EFFECT OF PHONETIC CONTEXT AND TIMING ON THE F-PATTERN OF THE VOWELS IN CONTINUOUS SPEECH

Jacqueline Vaissière

Departement RCP
Centre National d'Etudes des Télécommunications
Lannion (France)

ABSTRACT

Very importantly for automatic phonetic decoding of continuous speech in speech recognition systems (SRS), the acoustic contrast (in terms of the observed formant -F- values) between phonemically distinct phonemes can be enhanced, or reduced or even neutralized depending on specific contexts. For a given vowel V, the F-pattern (i.e. F-values and F-temporal evolution), is not only a function of ideal F-target values for each V, but also of (a) the articulatory contrast (i.e. distance) in terms of the tongue position among the phonemes in consonant-V-consonant-vowel sequence and (b) the timing between the required opposite movements of the tongue. The first part of this paper illustrates the observations done on the three most open oral vowels in French, extracted from 1040 sentences spoken by five male speakers. It is concluded that temporal constraints (TC) on the motion of articulators become a prominent factor in continuous speech. The second part concerns the consequences of such observations for SRS: (a) the TC are effective in explaining many of the confusions appearing in the actual SRS, where decision about V identity is mainly based on measuring a "rate- and context-insensitive" acoustic distance between the segment to be recognized and a set of reference templates; (b) some suggestions are provided on how to introduce the TC in the design of the future SRS.

INTRODUCTION

In automatic phonetic transcription of speech, the decision about the identity of the vowels is generally based on the measurement of an acoustic distance between the spectra sampled in the vowel to be recognized and a set of stored (speaker-dependant) reference templates (one or more template for each vowel) is often erroneous. The error rate is known to vary greatly depending on the speaking rate, on the particular speaker and of the number of distinctive vowels in the language. For a language which has a fairly rich vocalic system as French, 50 % of errors is very common. The inadequacy of the speech parametrisation and a poor distance metrics may be partly in fault, but clearly others facts are in cause. As often noticed in the literature, phonemically distinct vowels are not always entirely separated in the F1/F2 plane, when extracted from continuous speech even for a single speaker. This paper explores the formant (F) values of the French three most open vowels in continuous speech, in a fairly large number of sentences.

EXPERIMENTAL PROCEDURES

Two-hundred eight sentences spoken by 7 speakers (5 males and 2 females) were recorded in a sound-proof room at CNET (France), Firstly, high quality digital spectrograms for the 1456 sentences were calculated. Secondly, about 200 selected sentences were submitted to the acoustic-phonetic analyzer developped at CNET [1], to automatically transcribe them into a set of phonemic hypotheses. Thirdly, the sentences were further analyzed using the Spire facilities at the Speech Group at MIT (USA). The sentences were automatically segmented [2] and the F-values were calculated from LPC spectra (See Fig.1 for an example of the material used). The present study only concerns the 1040 sentences sopken by the male speakers: the data obtained for the female speakers were disregarded because of problem in F1 detection.

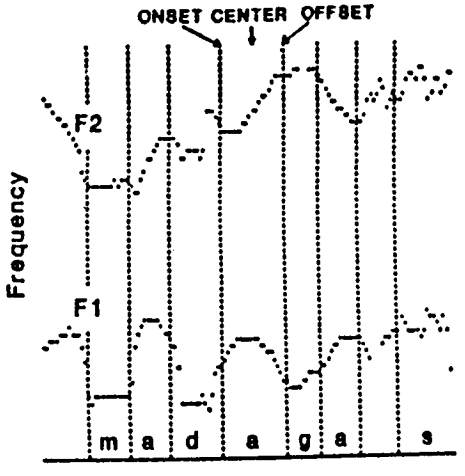

Fig. 1: MATERIAL USED
Display of the superimposed results of the segmentation program (vertical lines) and of the two first formants (F1 and F2) calculated from LPC spectra, every 5 msec.

I. RESULTS

I.1 TYPICAL PLACES OF ARTICULATION

The temporal course and the values of the F's in a V is influenced by the place of articulation of the surrounding phonemes [3]. The 16 French consonants and the three semi-vowels can be regrouped into the four traditionnal classes [4], depending on their places of articulation (See Fig. 2): (1) the bilabial and labio-dental, involving both lips or the lower lip in their production, (2) the dental, involving the tip or the fore part of the blade of the tongue, (3) the prepalatal, mediopalatal and velar, requiring an hightening movement of the dorsum of the tongue toward the hard palate and the velum, and (4) the uvular /r/ (backing of the tongue).

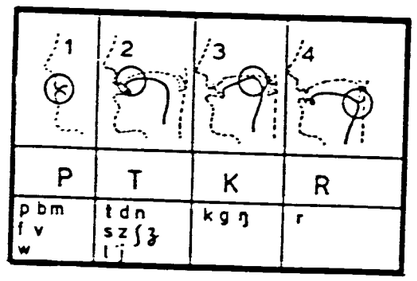

Fig. 2: THE FOUR CLASSES OF CONSONANTS
Classification of the 19 French consonants and semi-consonants into 4 articulatory classes (see text).

The letters P, T, K and R in this paper designate all the labial, dental, velar and /r/ consonants, respectively. A word like /Madagascar/ is thus considered equivalent (for what concerns the effect of the consonants on the vowel) to the sequence PaTaKaTKaR; the French words /femme/, /pape/, /bave/ to PaP, etc... V designates one of the three open vowels.

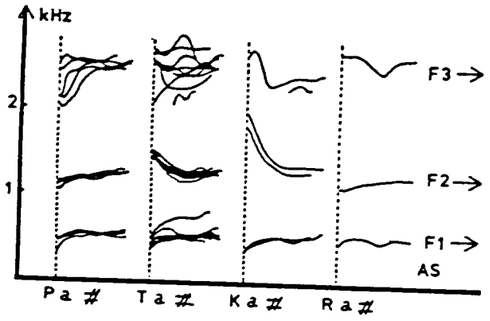

Fig.3: INVARIANT TARGETS
Display of the temporal course of the three first formants, in the vowel /a/, preceded by a labial, dental, velar or uvular consonant and followed by a pause, for the speaker AS. The data have been time-aligned on the vowel onset, found by the automatic segmentation program. The target F-values derived by visual inspection are indicated by arrows on the right part of the figure.

I.2 IDEAL TARGET IN LONG VOWELS

There is a well-known tendency, at least in controlled studies for non-sense sequences of speech, for the values of F1 and F2 of the vowels to be directed asymptotically toward the same target values. Such an ideal target is assumed to exist for each V, which is independant of consonantal context and thus can be rergarded as an invariant attribute of the vowel [5]. This tendency has been confirmed in our study of continuous speech, for vowels with relatively long duration (See Fig. 3 for illustration on /a/ lengthened because of prepausal lengthening).

I.3 THE EFFECT OF SHORTER DURATION

The F1 and F2 patterns for vowels with shorter duration exhibit the following characteristics.

a) Firstly, there is almost no reduction phenomena for V in labial context, even in short or very short occurences of V. In PVP context, the amplitude of the transitional F-movements from C to V and from V to C is small and the observed F-values all along V are closed to the F-target values. Compare the F-values at the F2 turning point in the sequences PaPa with the F-target values indicated by an arrow on Fig.4. Such an observation is expected by the fact that the lips movements for the production of the consonants does not interfere with the tongue movement for the production of the vowel.

b) The amount of reduction (in terms of difference between the F-target values and the observed values) can be predicted from articulatory considerations. When V and the surrounding consonants share about the same place of articulation (i.e. either anterior/front/dental or posterior/back/velar-uvular), the magnitude of F-transitions are reduced. In this case, the F-values observed at the F2 turning point are closed to the F-target values. It is the case for the front (anterior) /ɛ/ in dental (anterior) context (TɛT) and for the back (posterior) /ɔ/ in velar and uvular (posterior) context (KɔK and RɔR).

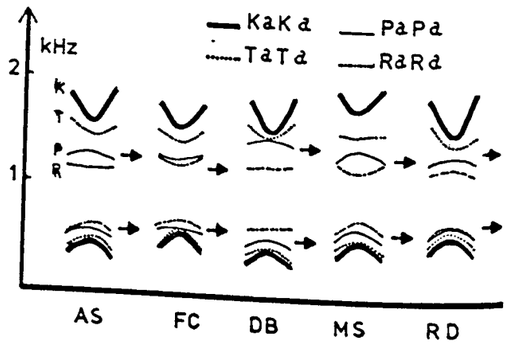

Fig.4: SYMETRICAL CONSONANTAL ENVIRONMENT
Range of the observed of F1, F2 and F3 variations observed in /PVPa/, /TVTa/, /KVKa/ and /rVra/ context for the 5 speakers. For all speakers, F2 is the highest and F1 the lowest in the velar context, and F2 the lowest and F1 the highest in the uvular context.

c) The F-target values are not reached in other cases. Great reduction in terms of differences between the observed F-values and F-target values has been observed for anterior vowel in posterior context (KɛK/ and posterior vowel in anterior context (TɔT). The shorter the vowel, the greater the amount of reduction. In very short occurences of TaT, for example, the values F2 at the vowel center are very closed to the values at the vowel onset and offset. (the "V" shape of F2 temporal course with appears in vowels with longer duration is almost reduced to a straigth line)

I.4 OPENNESS OF THE FOLLOWING VOWEL

The effects described above are not entirely sufficient to account for F-trajectories and values. For example, in a sequence like PaPi, where the closed vowel /i/ is following the open vowel /a/, and the intervocalic consonant is labial, two observations have been made. Firstly, there is always (for all speakers) a rising F2 (as the consequence of fronting of the tongue) Such a rise leads to higher F2 values in the mid region of the vowel, as compared to observed values in a sequence like PaPa. Secondly, there is some tendency of lowering of F1 (as a consequence of a lesser opening of the vocal tract). When the intervocalic consonant is dental (PaTi), the rising of F2 in /a/ is accentued in comparison with /a/ in PaTa.
The higher F2 values in /a/ at center and at offset when the following vowel is /i/ can be interpreted in two traditional ways; first, in terms of a palatalisation of the intervocalic consonant, secondly, as an anticipatory FRONTING-CLOSING due to vowel-to-vowel coarticulation [6].
More data are needed to check in a systematic way an eventual effect of anticipatory labialisation and velarization, on the three open vowels.

II. MODELLING THE INFLUENCES OF CONTEXT

Figure 5 schematizes our observations. The F-values observed at the onset, the mid region and the offset of a given vowel is the a function of the place of articulation of the adjacent consonants and the next vowel. This model is valid for the five speakers (the exact values of the F's are speaker-dependent).

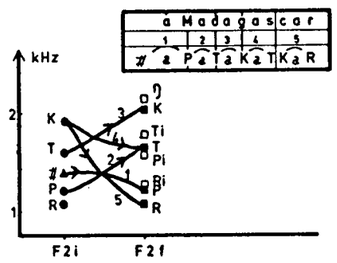

Fig. 5: MODELLING THE DIFFERENT INFLUENCES OF THE CONTEXT
F2 temporal course predicted for F2 during the vowel /a/, from F2 initial (F2i) to F2 final (F2f). For example, in the acoustic realisation of the sequence "à Madagascar" which is, on the articulatory equivalent to /aPaTaKaTkar/, F2 is sharply falling in kar (5), falling in gas (4), slightly falling in am (1), rising in mad (2) and dag (3), independantly of the speaker. When the following vowel is /i/ (as indicated by opened squares on the figure), F2f is higher in comparison with the cases where the following vowel is not /i/, as indicated by filled squares.

III. CONSEQUENCE FOR ASR

III.1 EXPLAINING PARTLY THE ACTUAL ERRORS

We have examined in an informal way the results obtained by the phonetic decoder of the KEAL system on a part of this sentences.and of a When the system is trained for the three V in their least reduced form (long V and PVPV context), good results are obtained in the test corpus for the V in similar contexts. Confusions often arise in different contexts, or in the cases the duration is shortened, in an expected way (See Fig.6). For example, the /a/ in a fast version of the word /quatre/ is generally identified as /ɛ/.

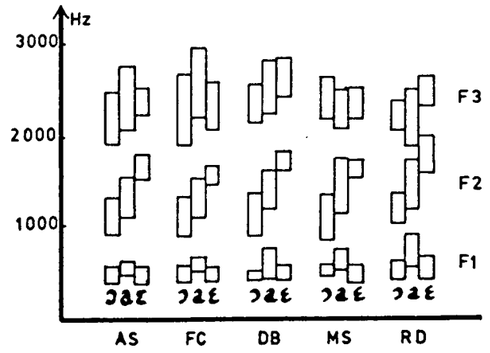

Fig. 6: RANGE OF F1, F2 AND F3 VARIATIONS
The vowels /a/, /ɔ/ and /ɛ/ are extracted from continuous speech. F1 tends to be higher for /a/, and F2 lower for /ɔ/ and higher for /ɛ/. There are, however, cases where the distinction between /a/ and /ɔ/ or /a/ and /ɛ/ is difficult to establish, at least from the formant values calculated on the spectrum sampled in the mid portion of the vowel.

In other terms, the results of the system depends directly on the particular allophones included in the training set, and of the duration of the vowels.

III.2 TRAINING THE SYSTEM

It is possible to create contexts maximizing or minimizing the coarticulatory effects [7,8]; The use of a very limited number of words for training allows to estimate, for each vowel and for each speaker, the range of possible variations for the F-values, at the center, but also at its onset and offest (See also [9]).

II.2 VERIFICATION RULES

The knowledge of physiological constraints is best to be put into the calculation of the fitness score between the output of an automatic decoding and the description of each word in the lexicon. Both the exact observed F-values and the description of the temporal course of the F during V (V shape, rising, falling, ...) are useful to verify the validity of an hypothesized CVCV phonetic sequences. For example, F2 in a sequence like /pak/ is obligatorily rising, independantly of the speaker (See Fig. 7): if the final consonant is not released, there is no burst and the F2 temporal course during the value (or the F2 value at the vowel offset) is the only way to distinguish between /pap/, /pat/ or /pak/. The acoustic distance between the vowel in the sequence /kak/, on one side, and the F-target values for the vowel /a/ and /ɛ/ should be expressed as a function of V duration: if V is less than 150 msec, it should be acoustically more closed to /ɛ/ than /a/.

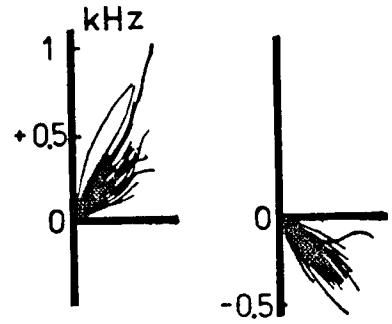

Fig. 7: OBSERVED F2 TEMPORAL COURSE DURING /a/
F2 temporal course during the vowel /a/ (five speakers plotted). On the right: PaK, PaT and TaK context, and on the left: TaP[-i], TaR[-i], and KaT[-i] context. The observed temporal course of F2 is in accordance with the model schematized in Fig. 5. The data have been aligned on the F2 value detected at the vowel onset.

CONCLUSION

First, for a general phonetic point of view, the study has confirmed for French and the usefulness of the notion of an invariant target for each vowel [5], the importance of the influence of the adjacent consonants on vowel formants [3], the existence of a vowel-to-vowel coarticulation phenomenum [6], and the importance of the time-factor for interpretating the formant patterns. Understanding of coarticulatory phenomena are essential in explaining the apparent variability in the acoustic realisations of a given vowel in continuous speech. It rises also questions as how the perceptual apparatus can deal with the distinction between phonemically distinct vowels which, through constraints due to the production apparatus, have become acoustically very closed in terms of F-pattern. The second conclusion concerns automatic speech recognition. Since there is exists a reasonably systematic relation between the F-pattern and the articulatory activity involved in the production, the systematic study of the F-pattern of the vowels in continuous speech provides a valuable tool for interpreting the limits of our actual phonetic decoders in terms of articulatory-acoustic relationship. Whether vowel identification is based on formant or not, the system has to deal in a proper way with the coarticulation phenomena in continuous speech.
Thirdly, it coins the usefulness of articulatory data concerning continuous speech. The possibility of obtaining articulatory data on large quantity of data and speakers [10] and the availability of a sound theory of speech production to interpret the link between articulatory data and the resulting acoustic signal [11] should lead to a proper theoritical framework for describing the range of possible acoustic variability for each sound. Such a framework is also needed to reduce the amount of training data necessary to adapt the systems to a new vocabulary or to new speakers.
Before such a concrete framework for describing coarticulatory phenomena is found, the knowledge can be only exploited in a rather ad-hoc manner, at the training level and/or at the level of phonetic acoustic decoder and lexical retrieval for a verification.

REFERENCES

[1] Mercier, G. and al, (1979), "The Keal speech understanding system", Proceedings of the NATO Advanced Study Institute, Simon J.C ed., 525-543.

[2] Leung, H. C., (1985), "A procedure for automatic alignement of phonetic transcriptions with continuous speech", Master of Sciences, Massachusetts Institute of Technology.

[3] House, A.S. and Stevens, K.N., (1963), "Perturbations of Vowel Articulations by Consonantal Context: An Acoustical Study", JSHR, 6, 111-128.

[4] Heffner, R-M, S., (1950), GENERAL PHONETICS, The University of Winconsin Press.

[5] Lindblom, B., (1963), "Spectographic Study of Vowel Reduction", JASA, Vol. 35, No. 11, 1773-1781.

[6] Ohman, S.E.G., (1966), "Coarticulation in VCV Utterances: Spectrographic measurements", JASA, Vol. 39, No. 1, 151-168.

[7] Abry, Ch. and Boe, L-J, (1984), "[i,a,u] Pas si fou? ou les lèvres des consonnes maximisent-elles l'espace vocalique des voyelles", 13èmes Journées d'Etudes du groupe de la communication parlée, Bruxelles, 28-30 Mai 1984, pp. 205-208.

[8] Liljencrants, J. and Lindblom,B., (1972), "Numerical simulation of vowel quality system: the role of perceptual contrast", Language 48, 839-862.

[9] Vaissière, J., (1985), "The use of allophonic variations of /a/ in automatic cotinuous speech recognition of French", 108Th Meeting of the Acoustical Society of America, Austin, Texas.

[10] Kiritani, S., Itoh, K. and Fujimura, O., (1975), "Tongue-pellet tracking by a computer-controlled x-ray microbeam system", JASA, Vol. 57, No. 6, 1516-1520.

[11] Fant, G., (1960), ACOUSTIC THEORY OF SPEECH PRODUCTION, The Hague: Mouton.