SEGMENTAL AND PROSODIC VARIABILITIES IN CONNECTED SPEECH. AN APPLIED DATA-BANK STUDY

GUNNAR FANT, LENNART NORD, ANITA KRUCKENBERG

Dept. of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH), Box 70014
S-100 44 Stockholm, Sweden

ABSTRACT

As a subset of the KTH data bank, we have recorded several subjects reading the same passages from a selection of various texts. We have studied variations in the realization of segmental and prosodic characteristics and to a less extent reading style. Data is reported on the degree of closure of voiced consonants, ambiguities in segmentation and vowel durations. Vowel-consonant contrasts may be highly reduced even in non-weak stress forms. The multi-cued realization of syntactic boundaries are discussed in relation to subjective assessments and to rhythmical structures. In addition to physical pauses, final lenghtening, formant-pattern changes and intonation contours, there are also local voice source features other than F0 to consider, e.g., creaky voice junctures.

INTRODUCTION

Advanced work on text-to-speech synthesis and speech recognition demands a continuous updating, extension and renewal of knowledge from speech analysis. We have to adopt a rule-oriented search to efficiently encode phonetic features, speaker typology and behavior. The data-bank storage and processing system of Carlson and Granström /1/ have provided a format and practical tool. A more complete account of the work is given in a report by Fant et al. /2/ which contains observations of speaker behavior under various conditions, not included here.

Within this limited frame, we have gained a fresh insight in several fundamental acoustic-phonetic problems and a view of what kind of problems we will encounter as the analysis proceeds. We have studied segmentation problems and underlying variabilities in articulatory gestures and, furthermore, the realization of syntactic boundaries, and how subjective juncture assessments correlate with acoustic factors. Our overall impressions is that of a richness of variability on all levels as well as potentials of structuring variabilities. One prevailing impression is that segmentals and prosodics share a common basis of acoustic correlates. Therefore; they should be treated together as seen from an underlying model of speech production. Our study has also provided some limited data on vowel durations and prosodic realizations which can be extended to support the up-dating of our synthesis rules.

When instructing subjects, we laid an emphasis on attaining a neutral but semantically distinct reading. In addition, we have also recordings of more engaged readings, occasional mannerisms and deliberately dramatized versions. However, even in the more normal readings, we observed a rather large span of intonation and overall prosodic patterns. Deviations from average and preferred patterns attain a subjective personality marking which attracts our attention without affecting the overall quality of the reading. It would be of interest to certify which prosodic factors remain intact and which are allowed to vary.

In the present pilot study we have concentrated on 14 subjects' readings of two sentences. Spectrograms and associated oscillograms, intensity and F0 plots were produced by means of our laboratory computer processing routines.

SEGMENTAL STUDIES

We have studied various coarticulation and reduction phenomena that affect the segmental composition of phonemes and complicate the task of boundary assignments.

Boundaries are more clearly realized by changes in "manner" cues than in "place" cues. Thus, it is easy to find the boundary between a fricative and a vowel but we have no clear rules for finding boundaries between vowels or between voiced consonants like /v/, /j/ and /r/ and their combinations with vowels. A voiced intervocalic stop is not always associated with a stop gap, and phonemically unvoiced stops in unstressed positions may attain voicing. Lack of oral closure may affect nasals as well as stop sounds or any consonant, and an incomplete abduction of the glottis in an /h/ causes a continuation of voicing.

In order to understand these ambiguities, we should consider a basic parameter of speech production related to the extent to which vocal-tract constriction targets are reached in connected speech. This parameter which has a strong descriptive power could be labeled "articulatory contrast" or more generally, "dynamic contrast". It affects not only the supraglottal articulators but also the glottal articulation. Thus, a sufficient adduction/abduction contrast is needed for preserving a voiced/voiceless boundary. Also, the boundary between a voiced /h/ and a following vowel becomes obscured by insufficient glottal contrast.

Articulatory contrast implies acoustic contrast in terms of envelope intensity modulation as well as an extended range of formant pattern dynamics. Decreased contrast, thereby, also affects the rate of change of formant patterns at segment boundaries.

Although these phenomena are by no means new in phonetic theory, we had not anticipated the full extent of their realizations. Thus, most speakers did not produce a full closure of the voiced stop /g/ in "legat". For one speaker, LN, the intensity modulation was marginal only and the formant pattern that of a connecting glide, see Fig. 1.

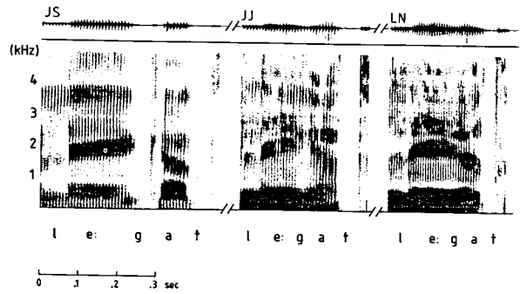

Fig. 1. Three degrees of articulatory contrast. The same word "legat" from three subjects' readings.

This is typical of voiced stops in fluent rapid speech and probably dependent on both the place of articulation and the vocalic context.

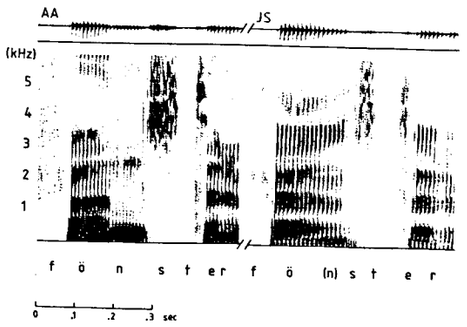

Fig. 2. Two subjects contrasting in oral closure of consonant [n].

Another example of incomplete closure is in the nasal consonant /n/ in the word "fönster" which often is realized by nasalization of the vowel only, see Fig. 2. The appearance of an orally closed segment for the /n/ of "i en" was even less frequent which is to be expected since "en" is a function word. On other occasions we have noticed this effect to be particularly strong in phoneme strings consisting of vowel-nasal-fricative. For American English, Malécot described this phenomenon in word pairs with nasal-homorganic unvoiced stop , like "camp" versus "cap", differing in nasalization only /3/.

The two-word string "han hade" in the initial part of the sentence "Han hade legat och skrivit det i en stor sal vars fönster vette mot Klarälven" is produced out of focus and with higher tempo and reduced articulatory contrast. The /n/ is realized by nasalization only, and the second /h/ is hard to detect being glottally coarticulated with the following vowel /a/. The second /h/, when present, can thus be said to be realized by aspiration of the following vowel. A further complication is that nasalization and aspiration share cues, e.g., the reduced F1 intensity. Only few speakers produced a sequence of clearly identifiable segments.

The Swedish /r/-sound appears in a variety of acoustic realizations ranging from a pronounced trill to a slight /r/ coloring of a neighboring vowel. /r/ also occurs frequently in consonant clusters with subsequent forms of coarticulation and reduction as a result /4/.

In "fönster vette" and "skrivit", the /r/-sounds are often reduced and segmentation becomes a problem. The acoustic cues become diffuse, a brief constriction phase is often found, but not always, and the same is true of the F1, F3 and F4 lowering cues. When present, the constriction phase of /r/ in "skrivit" may mark the right boundary of an inserted vowel after /k/. It may also reside in the unvoiced k-release. Segmentation rules for /r/-sounds are still undefined. Shall we concentrate on the stop gap if it is present or should we choose a larger domain of perceptual importance including a possible inserted vowel or a short segment of the same nature?

With the latter choice, the segmentation principle will deviate from that of handling stop sounds where, by convention, the voiced part of a following transition goes with the next segment. When the acoustic cues become weak, the auditive impression of the /r/ prevails though weakened.

More examples of variabilities of segmental realizations and segmentation ambiguities will be discussed in connection with the study of syntactic boundary regions in the following section.

SYNTACTIC BOUNDARIES AND PROSODICS

Our standard sentence with each word assigned a lexical stress pattern according to SAOB* attains the following structure:

4 3 2 3 2 4 3 2 4 4 4 4 4
HAN HADE LEGAT OCH SKRIVIT DET I EN STOR SAL

4 4 0 3 2 4 3 2 0
VARS FÖNSTER VETTE MOT KLARÄLVEN

This transcription of each word read in isolation is irrelevant to connected speech. Following established notations we transform it into a more realistic form omitting the stress of function words except the pronoun "det" which generally attains the prominence of its substitute.

*Swedish normative word dictionary.

HAN HADE LÄGAT OCH SKRIVIT DET | I EN STOR SAL |
VARS FÖNSTER VETTE MOT KLARÄLVEN

"i en stor sal" is a preposition phrase. Crosses denote grave accent. A vertical short bar denotes acute accent, if above the line, and the secondary syllable of grave accent, if below the line.

One object of the study was to study the realization of the syntactic boundaries before and after the preposition phrase. We found a considerable variation in both acoustic cues and subjective impressions. In a listening test, ten subjects assessed the degree of perceived boundaries on a scale from 0 to 5. The first boundary got an average rating of 2.2 with a standard deviation of 0.8 whilst the second boundary was rated 3.7 with a standard deviation of 0.75 within the jury. The standard deviation between speakers was 1.1 and 2.2, respectively.

The most prominent acoustic cue appeared to be segmental durations. Since the three words "det i en" in several cases merged to a single voiced gross segment [eI] with no clear boundaries, especially not in the formant juncture between /e/ and /I/, we selected an interval from the onset of voicing in the /e/ of "det", to the onset of the /s/ of "stor", thus potentially including final and initial lengthening effects. It should be noted that three of the 14 speakers omitted the /t/ of "skrivit" and produced a voiced stop gap of 40-70 ms duration for the /d/ of "det". The remaining 11 speakers' spectrograms showed an unvoiced stop gap of 70-140 ms duration appropriate for the /t/ plus /d/ with an uncertainty of whether the /d/ was realized at all and, if so, with no obvious boundary towards /t/. According to the sandhi rules of Gårding, /t+d/ are transformed to unvoiced /d/ /5/.

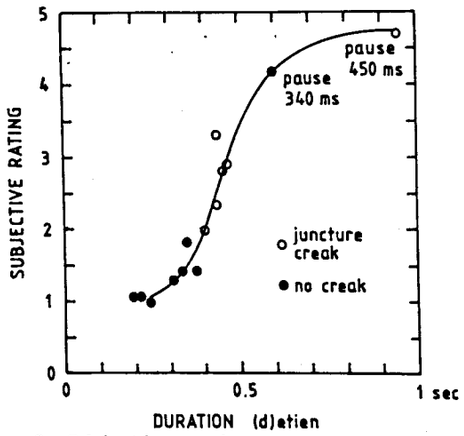

Fig. 3. Subjective rating versus durational measure of the first phrase boundary.

Fig. 3 shows a fair correlation between boundary region duration and the subjective boundary assessment. A tendency may be observed of a doubling of the subjective rating per 200 ms increase of the juncture duration. Deviations from this trend are within the standard deviation of listener judgements. A most apparent trend associated with more marked boundaries is the appearance of creaky voice, i.e., glottalization at the end of the /e/ which causes a local drop of F0 and/or an alternation between strong and weak glottal excitations which is especially apparent in the second and higher formants. These boundary cues have earlier been noted by Gårding /5/, Lehiste /6/ and Kreiman /7/. These alternations may cause an ambiguity in the definition of the local F0. The two speakers of the highest boundary rating produced a proper pause at the phrase boundary. A general phonological rule is to omit the /t/ of "det". Only one of the speakers, EJ, had a proper combination of /t/ unvoiced stop gap + release at the following vowel. Eleven speakers did not have a /t/ stop gap, and the two with proper pause after "det" omitted the /t/.

In absence of glottalization, most speakers produced a level or slightly rising F0 contour at the juncture. An exception was subject BB who had a falling F0 into the beginning of the second phrase. His reading style was in general more personal and affected than others.

As a durational measure, for the second phrase boundary "-----sal vars-----" we selected the /l/+occasional pause+/v/. There were seven subjective ratings between 3.0 and 3.6 with a duration of about 120 ms and six with a rating between 4.2 and 4.4 with associated durations from 200-540 ms. Of the later, four of these included a proper pause and two displayed a brief /l/-release. The F0-contour was mostly a fall+rise at the boundary with some correlation between magnitude of the movement and subjective rating. Here again, subject BB deviated from the rest by an F0 rise+fall. None of the subjects displayed a glottalization. Two subjects, however, had shown such tendencies in earlier informal recordings.

From a second sentence containing a sequence of enumerations, we found similar correlations between subjective boundary impressions and durational measures. Rather constant subjective ratings independent of durational measures were found when a boundary was terminated by strongly stressed syllables on either side.

Our data on vowel durations are summarized in Table I. They are compared with data from Carlson and Granström (ref. /1/) and from the text-to-speech (Rulsys) generated version of our sentence (in May 1986). A correction for overall tempo has been made, the Rulsys sentence being 20% longer. This comparison confirms our awareness of the insufficient contrast between present Rulsys-generated unstressed and stressed short and long vowels, see ref. /2/ for further details. This restricted study can only provide a tendency and more representative data will eventually be gathered.

Table I. Vowel durations in milliseconds

	Short unstressed	Short stressed	Long stressed
Present study	42	105	155
Rulsys	78	93	134
C & G (ref. /1/)	60	90	125

The average reading speed was five syllables or 14 phonemes per second. The standard deviation was rather low, 7% only.

We have looked into the rhythmical structure of the sentence. A rhythmical unit, "stress interval", has been defined as a subpart of the utterance located between the onsets of two successive vowels carrying main stress. Since function words are down graded, the main stresses are confined to content words. We find an overall tendency of two main stresses per second. Similar findings have been made by Goude and Malmström /8/ and Dauer /9/. Distances between vowel onsets in stressed syllables are, thus, of the order of 500 ms but vary with the number of phonemes typically from 350 ms for three phonemes to 600 ms for nine

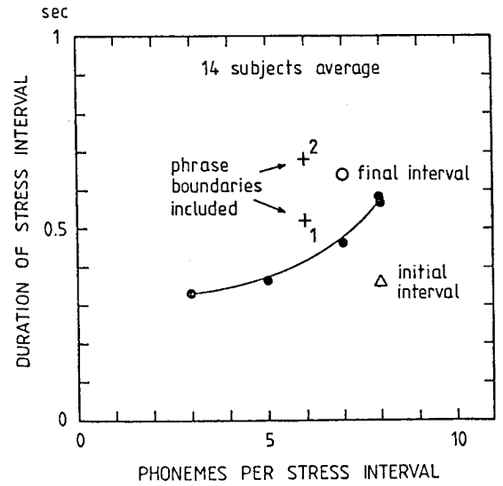

Fig. 4. Duration of stress intervals versus number of phonemes contained. Stress intervals that cut across a phrase boundary are lengthened.

phonemes. The weak tendency of isochrony in reading, see also the study of Strangert, is probably more a matter of constraints in number of phonemes per stress interval than an intention of the reader /10/. As an average for the nine sentences of the central passage, we find two words or eight phonemes per stress interval. It remains to quantify the actual performance of the reading of this and more extensive tests. Even though Fig. 4 refers to a single sentence, it exemplifies typical trends such as the relative weight of stressed vowels and that stress intervals which cut across phrase boundaries are longer whereas the sentence initial group, leading up the first stress, is shorter than within phrase stress intervals /11/. Further studies along these lines might give some insight in reading behavior.

REFERENCES

/1/ R. Carlson, B. Granström, "A search for durational rules in a real-speech data base", Phonetica 43, 140-154, 1986.
/2/ G. Fant, L. Nord, A. Kruckenberg, "Individual variations in text reading. A data-bank pilot study", STL-QPSR 4/1986 (KTH, Stockholm), 1-17.
/3/ A. Malécot, "Vowel nasality as a distinctive feature in American English", Language 36, 222-229, 1960.
/4/ L. Nord, "An acoustic and perceptual study of /r/ varieties in Swedish", forthcoming.
/5/ E. Gårding, "Internal juncture in Swedish, Gleerup, Lund, 1967.
/6/ I. Lehiste, "Perception of sentence and paragraph boundaries", in B. Lindblom, S. Öhman eds., Frontiers of Speech Communication Research, Academic Press, 191-201, 1979.
/7/ J. Kreiman, "Perception of sentence and paragraph boundaries in natural conversation", J. of Phonetics 10, 163-175, 1982.
/8/ G. Goude, S. Malmström, "Ett exempel på experimentalpsykologiskt studium av rytmupplevelse av poesi", University of Stockholm, 1970.
/9/ R.M. Dauer, "Stress-timing and syllable timing reanalyzed", J. of Phonetics 11, 51-62, 1983.
/10/ E. Strangert, "Swedish speech rhythm in a cross-language study", Almqvist & Wiksell Int., Stockholm, 1985.
/11/ I. Lehiste, "Isochrony reconsidered", J. of Phonetics 5, 253-263, 1977.