85	80	100	110
w	250	750	2100	3000	60	90	90	110
q	250	1600	2200	3200	75	95	120	80
m	250	1300	2300	3000	80	130	140	140
n	250	1450	2600	3300	90	50	150	130
r	250	1100	2300	3400	90	90	140	160
l	250	1650	2100	3400	90	120	145	190

5. CONCLUSIONS

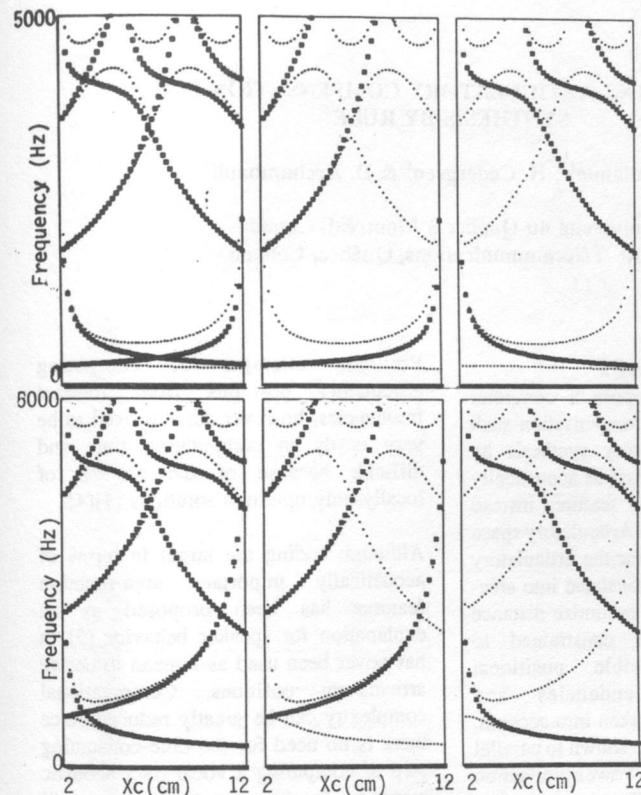
Spectrogram reading in light of resonance structures enables a clear acoustic-to-articulatory inversion [1] and thus enables an easy way of modeling formant trajectories. Consonant loci have to be revised to take account of this new vocalic structure. We think that this acoustic representation of the speech signal may suggest new investigation for a number of key problems such as effective formant calculation [8]: F2 could be the result of tracking of R2 instead of a large-scale integration.

Acknowledgements

We are very thankful to P. Badin for his suggestions on resonance structures.

REFERENCES

- [1] ATAL, B.S., CHANG, J.J., MATHEWS, M.V. & TUKEY, J.W. (1978) "Inversion of Articulatory-to-Acoustic Transformation in the vocal tract by a computer sorting technique", *J. Acoust. Soc. Am.*, 63, 1535-1555.
- [2] BAILLY, G., MURILLO, G., AL DAKKAK, O. & GUERIN B. (1988) "A text-to-speech synthesis system for French by formant synthesis", *7th FASE Symposium*, 255-260.
- [3] BAILLY, G. & TRAN A. (1989) "COMPOST: a rule-compiler for speech synthesis", *Eurospeech*, 136-139.
- [4] BADIN, P. & BOE, L.J. (1987) "Vocal tract nomograms: acoustic considerations", *Proc. of XIth Int. Cong. of Phon. Sci.*, 352-355.
- [5] BADIN, P. & FANT, G. (1984) "Notes on vocal tract computation", *STL-QPSR*, 2-3, 53-108.
- [6] FANT G. (1960), "Acoustic Theory of Speech Production", The Hague: Mouton & Co.
- [7] GUERTI, M. & BAILLY G. (1990) "Anticipation et retention dans les mouvements vocaliques du français", *XIII^e Journées d'Etudes sur la parole*, 292-295.
- [8] SCHWARTZ, J.L. & ESCUDIER, P. (1987) "Does the human auditory system include large scale spectral integration?", *The psychophysics of speech perception* (Schouten, M., Editor), Nato Asi series, Dordrecht, 284-292.



Figs.1a & b: Vocalic nomograms: on the nomogram for the 4-tube model are superposed from right to left: the nomograms of a) the back cavity alone, b) the front cavity and c) both cavities considered as independent. Top captions are for the closed case ($A_l=0.16 \text{ cm}^2$) and bottom for the open case ($A_l=4.0 \text{ cm}^2$).

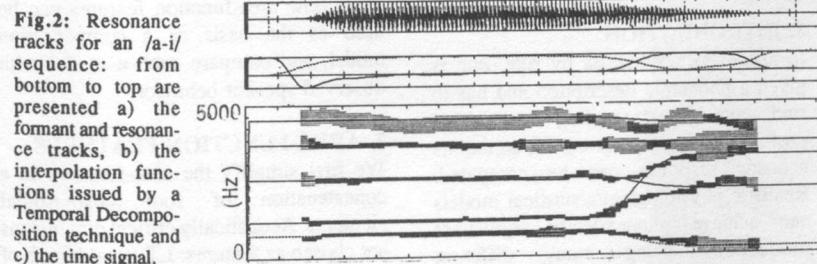


Fig.2: Resonance tracks for an /a-i/ sequence: from bottom to top are presented a) the formant and resonance tracks, b) the interpolation functions issued by a Temporal Decomposition technique and c) the time signal.

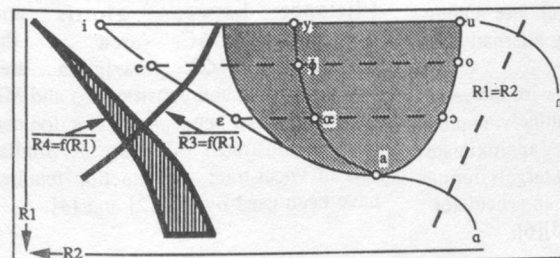


Fig. 3: Vocalic triangle in the R1-R2 plane: the classic F1-F2 vocalic triangle is figured in dark. Realization space for R3 and R4 are figured in heavy dark. Resonance targets for extreme back vowels /y/ and /a/ are suggested.

MODELLING ARTICULATORY COMPENSATION FOR SYNTHESIS BY RULE

G. Boulianne^{1,2}, H. Cedergren¹ & D. Archambault²

¹Université du Québec à Montréal, Canada

²INRS-Télécommunications, Québec, Canada

ABSTRACT

In this paper we propose a computer model of articulatory compensation such as needed in articulatory synthesis by rule. Targets are specified as acoustically important area-function features instead of formant frequencies. Articulatory space can then be searched for the articulatory position which, once translated into area-function features, will minimize distance to target. Search is constrained to physiologically possible positions; interarticulator dependencies and articulatory effort are taken into account. The model's behavior is shown to parallel that of real speakers in vowel production and bite-block experiments.

1. INTRODUCTION

In articulatory synthesis by rule, one is given a phonemic description and has to find the corresponding articulator positions of a computer model, so that its acoustic response can be computed. Realistic physiological/acoustical models can achieve phonemically equivalent productions using many different articulator configurations. The problem is then to find a strategy to choose among all the possible articulatory alternatives.

Human speakers are faced with the same problem; in fact they routinely exploit this freedom to closely approximate intended formant frequency targets during normal speech [8] or when an articulator is artificially constrained [5][6]:

Previous attempts at computing articulatory positions from formant frequencies, however, have proved to be very costly in computation time and difficult because of the number of locally-only optimum solutions [1][4].

Although coding the target in terms of acoustically important area-function features has been proposed as an explanation for speaker behavior [5], it has never been used as a mean to derive articulatory positions. Computational complexity can be greatly reduced since there is no need for the time-consuming step of computing a vocal tract acoustic response. In the next sections we will show how area-function features can be used as the basis of a compensation model, and compare such a model with observed speaker behavior.

2. AREA-FUNCTION FEATURES

We first simplify the area-function to a concatenation of four cosinusoidal elements. Acoustically critical dimensions are chosen as features: LT (total length of vocal-tract), OL (area at the lips), XC (distance between glottis and constriction), AC (area at the constriction), MG (maximum area between glottis and constriction) and ML (maximum area between constriction and lips) as illustrated in Figure 1. Similar sets of vocal-tract area-function features have been used by [1], [2] and [4].

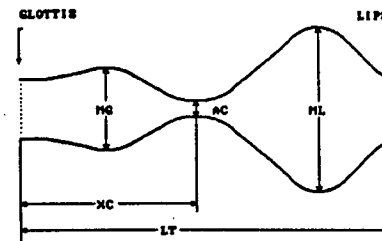


Figure 1. Area-function features

3. ARTICULATORY MODEL

In order to simulate inter-articulator dependencies and limit the search to physiologically realistic positions, we implemented the articulatory model of [7]. Only six major articulatory parameters were used: AM (jaw opening), AC and DC (tongue body position), AA (tongue tip angle), EL (vertical lips separation) and AL (lips protrusion). Most of the parameters are not absolute positions but measure articulator displacement relative to others (see [7]).

Given specific values for the parameters, the articulatory model generates a sagittal contour and, from it, area of sections along the length of the vocal-tract. Features are then derived from this area-function. The whole process is quite computation intensive, due to the articulatory model complexity. We developed a polynomial approximation to that process, such that area-function features are computed directly as a weighted sum of polynomial combinations of the articulatory parameters. This approximation has the benefit of smoothing out discontinuities in the articulatory-to-area-features relationship - that could prevent the search from reaching global optimum - as well as providing a tenfold reduction in computing time.

4. COMPENSATION STRATEGY

The model strategy is to seek for an optimal articulatory position, that is, one

which translated into area-function features will be closest to the feature target. The choice of a distance measure that defines "closeness" has a profound influence on the final behavior of the model.

4.1. Distance measure

We selected a simple weighted euclidean distance: the sum of the squared differences between features, where each feature is first divided by its standard deviation over a training set.

By itself, however, this distance measure makes no difference between "easy" and "difficult" to reach positions, and could accept unreasonable positions as good solutions. Modifying the distance measure to take articulatory effort into account can be done by simply adding to it a sum of the squared differences between each articulator and its rest position, where each articulator is first divided by its rest position.

Note that both "distance to feature target" and "effort" components of the total distance are expressed as ratios. Using such dimensionless units avoids introduction of arbitrary weighing coefficients, that would otherwise be needed to adjust each component's contribution to the total distance.

4.2. Validation of effort component

An experiment was run to ascertain the effectiveness of the "effort" component in producing positions resembling those which speakers would choose. We compared solutions obtained using this distance measure with radiographic evidence from French speakers. Figure 2 plots jaw angles given by the model as a function of incisor distance measurements made on radiographic data from continuous speech [9]. To obtain the jaw angles the compensation model was fed with features computed from the area-function data of [3]. The correlation of

79% with radiographic data is quite impressive, given the fact that the area-function features contained no specific information about jaw angle: the agreement between simulated and real data comes solely from the use of the "effort" component in the distance measure. Correlations for other articulatory parameters that have been measured on radiographic data in [9] are 83% for vertical lips separation and 88% for lips protrusion. In these results, both the "distance" and "effort" components come into play, since area features OL and LT do contain some information about lips position.

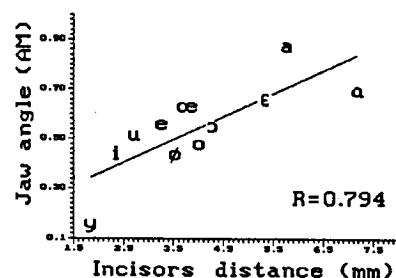


Figure 2. Vowel opening for computer and radiographic data

5. EXPERIMENTAL RESULTS

By constraining the jaw parameter to a single constant value during the search, we ran the computer equivalent of a bite-block experiment. Using the computer model and area-function data from [3], we obtained articulatory positions for the vowels /i,a,u/ in three cases:

1. *Normal case*: no articulator constraints.
2. *Compensated case*: imposing a specific "far from natural" jaw angle value while allowing other articulators free movement.
3. *Uncompensated case*: using articulators obtained in condition 1 and moving only the jaw to the imposed angle of condition 2. This shows the effect of the bite-block as if no compensation had taken place.

For /a,u/ the imposed jaw angle was $AM=0.25$ (equivalent to a 5.5 mm incisors distance), and for /i/ it was $AM=0.80$ (22 mm incisors distance). These conditions are similar to those used in [5] for human speakers.

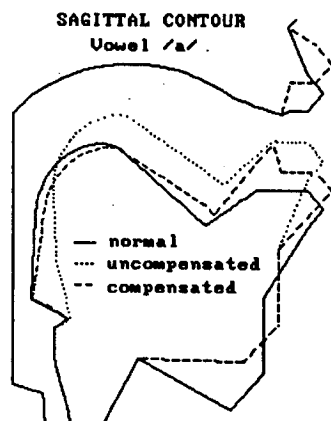


Figure 3. Articulators in bite-block experiment

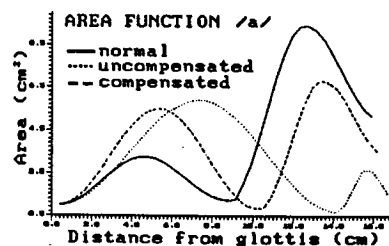


Figure 4. Area-function in bite-block experiment

For all three vowels area-function features in the uncompensated case are far from the normal case (average difference 66%), while values in the compensated case are close to the normal case (average difference 17%). Worst compensation occurred for vowel /a/. Figure 3 shows /a/ articulatory positions, and Figure 4 corresponding area-functions. Note that compensation is not perfect because the effort would be too

high. Main observations of [5] on human speakers also applies to our data:

1. Compensation is attained by "supershaping of the tongue relative to its attachments to the jaw" [5], in our case affecting parameters that define tongue body position relative to the jaw. A lip parameter (jaw-relative vertical lip position) is also affected.
2. Area and position of main constriction are better preserved (average difference 17%) than front and back cavity areas (average difference 25%).
3. Area at the lips is better preserved in /u/ than /i/ or /a/ (differences of respectively 5%, 28% and 33%).
4. Acoustic response computed from the area-functions shows that the first two formant frequencies in the compensated case approximate well those of the normal case (average difference 11%), while formants in the uncompensated case would be far from normal (average difference 40%).

6. CONCLUSION

Our model simulates articulatory compensation as an optimality seeking process. Coding production targets as acoustically-important area-function features is efficient and reproduces speaker behavior in bite-block vowels experiments. Adding an effort component insures that articulatory displacement trade-offs that occur during continuous speech vowel production are also correctly simulated.

As it stands, this model is currently used for both consonant and vowel production, but has only been validated for vowels. Dynamic effects like coarticulation are not modelled, but can easily be included by adding distances to past and future positions.

7. REFERENCES

- [1] ATAL, B.S., CHANG, J.J., MATHEWS, M.V., & J.W. TUKEY (1978) "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", *J.A.S.A.*, vol. 63, no 5, p. 1535-1555.
- [2] COKER, C.H. (1976) "A Model of Articulatory Dynamics and Control", *Proc. IEEE*, vol. 64, no 4, pp. 452-460.
- [3] FENG, G. (1987) "Etude articulatoire-acoustique des voyelles nasales du français", *Bulletin de l'Institut Phonétique de Grenoble*, vol. 16, p. 1-102.
- [4] FLANAGAN, J.L., ISHIZAKA, L., SHIPLEY, K.L. (1980), "Signal models for low bit-rate coding of speech", *J.A.S.A.*, vol. 68, no 3, pp. 780-791.
- [5] GAY, T., B. LINDBLOM & J. LUBKER (1981) "Production of Bite-Block Vowels : Acoustic Equivalence by Selective Compensation", *J.A.S.A.*, vol. 69, no 3, pp. 802-810.
- [6] LINDBLOM, B., J. LUBKER & R. MCALLISTER (1977) "Compensatory Articulation and the Modeling of Normal Speech Production Behavior", dans R. Carré, R. Descout & M. Wajskop, eds., *Modèles articulatoires et phonétiques*, Grenoble, pp. 147-161.
- [7] MERMELSTEIN, P. (1973) "Articulatory Model for the Study of Speech Production", *J.A.S.A.*, vol. 53, no 4, pp. 1070-1082.
- [8] PERKELL, J.S. (1989) "Testing Theories of Speech Production : Implications of some Detailed Analyses of Variable Articulatory Data", in Hardcastle & Marchal, eds., *Speech Production and Speech Modelling*, Kluwer, Dordrecht, 448 p.
- [9] SIMARD, C. (1985) *Etude des séquences du type consonne constrictive plus voyelle en français, à l'aide de la radiocinématographie et de l'oscillographie*, CIRB, Québec, 403 p.

CONSTITUTION SEMI-AUTOMATIQUE DE LEXIQUES DE CONTOURS PROSODIQUES POUR LA SYNTHÈSE A PARTIR DU TEXTE

V. Aubergé et M. Contini

ICP - Université Stendhal, UA 368-
38000 Grenoble - France

ABSTRACT

Each application of text-to-speech synthesis requires a specific prosodic strategy (depending on the context). Furthermore, the elaboration of these strategies is tightly linked to the analysis capabilities of the automatic text processing, which is then to reproduce the prosody of the application-specific corpus. A complete system was elaborated to handle at the same time the textual data and the prosodic data of the corpus. The successive processing modules of such a system are : automatic extraction, visualization and averaging of prosodic contours.

The next step of this work is the constitution of different prosodic modules, suitable for varying linguistic situations.

1. INTRODUCTION

Lorsque l'on aborde les technologies vocales, et en particulier la synthèse à partir du texte, on s'aperçoit que la qualité segmentale de la voix synthétique est devenue suffisante pour que la qualité de l'intonation synthétique soit maintenant un problème prioritaire, puisque l'intonation est directement reliable à l'intelligibilité et pas simplement à des critères de naturalité ou d'agrément.

Dans un système artificiel, l'entrée est un texte, c'est-à-dire une projection sémiotique du langage. Pour transformer ce message symbolique en message parlé, on peut retrouver dans l'écrit certaines des structures du langage mais on ne sait rien d'éventuels modèles psycholinguistiques du locuteur. C'est pourquoi, afin d'améliorer la qualité prosodique d'un système de synthèse on propose ici une

méthodologie, basée sur l'analyse inductive de corpus [1] & [2] qui a permis la conception d'un module de génération de l'intonation pour la synthèse.

Le corpus est construit *a priori* selon des critères linguistiques que l'on qualifiera d'hypothético-déductifs. Dans un premier temps, des hypothèses sont émises sur des points d'ancrage des structures intonatives autour d'indices déductibles de l'écrit. A partir de ces indices linguistiques est construit un corpus symbolique, qui devient un corpus oral après enregistrement. Ensuite, toutes les données du corpus décomposées et classifiées sont moyennées et constituent une base de contours paramétrés par des indices linguistiques. Si on est capable de bien choisir les liens de coïncidence entre les structures de l'écrit (hypothèses *a priori*) et celles de l'intonation, nous pouvons espérer que les représentants des classes intonatives permettront de reconstruire une structure anthropophonique. La génération de l'intonation se résume ensuite à une application hiérarchisée de ces contours pendant la phase de synthèse.

2. LE CORPUS

Le choix du corpus est fondamental. Il doit être le résultat d'une démarche logiquement corrélée avec la méthode d'analyse qui doit ensuite s'appliquer au corpus. Sa qualité de représentativité est déterminante quelle que soit la finesse de l'analyse qu'on en fera. Puisque l'entrée du synthétiseur est un texte, le choix des indices pertinents est en pratique fortement limité par les capacités des analyses textuelles automatiques avec lesquelles il est nécessaire d'entretenir des

relations étroites.

L'hypothèse la plus forte retenue *a priori* est celle de "rendez-vous" structurels entre intonation et syntaxe [3]. Cette relation sera d'autant plus émergente du corpus que le locuteur est mis dans une situation de lecture de phrases isolées non marquées sémantiquement. Parallèlement, il est possible, avec les analyseurs du CRISS de produire une analyse morpho-syntaxique automatique de la phrase [4].

Le corpus a été construit pour être analysé hiérarchiquement par niveaux syntaxiques successifs : l'hypothèse étant que non seulement l'intonation coïncide avec la syntaxe par morceaux, mais surtout qu'il est possible d'associer une unité intonative caractéristique à chaque niveau considéré.

Le niveau maximal représenté dans notre corpus est la phrase. Les niveaux inférieurs étudiés sont dans l'ordre la proposition, le syntagme, l'unité lexicale et enfin la syllabe.

A chacun de ces niveaux sont étudiées l'influence de la longueur (en nombre de syllabes) de l'unité considérée et sa position syntagmatique à l'intérieur du ou des niveaux supérieurs. Chaque unité est caractérisée par des attributs spécifiques qu'on suppose distinctifs pour les unités intonatives : e.g. la modalité pour la phrase, la fonction de dépendance pour la proposition, la nature ou la fonction syntaxique pour le syntagme, son organisation en constituants... Le niveau sémantique est abordé par l'utilisation de traits lexicaux.

A chaque patron linguistique produit, on choisit un représentant (i.e. un énoncé de phrase textuel) qui respecte des contraintes phonétiques afin de garantir un étiquetage correct du corpus acoustique incident.

Un premier corpus de 164 phrases a été enregistré selon une ergonomie qui, par expérience, laisse espérer une bonne cohérence interne dans la stratégie prosodique du locuteur. Il est à noter que ce locuteur est celui du dictionnaire de polysons qui constitue la méthode de synthèse du système sur lequel a été testé cette méthodologie.

3. L'ANALYSE DU CORPUS

Les paramètres physiques choisis sont classiquement Fo et durée. Le paramètre

durée est bien sûr délicat à acquérir puisque s'y projettent aussi bien les structures segmentales que suprasegmentales. Quant à Fo, il a été montré (et nous avons pu le vérifier *a posteriori*) que l'acquisition de trois points par voyelle (début, *extremum*, fin) est suffisante pour restituer l'évolution suprasegmentale de ce paramètre pour le français.

Il a été obtenu ainsi une base de données simplifiées, où sont associés le corpus symbolique étiqueté (par les paramètres linguistiques selon lesquels il a été construit), et le codage de Fo et durée. Un ensemble de logiciels de gestion du corpus manipulent et traitent les données en fonction des étiquettes linguistiques qu'elles contiennent. Des outils de visualisation des paramètres acoustiques permettent d'observer les données intonatives manipulées, avec diverses possibilités de normalisation, de myennage et de superposition des contours. La méthodologie analytique du traitement passe à la fois par la classification des contours homogènes et par une analyse contrastive de ces contours :

- Pour décider de la qualité d'une structure du texte à coïncider avec une structure intonative, toutes les structures correspondantes dans le corpus sont rassemblées en une même classe. Un contour-moyen est conservé dans la base en association avec les paramètres linguistiques qui spécifient la structure textuelle de cette classe.

- L'autre méthode d'investigation systématiquement utilisée avant la classification puis ensuite comme vérification sur les contours-moyens calculés est une méthode contrastive par paire minimale. Lorsqu'un paramètre linguistique est supposé pertinent, alors dans le corpus sont choisis deux représentants pour lesquels ce paramètre varie, toute chose étant égale par ailleurs. Si la paire minimale est effective alors l'ensemble des paramètres (les paramètres étudiés et contextuels) seront spécifiques d'une classe de contours.

Voici en exemple quelques extraits du corpus symbolique à partir desquels seront comparés les contours intonatifs

correspondant aux éléments notés entre " / " :

paramètre étudié : position respective Nom-Adjectif (construction du groupe)

paramètres contextuels : Groupe Nominal (GN), début de phrase assertive, 4 syllabes

/ Ce beau passant / chantait.

/ Ce passant fou / chantait.

/ Ce fantastique passant / chantait.

/ Ce passant fantastique / chantait.

paramètre étudié : longueur

paramètres contextuels : adjectif, dans la structure GN construit selon le modèle de la phrase 2, début de phrase assertive.

/ Ce beau passant / chantait.

(1 syllabe)

/ Ce petit passant / chantait.

(2 syllabes)

/ Ce fantastique passant / chantait.

(3 syllabes)

paramètre étudié : fonction de dépendance de proposition

paramètres contextuels : 7 syllabes, fin de phrase assertive.

Je vois ces enfants ; / ils jouent avec un balai !.

Quand je vois ces enfants, / ils jouent avec un balai !.

Je vois ces enfants / quand ils jouent avec un balai !.

Avec une telle méthode, il est à chaque instant possible d'affiner l'analyse en intégrant de nouveaux paramètres. La cohérence avec les capacités des analyses textuelles automatiques est ainsi assurée par une simple mise à jour de la base des contours.

4. LA BASE DE CONTOURS

Au niveau de la phrase, l'unité intonative conservée est la ligne de déclinaison pour chaque longueur de phrase et pour chaque modalité. La cohérence à l'intérieur de chaque classe est vérifiée (écart-type < 1/4 ton). Les lignes-moyennes sont, elles aussi, très cohérentes (écart-type < 1/4 ton pour les assertives et 1/3 pour les interrogatives), c'est donc la pente qui varie avec la longueur de la phrase, les valeurs des bornes restent constantes. Des

variations locales entre chaque classe sont utilisées en synthèse pour une rupture de la monotonie.

Au niveau de la proposition, la fonction de dépendance a un intérêt distinctif évident. C'est également une ligne (attaque, finale) qui est calculée pour chaque classe paramétrée par la fonction de dépendance, la position dans la phrase et la longueur.

Au niveau inférieur, celui du syntagme, intervient pour chaque groupe un paramètre booléen (fin, non fin) de phrase assertive. Les contours d'un même groupe en position fin et non fin sont similaires jusqu'à l'avant dernière syllabe. Quant à la dernière syllabe, elle présente une pente négative de valeur à peu près identique quelque soit le groupe. Plutôt que de calculer une règle de transformation d'un contour de non fin à fin, ont été stockés les deux types de contours. L'unité la plus finement représentée est le groupe nominal (GN). On retrouve dans ce corpus la limite de quatre syllabes pour le groupe intonatif. Les constituants ne sont pas particularisés à l'intérieur d'un GN de moins de quatre ou cinq syllabes (quelque soit la construction du GN, son contour est similaire) et deviennent des unités intonatives dans les GN plus grands (les contours du GN s'opposent selon sa construction). Pourtant, si l'on observe le comportement de l'adjectif ou du nom, lorsque sa longueur varie à l'intérieur d'un même GN, son évolution reste très cohérente et tout à fait indépendante de la longueur du GN. Les contours-moyens des GN sont stockés, paramétrés par une variable de position dans la phrase (fin, non fin), par la situation par rapport au verbe et par une indication booléenne sur la fonction syntaxique du GN en valeur de complément circonstant ou pas. Pour les GN de un, deux, trois et quatre syllabes, ce contour sera maximal ; pour les GN de plus de quatre syllabes, les contours des constituants interviendront dans la reconstruction de la structure pendant la phase de génération.

Les groupes verbaux ont été classifiés selon leur appartenance ou non aux verbes outils, leur longueur et leur position dans la phrase.

Une étude restreinte des adverbes autonomes a donné des contours particularisés par leur position dans la

phrase et par rapport au verbe.

5. CONCLUSION

Cette démarche a pour principe de formuler des hypothèses linguistiques les plus faibles possibles (mais alors taille du corpus ?) et de laisser les données "s'auto-classifier". Le rôle de l'expert est isolé et une étape ultérieure consistera à le substituer par un système formel. Des tests de validation de l'intonation générée sont en cours. Les premiers résultats sont encourageants : l'intonation obtenue est contrastée et de plus elle paraît effectivement se rapprocher de l'intonation de notre locuteur. Il s'est avéré que sa stratégie s'est bien organisée selon les hypothèses posées à la construction du corpus, sans doute en partie parce que ce locuteur a été placé en situation non spontanée de lecture. En dehors de tout contexte discursif, il n'a eu comme support formel que la syntaxe de phrase.

Cette première expérience a pu valider cette méthodologie pour la synthèse. On peut facilement imaginer que la même méthode de constitution et d'analyse de corpus soit appliquée à d'autres langues avec l'ensemble des outils mis en place. L'intonation est un facteur de qualité essentiel de la synthèse vocale si elle devient spécifique par exemple d'une situation de messagerie, de dialogue homme-machine, de bureautique multimedia ou de commandes à distance, chacune de ces applications étant caractérisable par une base de contours. Cette méthode sera sera toujours limitée par le choix de paramètres linguistiques pertinents, en fonction du contenu intonatif du corpus et de la capacité des analyses linguistiques automatiques, mais on peut supposer que plus les spécifications linguistiques seront précises, moins les structures intonatives seront étendues.

6. REFERENCES

- [1] CONTINI M. & PROFILI O. (1987), "Génération automatique de schémas macroprosodiques en italien, à partir d'un texte écrit.", 16^{èmes} JEP - GALF, 245-248.
- [2] EMERARD F. & BENOIT C. (1987), "De la production à

l'extraction, l'état d'un chantier.",

16^{èmes} JEP - GALF, 224-228.

[3] ROSSI M. DI CRISTO A. HIRST D. MARTIN P. & NISHINUMA Y. (1981), "L'intonation, de l'acoustique à la sémantique", Klincksieck Ed., Paris.

[4] ROUAULT J. (1988), "Linguistique automatique, applications documentaires", Sciences pour la Communication, Peter Lang, Berne.

3.2 The Rule Syntax

In our approach a context-sensitive rule consists of a condition part and an action part. The first comprises one or more conditions, which may be combined by the logic operators '&' (and) and 'r' (or). The latter contains one or more actions. In the first place these can be thought of as transformations; however, in view of the procedure of rule application we need additional functions to control both the scanning of the input and the transformation process.

3.3 The Rule Interpreter GRIP

The described formalism has been put into practice by means of the rule interpreter GRIP (Graphon Rule Interpreter) which translates a source textfile of rules into executable instructions.

3.4 Co-operation with LIFT

GRIP rules operate on a LIFT representation of a given text. They are applied once in a serial order, and every single rule scans an entire text unit (normally a sentence). To be more precise, via a so-called condition pointer a particular tier of LIFT is selected to constitute the input for the rule. (Note that, thanks to the links between elements of different tiers, any structural context associated with a particular element of the selected tier can be tested in the rule's condition part.) A second pointer, the so-called action pointer, directs the actions to a second (possibly the same) tier. During rule application both pointers are moved further along the tiers.

4. APPLICATION IN PHONOLOGY AND PHONETICS

4.1 Example

To illustrate syntax and processing of rules, consider the following phonological transformations encountered in standard Austrian German:

- ə-deletion (1)
- progressive nasal assimilation (2)
- syllabification of the nasal (3)

These can be formalized as follows:

$$ə \rightarrow 0 / [+obstr] \text{ — } n \quad (4)$$

$$n \rightarrow [\alpha \text{ place }] / [+obstr] \text{ — } (5)$$

$$[+nas] \rightarrow [+syll] / [+obstr] \text{ — } (6)$$

Thus:

<i>Lippen</i> (lips)	→	['lɪpm]
<i>laufen</i> (to run)	→	['lɔʏfm]
<i>Regen</i> (rain)	→	['re:gn]

Using GRIP the rule for e.g. *Lippen* looks like the following:

$$[+E(-1,p) \& +E(0,ə) \& +E(1,n)] \\ \text{del skip}(1,r) \text{ chg-el}(m) \quad (7)$$

The condition part (in square brackets) of (3) consists of three conditions that are concatenated by the operator '&'. Whenever processing of a rule is initiated, the condition pointer addresses the first element of a specific tier, in this case the phonetic tier. Subsequently the condition pointer is moved along this tier. As soon as the element 'ə' is addressed, all three is-element conditions (+E) are met since the preceding element is 'p' and the succeeding element is 'n'. This causes execution of the action part. First, the element addressed by the action pointer ('ə') is deleted (del); next, the action pointer is moved one element further (skip) (note that whenever the element addressed by the action pointer is deleted, the latter addresses the preceding element.); finally the element 'n' is changed into 'm' (chg-el).

Taking into account the context 'pən' and 'bən' only, rule (7) is an elegant way to implement the transformations (1,2,3). Since the rules (4,5,6) are valid for any context, (7) has to be extended. By exploiting a feature-based representation of phonemes GRIP allows to combine (4) and (6) in a single rule (8).

$$[+E(0,ə) \& \\ +F(-1,CONS) \& -F(-1,NAS) \& \\ +F(1,CONS) \& +F(1,NAS) \& \\ -E(-1,r) \& -BEG(1,SYLL)] \\ \text{del}() \text{ skip}(1,r) \text{ chg-F}(SYLL) \quad (8)$$

(8) reads as follows:

Every 'ə', preceded by a non-nasal consonant (+/-F denotes presence/absence of the specified feature), succeeded by a nasal consonant, is deleted (del) and the subsequent nasal becomes syllabic (chg-F(SYLL)).

Note that the sequence 'ər' is treated within a separate rule and thus is excluded in (8). Finally the condition -BEG(1,SYLL) serves to prohibit a syllable boundary between 'ə' and the nasal, e.g. *genommen* (taken): [gə'nɔmən] and not *[gɔmən].

With regard to the implementation of (5), basically three separate rules would have to be written, in order to account for each place of articulation (velar, labial, and labiodental). The elegant notation of (5) is due to the notion of "α-place". In (9) we therefore introduce "accept", a GRIP action to copy feature bundles from neighbouring phonemes.

$$[+E(0,n) \& +F(-1,OBSTR) \& \\ (+F(-1,ANT) | \\ +F(-1,HIGH \& BACK))] \\ \text{accept}(-1, HIGH/BACK/LAB/ \\ ANT/COR) \quad (9)$$

(9) reads as follows:

'n' preceded by an obstruent which is either anterior or high and back, accepts the features high, back, labial, anterior and coronal from the obstruent. Since these 5 features serve to describe the place of articulation, the nasal is assimilated, yielding 'm', 'ɱ', 'ŋ', or 'ŋ'.

Note that in the condition part the palato-alveolar articulation ('j' 'ç...') is excluded. In fact, this should have been done in (5) as well. IPA does not provide for a palato-alveolar nasal, thus (5) takes for granted that in such a case the nearest possible place of articulation will be

chosen. With regard to a computer implementation implicit assumptions of this kind have to be analysed very carefully.

4.2 Conclusion

The above rules primarily refer to the phonetic tier. However, other rules, in particular rules concerning supra-segmentals, obviously depend on various kinds of linguistic information (e.g. morphological and syntactic structure).

Within the text-to-speech-synthesis system GRAPHON, phonological and phonetic rules fill up LIFT, exploiting the information previously generated in the morphological and syntactic analysis (cp. fig. 1). To this end neither condition part nor action part of GRIP rules are bound any longer to a single tier as it was mostly the case in the introducing example in 4.1.

The joint processing of context conditions making reference to several structural levels at the same time significantly extends the linear representation of segments in SPE rules. Besides their practical relevance within text-to-speech synthesis, LIFT and GRIP provide the linguist with a powerful tool for rule development and test.

ACKNOWLEDGEMENTS

The research project has been receiving financial support from the "Fond zur Förderung der wissenschaftlichen Forschung (FWF)".

REFERENCES

- [1] CHOMSKY, N. & HALLE, M. (1968), *The Sound Pattern of English*, New York: Harper & Row.
- [2] CLEMENTS, G. (1985), *The geometry of phonological features*, Phonology Yearbook 2, 225-253.
- [3] HERTZ, S. R. (1988), *The Delta programming language: an integrated approach to non-linear phonology, phonetics, and speech synthesis*, Working papers of the Cornell Phonetics Laboratory 2, 69-122.
- [4] KOMMENDA M. & POUNDER A. (1986), *Morphological Analysis for a German text-to-speech System*, Proc. COLING'86, Bonn, p. 263-268.

A COMPARISON OF THE INTELLIGIBILITY SCORES OF CONSONANTS AND VOWELS USING CHANNEL AND FORMANT VOCODED SPEECH

R.H. Mannell and J.E. Clark

Speech, Hearing and Language Research Centre
Macquarie University, Sydney, Australia

ABSTRACT

The intelligibility of various phonetic classes is examined following vocoding by a formant vocoder and various (Hz-scaled and Bark-scaled) implementations of a channel vocoder. The results suggest that particularly in the case of various consonant classes the 1 Bark channel vocoder performed a little (but significantly) better than the Hz-scaled channel vocoders and much better than the formant vocoder. The 1 Bark vocoder achieved intelligibility results equivalent to natural speech for most phonetic classes. The results support the idea that channel vocoding techniques are intrinsically capable of achieving natural speech intelligibility and suggest that formant systems may be intrinsically incapable of achieving natural intelligibility for certain phonetic classes.

1. INTRODUCTION

This work arises from a general interest in synthesiser performance and particularly in the performance of competing parametric encoding strategies. The limitations of speech synthesis performance is well recognised and there is a growing body of quantitative evidence [1,3,6,7] as to the nature of these performance limitations. Synthetic vowel intelligibility is often equivalent to that of natural vowels whilst consonants, on the other hand appear to consistently demonstrate a shortfall in synthetic relative to natural intelligibility [1,6,7]. Most of these studies examined the intelligibility of synthesis-by-rule and text-to-speech systems and all examined the performance of formant synthesisers. One question that this study attempts to address is whether these findings reflect merely our limited ability to formulate rules for the generation of consonants using a formant-based synthesis-by-rule system or whether there is a more fundamental limitation in the potential performance of formant synthesis with respect to consonants. An even greater motivation for the present

study is an examination of whether channel synthesis also shares with formant synthesis any fundamental limitation in its ability to synthesise consonants and further, which filter configurations produce consonant intelligibility performance approaching that of natural speech. Vocoders were utilised in this study as they allow a direct examination of various encoding strategies without the confounding effects of rule and database defects potentially inherent in synthesis-by-rule systems. Further, vocoder software simulations are very flexible allowing easy modification of filter configurations etc.

The present paper is a further progress report on a study outlined at the Tallin conference [2].

2. PROCEDURE

The primary means for manipulating the parametric information content of the resynthesised speech was via a classical channel vocoder, first described by Dudley [4]. A channel vocoder was used because it is relatively free of any major a priori assumptions about the primacy or otherwise of particular spectral features such as energy peaks or depressions as bearers of phonological information. A channel vocoder comprehensively encodes all spectral components able to be resolved by the frequency resolution (bandwidth) of the filters. Being a vocoder the speech is resynthesised directly from spectral information extracted from a natural speech signal and not from rules and databases as would be the case with synthesis-by-rule systems.

The vocoder (figure 1) is a software simulation residing on a VAX 11/750 computer consisting of band pass (BP) and low pass (LP) FIR filters designed using the well known window synthesis technique. This allowed for considerable flexibility in

filter design and numerous filters of different time and frequency domain characteristics have been designed and used over this project. The present paper will only examine a subset of these filter configurations in which the BP channel filter bandwidths are varied in both the Hz-scale and the Bark scale. The pitch and excitation detection algorithms were adapted to the limited input data and all decisions made by that module were confirmed by a experienced phonetician. It is unlikely that the pitch/excitation module contributed in any way to the final intelligibility results.

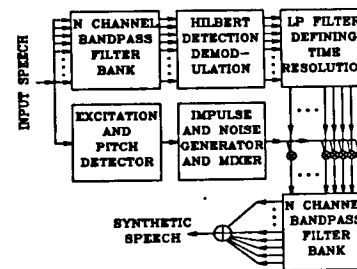


FIGURE 1. S.H.L.R.C. CHANNEL VOCODER

The speech was digitised and bandlimited to 0 to 5 kHz. Four Hz-scaled BP channel filterbank configurations (Bandwidths: 100, 200, 400 and 800 Hz) and five Bark-scaled filterbanks (Bandwidths: 0.75, 1.0, 1.5, 2.0 and 3.0 Bark) were utilised in the part of the project described in this paper. The outputs of these filters were passed through identical Hilbert transforms. These Hilbert filters were inserted before the following LP filters to maintain a constant spectrum envelope demodulator (deemed desirable for conditions in which the LP filter was varied to manipulate the time resolution of the total system). For the conditions reported upon in this paper the LP filters were fixed at 50 Hz (a time resolution of 10 msec as defined by the sampling theory) and as this filter was the "slowest" filter in the system it defined the time resolution of all the systems as a constant 10 msec. In all conditions the BP synthesis filters were identical to the BP analysis filters.

The formant vocoder used in this study was developed at the Joint Speech Research Unit (JSRU) in the United Kingdom and was based on an automatic formant analysis system described by Dupree [5] and coupled

with the highly regarded JSRU formant synthesiser described by Rye and Holmes [8]. The time resolution of this system was the same as that used in the channel vocoder configurations (ie. 10 msec).

The test items were 11 vowels in an /h_d/ frame and 19 consonants in a CV frame (V=/a:/) spoken by a native speaker of Australian English. These tokens were recorded to professional audio standards in an echo free room and digitised (16 bits) onto and vocoded on a VAX 11/750 computer. The tests were conducted in a sound treated room using calibrated TDH-49 headphones with standard cushions and circumaural seals. The level normalised test tokens were presented both in silence and in +6, 0 and -6 dB S/N (utilising USASI speech-shaped noise) at a presentation level of 70 dB s.p.l. (ref. 20 µPa). There were 20 different listeners for each condition all of whom were native speakers of Australian English, none of whom had any experience with synthetic speech, and none of whom had any history of hearing or speech pathology. All subjects were screened with a simple speech discrimination test which ensured that they were reliably able to identify monosyllabic words presented at 40 dB s.p.l. All relevant pairs of conditions were compared using the chi square test and tested for significant difference at the 0.01 level.

3. RESULTS AND DISCUSSION

The intelligibility results for the unmasked conditions are summarised in figures 2 to 5. Since the 0.75 and 1.0 Bark conditions are not significantly different for any phonetic classes the 1 Bark condition is the optimal Bark-scaled condition (ie. fewer filters for no loss of intelligibility) and is used in the following discussion as the Bark-scaled reference condition. The 100 Hz condition is significantly higher in intelligibility than the 200 Hz condition for some phonetic classes and so the 100 Hz condition is considered the optimal Hz-scaled condition.

The difference in vowel intelligibility between the natural condition and the 100 Hz and the 1 Bark conditions is not significant when presented unmasked but the performance of masked 100 Hz vowels is significantly lower than the performance of both natural and 1 Bark vowels. The formant vocoded vowels are moderately, but significantly, lower in intelligibility than the natural vowels. Further, the formant vocoded vowels are significantly lower in

intelligibility than the 100 Hz vowels which are in turn not significantly different from the 1 Bark vowels. It should be noted that intelligibility significantly deteriorates (for the vowels) when the frequency resolution drops below 200 Hz and 1.0 Bark.

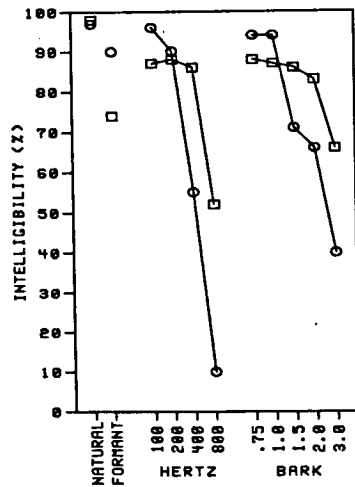


FIGURE 2. ○ ALL VOWELS
□ ALL CONSONANTS

The intelligibility of the formant vocoded consonants is considerably (and significantly) lower than that of the natural, 100-200 Hz and 0.75-1.0 Bark consonants. There is no significant difference between the 100 Hz and the 1 Bark unmasked consonant conditions. Consonant intelligibility does not deteriorate significantly up to 400 Hz and up to 2.0 Bark indicating that consonants are generally able to withstand poorer frequency resolution than are vowels.

In figure 3 it is clear that the intelligibility of channel vocoded affricates is unimpaired at all bandwidths whilst the formant vocoded affricates are significantly and markedly lower in intelligibility compared with all other conditions. The stops show a somewhat unpredictable pattern but there is little significant difference between the voiceless and voiced stops and so they will be dealt with here as a single class. Generally all vocoded stops are lower in intelligibility than the natural condition. The formant vocoded stops are generally lower in intelligibility than the channel vocoded stops, but not significantly lower than the 100 Hz and the 1 Bark conditions. Masked 100 Hz voiceless stops are significantly lower in intelligibility

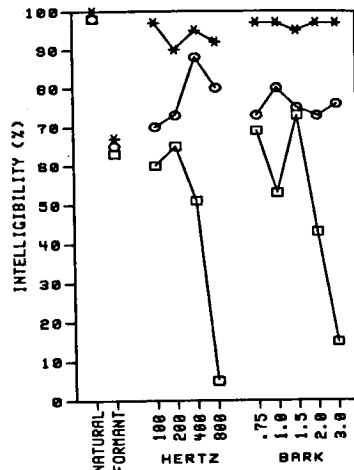


FIGURE 3. * AFFRICATES
○ VOICELESS STOPS
□ VOICED STOPS

than masked 1 Bark voiceless stops. It is interesting to note that the voiceless stops do not deteriorate in intelligibility with increasing bandwidth whilst voiced stops do.

In figure 4 it can be seen that there is no significant difference in intelligibility for both voiceless and voiced fricatives between the natural condition and the 100-400 Hz and the 0.75-2.0 Bark conditions. The formant vocoded fricatives, on the other hand are significantly lower than the natural and the channel vocoded fricatives in intelligibility. There is no difference between the best Hz-scaled and Bark-scaled conditions for both unmasked and masked presentation. Only the voiced fricatives are effected by increasing bandwidth and only for the 800 Hz condition.

Figure 5 details the nasal and approximant results which display very similar patterns to each other and be treated in the following comments together. Firstly, (and predictably) the intelligibility curves behave very much like that of the vowels as bandwidth increases. The best Hz-scaled and Bark-scaled conditions are not significantly different from the natural intelligibility, and intelligibility drops off significantly when bandwidth exceeds 400 Hz or 2 Bark. The formant vocoded condition is not significantly less intelligible than natural or the best channel vocoded conditions when heard unmasked but is significantly lower in intelligibility than natural tokens when heard

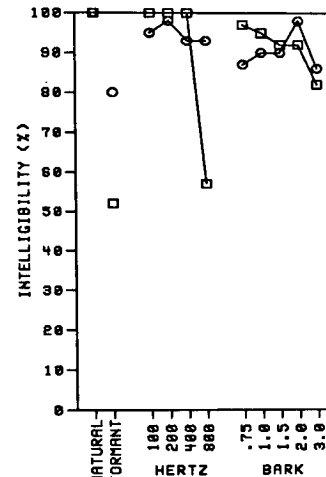


FIGURE 4. ○ VOICELESS FRICATIVES
□ VOICED FRICATIVES

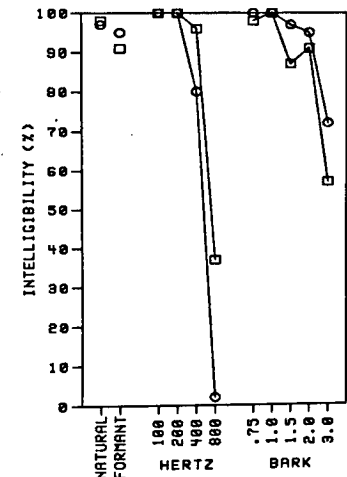


FIGURE 5. ○ NASALS
□ APPROXIMANTS

masked. The 1 Bark condition is not significantly less intelligible than natural for both nasals and approximants when presented either unmasked or masked. This is also true for the 100 Hz condition but with the exception that masked approximants are significantly lower in intelligibility than masked natural approximants.

4. CONCLUSIONS

This study presents evidence that channel vocoders with a 1 Bark bandwidth filterbank perform significantly better than formant vocoders. A 1 Bark filterbank vocoder is equivalent in intelligibility to natural speech for all phonetic classes except the stops and the approximants (and in the second case only when masked). It also performs marginally better than the 100 Hz filterbank vocoder (this is only evident when tokens are masked). It cannot be stated with complete confidence that formant systems are inherently less able to parametrically encode speech than 1 Bark channel vocoders (and channel vocoders in general) however this study supports that conclusion and there does not seem to be any evidence in the literature to support the opposite conclusion (particularly with respect to consonants). It seems likely that a 1 Bark channel vocoder has the intrinsic ability to adequately encode phonologically relevant parametric detail with sufficient accuracy to produce intelligibility approaching, or equal to, natural speech.

5. REFERENCES

- [1] CLARK, J.E. (1983), "Intelligibility comparisons for two synthetic and one natural speech source", *J. Phon.* 11, 37-49.
- [2] CLARK, J.E. MANNELL, R.H., & OSTRY, D. (1987), "Time and frequency resolution constraints on synthetic speech intelligibility", *Proc. XI ICPHS*, Tallin Estonia, Aug. 1-7, 1987.
- [3] CLARK, J.E., DERMODY, P. & PALETHORPE, S. (1985), "Cue enhancement by stimulus repetition: Natural and synthetic speech comparisons", *J.A.S.A.* 78, 458-462.
- [4] DUDLEY, H. (1939), "Remaking speech", *J.A.S.A.* 11, 169-177.
- [5] DUPREE, B.C. (1978), "Automatic formant analysis", *Proc. Institute of Acoustics*, Spring meeting, Cambridge.
- [6] LOGAN, J.S., PISONI, D.B. & GREENE, B.G. (1985), "Measuring the segmental intelligibility of synthetic speech: results from eight text-to-speech systems", *Research on Speech Perception*, Progress Report No. 12, 319-334.
- [7] PISONI, D.B., NUSBAUM, H.C., & GREENE, B.G. (1985), "Perception of synthetic speech generated by rule", *Proc. IEEE* 73, 1665-1675.
- [8] RYE, J.M. & HOLMES, J.N. (1982), "A versatile software parallel-formant speech synthesiser", *JSRU Research Report No 1016*

PARSING UNRESTRICTED TEXT: A MULTI-PHASE APPROACH

A. I. C. Monaghan

Centre for Speech Technology Research, University of Edinburgh,
80, South Bridge, Edinburgh EH1 1HN, Scotland.

ABSTRACT

This paper describes work done on the Edinburgh University CSTR text-to-speech (TTS) system, a linguistically sophisticated speech output system based around a morph lexicon and a complex morphological decomposition module. The major problem with this and many other TTS systems is the lack of a reliable syntactic parse: this paper outlines a strategy designed to remedy that problem. The approach described here is, of course, still to be proven in application, but it is intended to produce a practical, efficient and flexible parsing strategy for unrestricted text by combining the best of both statistical and linguistic approaches.

1. PARSING FOR TTS

From the very crude linguistics-based island parsing of MITalk [1] to the highly-sophisticated statistical knowledge used in systems such as CLAWS and UCREL [2], almost every conceivable combination of parsing techniques has been applied to the problem of analysing unrestricted text. Until very recently, however, the criteria for deciding what parsing techniques would be implemented in a given TTS system had more to do with the researchers' interests in syntax than with the requirements of text-to-speech conversion: it is only in the last couple of years that some TTS workers [3,4] have advocated sacrificing full syntactic parsing in order to achieve efficient extraction of the information most important to TTS systems, and even this work has so far concentrated in each case on applying a particular parsing technique which has been roughly tailored to the perceived needs of a TTS system. There is an important unanswered question at the root of this

approach: what does a TTS system require from syntax? The obvious things are word-class disambiguation ("Is it a noun or a verb?") and syntactic dependencies ("What does this NP dominate?" "Does the PP go with the noun or the verb?"): the former is required to determine stress and pronunciation for many orthographic forms, and the latter is assumed to be crucial for assigning prosody. However, it is clear that neither disambiguation nor dependency relations can be obtained from a purely syntactic analysis. For example, given an NP such as *The damned* no amount of syntactic analysis can determine with certainty whether *damned* is a noun, an adjective or a verbal participle: it is an arbitrary choice depending on which rule the parser finds first. Similarly, the well-known example sentence *I saw the man in the park with the telescope* demonstrates the impossibility of assigning PPs a place in a syntactic tree on any principled basis, and hence the impossibility of determining what the NPs are. These are serious problems for any full parse which relies on deterministic rules, but the point is that for any TTS system currently under development it doesn't much matter which of the possible analyses is chosen: the overall performance of the system will not be significantly altered. Unrestricted text includes much more problematic examples than these, of course, but the argument put forward here is that, at least until the end of this century, they are not worth worrying about.

What of the claim that prosody will suffer if such cases are not resolved? Aside from the fact that they CANNOT be resolved by syntax, there are various reasons for believing that the "correct"

syntactic structure is not essential for producing good prosody. Firstly, it is widely accepted that prosodic structure is much flatter than syntactic structure [5,6]. There must therefore be levels of structure in syntactic analyses which are not relevant to prosody, i.e. a many-to-one syntax-to-prosody mapping, and if these levels are omitted in the syntactic analysis there will be no corresponding degradation in the prosodic realisation. Secondly, there is the long-standing problem of the one-to-many syntax-to-prosody mapping [6,7] which results in different accent patterns on what is syntactically the same sentence and provides the basis for the less-widely-accepted view that syntax has very little to do with prosody [7,8] which is given at least lip service in many TTS systems [3,9,10]. According to this view, it is pragmatics and semantics which determine prosody and any correlation with syntax is an artefact of the syntax-semantics correlation. Thirdly, the syntax of spontaneous speech is known [11] to be much less complex and varied than that of written text: it is not clear that human readers actually realise the types of syntactic structure which can be found in technical writing, and indeed professional broadcasters make frequent errors in reading aloud from even moderately complex material with which they are unfamiliar. To what extent TTS systems should be expected to cope with text which was not designed to be spoken is a difficult question, but it is obviously unrealistic to expect them to perform better than humans and it may well be that users of any successful TTS system would simply not produce such text.

In any case, speech output systems must be able to respond reasonably to any input if they are to claim unrestricted applicability: even if this does not mean that they should read T. S. Eliot's poetry as well as Eliot himself would have done, it does mean that the parser should be failsafe and that the information available at every stage should be exploited to the full. The strategy which is proposed in the remainder of this paper is an attempt to do just that, and is justified if at all on purely pragmatist grounds.

2. MULTI-PHASE PARSING

Given that the information necessary to produce a "perfect" acoustic realisation of a text sentence is not available to automatic systems, and given that the system must produce as acceptable a realisation as possible for any input text without ever failing altogether, it seems obvious that a simple phrase-structure parser will not suffice: such parsers are quite capable of failing for trivial reasons, and are prone to serious errors if given word-class-ambiguous input of the type generally produced by TTS systems. There is also the question of punctuation, abbreviations, and other non-words. These problems need to be handled before any phrase-structure analysis is attempted, i.e. by some sort of pre-processor, so that the parse is guaranteed not to fail. As was stated above, a purely syntactic analysis cannot determine the attachment of constituents such as PPs or adverbs, and so this level of structure must also be supplied by heuristics. The final analysis must then be passed to phonetic or phonological modules, and there are bound to be elements of structure which the syntax has built up but which are irrelevant to the flatter, more linear prosodic structure: some sort of post-parse interface is therefore needed to ensure that the syntactic information is passed on with minimal redundancy. These observations are the basis for the CSTR multi-phase parsing strategy, which includes the following components:

PRE-PARSER: This phase makes use of reliable collocational and other statistical or heuristic information to remove needless word-class ambiguity (e.g. noun/verb ambiguity after determiners, main/aux verb ambiguities), and recognises and pre-process elements which the parser cannot handle (sentential adverbs, impermissible sequences (e.g. determiner + verb), clitics, punctuation, etc.). It is essential that the input to this phase from dictionaries, morphology, text pre-processors, etc. is optimised: for example, the word class of *damned* in (1) above could conceivably come out of some morphological analysis as four-ways (or more) ambiguous (MAINVERB, PARTICIPLE, ADJECTIVE, NOUN) but in view of the practical limitations of the parser it is advisable to apply a

morphological rule along the lines of

MAINVERB + ed --> EDFORM which would produce the unambiguous wordclass EDFORM as output and leave the parser to determine what type of phrase will be built from the syntactic context. An initial version of this pre-parser is implemented in our current system. The output of this phase should be guaranteed not to produce fatal errors in subsequent phases, and the pre-processed elements should be passed without further processing to the post-parse phase.

PHRASE-LEVEL PARSE: This needs to be failsafe, so it must be kept simple. An intelligent control structure (i.e. not just ordered rewrite rules) would also be advantageous: the problem of disambiguating noun/verb ambiguous items so as to identify the correct VPs requires a solution in terms of search strategies and control structure (stated informally, "Use breadth-first search, and look for verbs before nouns."). The depth of embedding, order of search, and so forth will ideally be variable, at least for development purposes until the rules have been satisfactorily tuned. The main purpose of this phase is to parse its input unambiguously into minimal constituents (NP, VP, PP). Each constituent may contain only one possible head, so that a sequence of three nouns produces three separate NPs: the general principle to be observed is that no spurious structure should be generated at this stage which has to be dismantled by subsequent processes. This phase is currently being implemented as a set of phrase-structure rules taking wordclass-ambiguous input and producing a string of constituents spanning the input in which all word-classes have been disambiguated. The disambiguation of word-class is determined largely on the basis of frequency information, in that if the most frequent word-class for that item results in a possible parse then that word-class will be taken. The phrase-structure rules are intentionally limited in coverage, so that only the most common phrase-types of English are covered (all other constructions must be handled by later phases) and distinctions such as that between adjectives and participles are not preserved: we have found [10] that the effect of such distinctions on prosody is negligible, whereas their effect on parsing time is considerable.

CLAUSE-LEVEL PARSE: This will probably be based more on statistical than on syntactic knowledge, but given a phrase-level parse of the sort detailed above we can certainly make an intelligent guess at the location of clause boundaries. Together with a heuristic approach to the construction of major phrases and their attachments, a flattened clause structure will be produced. This phase will use information such as punctuation and verb subcategorisations, and can be as simple or as complex as is practical, although it must be failsafe and sensitive to the capabilities of the prosody modules. The first step is to collapse the minimal phrases identified in the phrase-level parse into larger phrases, and then the head of the clause must be identified: finally, pre- and post-modifiers will be attached according to subcategorisation information and general default rules. An initial version of this is under development, based on the assumption that there is one VP per clause. This is the phase where PP attachment and compounding are performed. Our current heuristic approach to PP attachment is simply to attach PPs in linear order and as low as possible in the tree, and this seems to produce reasonable prosody most of the time. *N*-noun compounds are a more serious problem, as identifying the head is virtually impossible except on a per-case basis [12] and yet incorrect accent placement results in very low acceptability of output. Our current approach involves a version of the Compound Stress Rule [14] with various exception clauses.

POST-PARSE: This phase is necessary to ensure compatibility between syntax and prosody. Its main purpose is to remove any prosodically-irrelevant syntactic structure (e.g. internal structure of adjective phrases, complex prepositions and the like) which has been built up during parsing, and to ensure that the structure which is passed on to the prosodic rules is concise and coherent. This phase also slots adverbs, abbreviations, etc. back into place on the basis of the original linear order. The post-parse phase will subsume our current syntax-intonation mapping rules [15], and will contain additional rules to integrate discourse-level information whenever this is available. The eventual shape of this phase depends to a very large extent on the details of the

prosodic processing which it feeds, so that notions of prosodic well-formedness and statistical heuristics would be equally appropriate in, say, determining at what level adverbs were attached.

3. CONCLUSIONS

The multi-phase parsing strategy discussed above is presented as an alternative to the single-pass or double-pass, purely linguistic or purely statistical parsers common in TTS systems. This strategy is designed both to maximise the use of linguistic and statistical knowledge at each phase and to be a development tool which can be extended as and when required: many phases are under construction already, and the other elements can be functioning/deliverable in a very short time but will allow for development over a longer term.

The pre- and post-parse stages above are clearly highly application-specific, in that they serve as interfaces between the syntactic processing and a specific system: the other phases, however, are seen as application- and domain-general, being limited to a core syntax and being relatively unambitious in the structures they produce. It is therefore anticipated that this strategy could be applied to any TTS system with the minimal problems of designing specific interfaces.

We make no apologies for the lack of theoretical syntactic motivation in this presentation: in our view, the role of syntax in TTS systems is largely as a woefully inadequate substitute for semantic and pragmatic analyses. We therefore consider the mixing of different approaches to parsing as perfectly justifiable insofar as they complement each other, and in the absence of discourse information we believe it is essential for a high-quality speech-output system to make use of all the available knowledge sources in analysing written text.

4. REFERENCES

- [1] ALLEN, J., S. HUNNICUTT & D. KLATT (1987), *From Text to Speech: The MITalk System*. Cambridge: CUP.
- [8] BOLINGER, D. (1986), *Intonation and its Parts*. Stanford: University Press.
- [11] BROWN, G., A. ANDERSON, R. SHILLCOCK & G. YULE (1984), *Teaching Talk*. Cambridge: CUP.
- [14] CHOMSKY, N. & M. HALLE (1968), *The Sound Pattern of English*. New York: Harper & Row.
- [2] GARSIDE, R., G. LEECH & G. SAMPSON (eds) (1987), *The Computational Analysis of English: A Corpus-Based Approach*. London: Longman.
- [7] LADD, D. R. (1980), *The Structure of Intonational Meaning: Evidence from English*. Bloomington: Indiana University Press.
- [15] MONAGHAN, A. I. C. (1989), "Phonological Domains for Intonation in Speech Synthesis", *Proceedings of Eurospeech 1989*, vol. 1 pp. 502-505.
- [10] MONAGHAN, A. I. C. (1990), "Rhythm & Stress Shift in Speech Synthesis", *Computer Speech and Language* 4 (1), pp. 71-78.
- [5] PIERREHUMBERT, J. B. (1980), *The Phonology and Phonetics of English Intonation*. Ph.D. dissertation, MIT.
- [9] QUAZZA, S., G. VARESE & E. VIVALDA (1989), "Syntactic Pre-Processing for High Quality Text-to-Speech", *Proceedings of Eurospeech 1989*, vol. 1 pp. 506-509.
- [3] QUENE, H. & R. KAGER (1989), "Automatic Accentuation and Prosodic Phrasing for Dutch Text-to-Speech Conversion", *Proceedings of Eurospeech 1989*, vol. 1 pp. 214-217.
- [6] SELKIRK, E. O. (1984), *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, Mass.: MIT Press.
- [12] SPROAT, R. W. & M. Y. LIBERMAN (1987), "Toward Treating English Nominals Correctly", *Proceedings of the 25th Annual Meeting of the ACL*, pp. 140-146.
- [4] WILLEMSE, R. & L. BOVES (1989), "Context Free Wild Card Parsing in a Speech-to-Text System", *Univ. Nijmegen Dept. of Language & Speech (Phonetics) Proceedings* 13, pp. 65-81.

SESSIONS ORALES / ORAL SESSIONS

SESSION 1 : Production

(Théories, modèles, méthodes / Theory, models, methodology)

- 1 **Préparation motrice et sélection de cibles articulaires accentuées.**
Lise Crevier-Buchman, J.F. Bonnot, Claude Chevrier-Muller, Catherine Arabia-Guidet 2:2
- 2 **The distinction of central and peripheral speech timing mechanisms**
Eric Keller 2:6
- 3 **Constraints on the behavior of the tongue body : vowels and alveolar stop consonants.**
Simon Levy 2:10
- 4 **Word- and phrase-level aspects of vowel reduction in Italian.**
Edda Farnetani, Mario Vayra 2:14✕
- 5 **Production des voyelles et modèle à régions distinctives.**
René Carré, Mohamed Mrayati 2:18
- 6 **Correcting erroneous information in spontaneous speech : cues for ASR.**
Peter Howell 2:22
- 7 **Prosodic effects on articulatory gestures. A model of temporal organization.**
Osamu Fujimura, Donna Erickson, Reiner Wilhelms 2:26
- 8 **Stuttering as indication of speech planning.**
Marina Koopmans, Iman Slis, A. Rietveld 2:30
- 9 **MRI (Magnetic Resonance Imaging) for filming articulatory movements.**
Arne-Kjell Foldvik, Olaf Husby, Jorn Kvaerness, Ingrid C. Nordli, Peter A. Rinck 2:34

SESSION 2 : Production
(Analyse et description / Analysis and description)

- 1 **Stability of voice frequency measures in speech.**
William Barry, Michael Goldsmith, Adrian Fourcin, Hilary Fuller 2:38
- 2 **Electromyographic study on Laryngeal adjustment for whispering.**
Koichi Tsunoda, Seiji Niimi, Hajime Hirose, Katherine S. Harris, Thomas Baer 2:42
- 3 **The intrinsic fundamental frequency of vowels and the activity in the cricothyroid muscle.**
Niels Dyhr 2:46
- 4 **Laryngeal and oral gestures in English /P, T, K/**
André M. Cooper 2:50
- 5 **Devoicing of Japanese /u/ : an electromyographic study.**
Zyun'ici B. Simada, Satoshi Horiguchi, Seiji Niimi, Hajime Hirose 2:54
- 6 **Is subglottal pressure a contributing factor to the intrinsic F0 phenomenon ?**
Erkki-Antero Vilkman, Ilkka Raimo, Olli Aaltonen 2:58
- 7 **Modelling of speech motor control and articulatory trajectories.**
Pascal Perrier, Raphaël Laboissiere, Laurent Eck 2:62
- 8 **Trying to determine place of articulation of plosives with a vocal tract model.**
Alain Soquet, Marco Saerens, Paul Jospa 2:66
- 9 **Complex nature of the seemingly simple vocal fold cycle.**
Josef Pesak 2:70

SESSION 3 : Perception

(Théories, modèles, méthodes / Theory, models, methodology)

- 1 **An investigation of locus equations as a source of relational invariance for the stop place dimension.**
Harvey Sussman 2:74
- 2 **Modelling speech perception in noise : a case study of the place of articulation feature.**
Abeer Alwan 2:78
- 3 **Mechanisms of vowel perception : evidence from step vowels.**
Frank Gooding 2:82
- 4 **Two processing mechanisms in rhythm perception.**
Morio Kohno, Asako Kashiwagi, Toshihiro Kashiwagi 2:86
- 5 **Lexical stress in a "stressless" language : judgments by Telugu-and English-speaking linguists.**
Leigh Lisker, Bh. Krishnamurti 2:90
- 6 **Connectionist models of Speech Perception.**
Dominic W. Massaro 2:94
- 7 **Perception of spectrally compressed speech.**
Richard Hurtig 2:98
- 8 **A model of optimal tonal feature perception.**
David House 2:102
- 9 **On the perception of duration of the Czech vowels.**
Jevgenij Timofejev 2:106

SESSION 4 : Aspects Linguistiques/Linguistic Aspects Phonologie / Phonology

- 1 **Differentiating between phonetic and phonological processes : the case of nasalization.**
Maria-Josep Solé, John J. Ohala 2:110
- 2 **Spirantisation, charme et gouvernement : l'identité et les métamorphoses de [g] en mooré.**
Georges Hérault 2:114

- X3 About the phonetics / phonemics interface : the case
 of Kriol stress.
Bernard Laks, Alain Kim 2:118 X
- X4 Problèmes de syllabation automatique du français.
Marc Klein, Bernard Laks 2:122 X
- 5 A propos de "h" final en coréen.
Jung Won Lee 2:126
- X6 Underspecification and phonological assignment of
 phonetic strings : the case of classical mandaic
 [qen:a:] 'nest'.
Joseph L. Malone 2:130 X
- 7 Low level phonetic implementation rules : evidence
 from Sindhi.
Paroo Nihalani 2:134
- 8 Evidence for final devoicing in German ? An
 experimental investigation.
*Hans Georg Piroth, Lieselotte Schiefer, Peter Michael Janker,
 Birgit Johnne* 2:138
- 9 Sound distinction : universal inventories of phonic
 substance or language-specific systems ?
Vulf Y. Plotkin 2:142

SESSION 5 : Aspects linguistiques / Linguistic Aspects
 Typologie et universaux / Typology and universals

- 3 Medieval and early modern English systems of
 vowel order : from alphabetic to organic schemes.
Horst Weinstock 2:146
- 4 Evaluation quantitative de l'alternance phonétique
 du /ə/. Importance de l'entourage consonantique.
Joëlle Van Eibergen 2:150
- 5 Typology of Germanic morphosyllabism.
Yuri Kuzmenko 2:154

- 6 **Isomorphous and allomorphous characteristics of the Caucasian and some Indo-European languages in the field of phonetics and phonology.**
Lily Alexandrovna Ponomarenko 2:158
- 7 **The rhythmic organization of speech in Slavonic languages.**
Vsevolod Potapov 2:162
- 8 **The typology of speech segment units.**
Rodmonga Potapova 2:166
- 9 **Phonological component in the quantitative language typology.**
Ludmyla Zubkova 2:170

SESSION 6 : Sociophonétique / Sociophonetics

- 1 **Neutralisation des voyelles nasales chez des enfants d'Ile de France.**
Isabelle Malderez 2:174
- 2 **Evaluation of voice and pronunciation characteristics of men and women.**
Mirjam Tielen, Florian J. Koopmans-van Beinum 2:178
- 3 **Social distribution of long-term average spectral characteristics in Vancouver English.**
John H. Esling, Bernard Harmegnies, V. Delplanq 2:182
- 4 **Observations sur la chute du L dans le français de North Bay (Ontario).**
Jeff Tennant 2:186
- 5 **Foreign accent and the native speaker.**
Una Cunningham-Andersson 2:190
- 6 **Evolution de l'accent méridional en français niçois : les nasales.**
Alain Thomas 2:194
- 7 **Identifying foreign languages.**
Zinny S. Bond, Joann Fokes 2:198

- 8 **Comprehension of vocalizations across species.**
Reijo Aulanko, Lea Leinonen, Ilkka Linnankoski, Maija Laakso 2:202
- 9 **Speech types, speech culture and their segmental correlates.**
Natalja Geilman 2:206

SESSION 7 : Prosodie / Prosody
Accents et tons / Stress and tone

- 1 **Moraic nasal and tonal manifestation in Osaka Japanese : implications for the representation of Mora.**
Yasuko Nagano-Madsen 2:210
- 2 **Durational complementation within the syllable in Chinese and Sui.**
Shi Feng 2:214
- 3 **Phonological interpretation of Fo variations in a Bantu language : Kinyarwanda.**
Thierry Chambon 2:218
- 4 **Interaction between suprasegmental features.**
Jialu Zhang 2:222
- 5 **Temporal location of stress cue and its nature.**
Asoke Kumar Datta 2:226
- x 6 **Downstep et downstep.**
Annie Rialland 2:230 x
- 7 **Dephonologization of syllabic intonation in Lithuanian urban sociolects.**
Laima Grumadiene 2:234
- 8 **Beats and binding laws instead of the syllable.**
Katarzyna Dziubalska-Kolaczyk 2:238
- 9 **Syllable tonemes in Latvian.**
Dace Markusa 2:242

SESSION 8 : Prosodie /Prosody
Intonation

- 1 Modelization of intonation patterns in Spanish for automatic recognition.**
Juan Maria Garrido Alminana, Francesc Gudayol i Portabella 2:246
- 2 Intonation and ambiguity.**
Maria Carmen Fernandez Leal 2:250
- 3 Exploiting the secondary accent in a prosodic model for French synthesis.**
Valérie Padeloup 2:254
- The acoustic characteristics of boundaries used in uttering telephone numbers in Mandarin Chinese, Japanese and English.**
Yoshimasa Tsukuma, Junichi Azuma 2:258
- Fo-declination in German declarative utterances.**
Isolde Wagner 2:262
- The representation of intonation in Mandarin Chinese.**
Jialing Wang 2:266
- Prosodic Italics : functions and phonetic realization.**
Alexander Panasyuk, Irina Panasyuk 2:270
- Locutions phraséologiques intonatives.**
Natalia Svetozarova 2:274
- 9 Interférences phonétiques au cours de l'apprentissage du français (groupe linguistique slave).**
Nadejda Evtchik, Galina Roudzit 2:278

SESSION 9 : Prosodie / Prosody
Organisation temporelle et rythme / Timing and rhythm

- 1 FO and the perception of duration.**
Wim Van Dommelen 2:282

- 2 **Durational shortening and anaphoric reference.**
W. N. Campbell 2:286
- 3 **Theories of prosodic structure : evidence from syllable duration.**
D. R. Ladd, W.N. Campbell 2:290 X
- 4 **On vowel quantity and post-vocalic consonant duration in Dutch.**
Allard Jongman, Joan A. Sereno 2:294
- 5 **Rhythmic categories : a critical evaluation on the basis of Greek data.**
Amalia Arvaniti 2:298
- 6 **On the acquisition of segmental duration in normal and articulation disordered 4-year-olds.**
Jacqueline Bauman-Waengler, Hans-Heinrich Waengler 2:302
- 7 **Les contraintes temporelles des types consonantiques sur le timing mandibulaire de la quantité en arabe tunisien.**
Fatma Boussaffa, M. Jomaa, Rudolph Sock 2:306
- 8 **The durational patterns of syllables in standard Chinese.**
Jianfen Cao 2:310
- 9 **The Rhythm of Tanka, Short Japanese poems : read in prose style and contest style.**
Yayoi Homma 2:314

SESSION 10 : Applications
Didactique / Foreign language teaching

- 1 **Lateral consonants production in bilingual speakers learning a third language.**
Joaquim Llisterri, Gemma Martinez-Dauden 2:318
- 2 **Pour que l'enseignement des langues étrangères donne les moyens de les comprendre à l'étranger.**
Regina Llorca 2:322

- 3 **A comparison of English and Greek alveolar fricatives.**
Lefteris Panagopoulos 2:326
- 4 **La valeur et les fonctions de la pause dans le parler oral monologique spontané dans la langue étrangère et dans la langue maternelle.**
Marina Avdonina 2:330
- 5 **Learning English segments with two languages.**
Mohamed Benrabah 2:334
- 6 **Problématique de l'enseignement de la prosodie du russe aux francophones.**
Michel Billières 2:338
- Contrastive phonetics and teaching foreign language pronunciation : theory and practice.**
Mukhamedjan Kireevich Isaev 2:342
- Les modèles rythmiques et la fréquence des substantifs avec l'accent mobile en russe.**
Elena Jasova 2:346
- Towards designing an intonation training device based on speech signals clustering.**
Leonid A. Kanter, Alexander V. Savin, Ksenia G. Guskova 2:350

SESSIONS AFFICHEES / POSTER SESSIONS

SESSION 11 : Production

- 1 **Cross-sectional tongue movement and tongue-palate movement patterns in [s] and [ʃ] syllables.**
Maureen Stone, Alice Faber, Marc Cordaro 2:354
- 2 **An examination of the jaw's contribution to lingual stability.**
Eric Vatikiotis-Bateson, Maureen Stone, Michael Unser 2:358
- 3 **An articulatory investigation of front rounded and unrounded vowels.**
Philip Hoole, Hans-G. Tillmann 2:362
- 4 **Tract shapes of tense and lax vowels : a comparison of x-ray microbeam and EMG data.**
Katherine S. Harris, Eric Vatikiotis-Bateson, Peter J. Alfonso 2:366
- 5 **Modeling vowel articulation/modélisation de l'articulation des voyelles.**
Michel Tam Tung Jackson 2:370
- 6 **What does the laryngograph measure ?**
David Miller, S. Nevard, Remi Brun, Adrian Fourcin 2:374
- 7 **Des paramètres formantiques au profil articulatoire.**
Paul Jospa 2:378

SESSION 12 : Acoustique / Acoustics

- 1 **A functional model of dynamic characteristics of formant trajectories.**
Satoshi Imaizumi, Hiroshi Imagawa, Shigera Kiritani 2:382
- 2 **Comparison of the modified hermite transformation with other unitary transformations in a predictive transform coding scheme.**
Victoria E. Sanchez, Jose C. Segura, Antonio M. Peinado, Juan M. Lopez, Antonio J. Rubio 2:386

- 3 **Numerical simulation of the glottal flow and glottal excitation.**
Gabriele Hegerl 2:390
- 4 **Speech production and chaos.**
Hans Peter Bernhard, Gernot Kubin 2:394
- 5 **Effects of language change on voice quality. An experimental study of Catalan-Castilian bilinguals.**
Marielle Bruyninckx, Bernard Harmegnies, Joaquim Llisterra, Dolors Poch-Olive 2:398
- 6 **A new high resolution time-bark analysis method for speech.**
Unto-Kalervo Laine 2:402
- Extracting nasality from speech signals.**
Henning Reetz 2:406
- Two-Formant model of the acoustic description of speech articulation.**
Degtyorev Nikolay 2:410
- Acoustic description of the Spanish nasal consonants in continuous speech.**
Maria Jesús Machuca Ayuso 2:414

SESSION 13 : Prosodie / Prosody

- 1 **Disambiguating sentences using Prosody.**
Patti Price, Mari Ostendorf, Stefanie Shattuck-Hufnagel 2:418
- 2 **Signalisation prosodique de la structure informationnelle dans le discours radiophonique en finnois et en français.**
Veijo V. Vihanta 2:422
- 3 **L'organisation rythmique du discours d'une personne bilingue (russe - turc).**
Lidia Ignatkina, Ludmila Shcherbakova 2:426
- 4 **Les marqueurs acoustiques de l'énoncé en français québécois.**
Conrad Ouellon, Claude Paradis, Louise Duchesne 2:430

- 5 **L'accent en français québécois spontané : perception et production.**
Denise Deshaies, Conrad Ouellon, Claude Paradis, Sylvie Brisson 2:434
- 6 **Modelling of Russian intonation : a "contour interaction" based algorithm.**
Einar Meister 2:438
- 7 **Interaction between pauses and secondary word stresses in the rhythmic system of an archaic Bulgarian dialect.**
Petar Vodenicharov 2:442
- 8 **Analyse numérique des tons du vietnamien.**
Ngoc Quach Tuan 2:446
- 9 **La proéminence et la perspective fonctionnelle de la communication /PFC/.**
Jan-Jaroslav Sabrsula 2:450
- 10 **Symmetry and asymmetry in multi-dimensional prosodic system as cues of textual expressiveness.**
Emma Nushikyan 2:454

SESSION 14 : Technologie / Technology

- 1 **Acceptability of several speech pausing strategies in low quality speech synthesis ; interaction with intelligibility.**
Vincent-Johan Van Heuven, Peter J. Scharpff 2:458
- 2 **Phonotactic knowledge acquisition by syllable structure modelling.**
Alessandro Falaschi 2:462
- + 3 **Phonetic data bases for German.**
Klaus Kohler 2:466 ✕
- 4 **Acoustical data base as a tool for the research of vowel systems.**
Anttik Iivonen 2:470
- 5 **An interactive system for language identification**
Vladimir Kuznetsov 2:474

- 6 **Modélisation explicite de la coarticulation pour le décodage acoustico-phonétique : les triplets phonétiques.**
Yves Laprie, François Lonchamp 2:478
- 7 **Workstation for speech analysis.**
Sergey Andreyev, Vladimir Chuchupal 2:482
- 8 **Workstation and signal processing software for experimental phonetics.**
Michael Scheffers, Werner Thon 2:486
- 9 **L'éditeur de signal 'à distance' MAPSIGNAL.**
Alain De Cheveigné 2:490

SESSION 15 : Technologie / Technology
Synthèse de la parole / Speech synthesis

- Speech synthesis computer for the Blind.**
George Losik 2:494
- The automatic Russian text transcriber.**
Elena Ovcharenko, J. Ipatov, S. Stepanova 2:498
- Couplage entre le modèle à deux masses et un modèle analogue du conduit vocal à réflexion : théorie et implantation.**
Loan Trinh Van, Bernard Guerin, Eric Castelli 2:502
- 4 **Synthesis-by-rule for French.**
Gérard Bailly, Mhania Guerti 2:506
- 5 **Modelling articulatory compensation for synthesis by rule.**
Gilles Boulianne, Henrietta Cedergren, Danièle Archambault 2:510
- 6 **Constitution semi-automatique de lexiques de contours prosodiques pour la synthèse à partir du texte.**
Véronique Aubergé, Michel Contini 2:514

- 7 **A multi-linear representation and rule formalism for Phonology and Phonetics in text-to-speech synthesis.**
Sigismund Frenkenberger, Markus Kommenda, Sylvia Moosmüller 2:518
- 8 **A comparison of the intelligibility scores of consonants and vowels using channel and formant vocoded speech.**
Robert Mannell, John Clark 2:522
- 9 **Parsing unrestricted text : a multi-phase approach.**
Alex Monaghan 2:526