

ACOUSTIC-ARTICULATORY INFORMATION IN A SPEECH DATA BASE

D.Autesserre*, C.Barrera**, R.Espesser*, G.Perennou**, M.Rossi*, B.Teston*,
N.Vigouroux**.

ABSTRACT

This paper details the present state of a research programme aimed at setting up a multi-media acoustic-articulatory data base. The research is being carried out jointly as part of an 'acoustic-phonetic decoding' operation within the 'Speech Communication' section of the GRECO by two laboratories:

- the CERFIA, where homogeneous infra-phonemic segments are isolated by means of a pre-segmentation of the speech signal.
- The Institute of Phonetics at Aix-en-Provence, where the movements of the velum, the lateral walls of the pharynx and the lips during phonation are measured on pictures from video films.

We will be describing the various processes used to establish a relation between the two sets of data and presenting the initial results of this confrontation.

INTRODUCTION

Minute segmentations, of time and frequency, of the acoustic signals on the 'BDson' provide data base users with a set of infra-phonemic segments. These are subject to the two main sources of known variations - interindividual variations, which are dependent on idiosyncratic factors and intraindividual variations, as the combination of phonic units in the spoken sequence entails a contextualisation of the acoustic cues. In order to achieve a better definition of such variability, it seems necessary to take into account speech production processes. It would appear that future users of acoustic data bases should be given additional physiological information. How are we, however, to correlate the discontinuities of the speech signal with articulatory events of varying duration? In fact, two types of segmentation are involved and although both will eventually establish the existence of invariants necessary to speech communication, each will do so according to the particular limitations of its own operative logic. Such a correlation can be achieved through the comparison, at given intervals, of the evolution (represented, for instance, by curves) of the acoustic and articulatory cues, any discontinuity being located by appropriate labelling.

* Institut de Phonétique d'Aix, UA 261 CNRS, Université de Provence I, 29 avenue R. Schuman, 13621 Aix Cedex, France.

** Laboratoire CERFIA, UA 824 CNRS, Université P. Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France.

EXPERIMENTS AND METHODS

1. Isolation, measuring and labelling of articulatory data.

The physiological cues selected are related to the movements of the velo-pharynx and the lips as filmed by two video cameras. In order to collect pictures of the velum one of the cameras is connected to a flexible fiberscope placed in the nose (ref 1). A corpus of a hundred or so sentences containing the bi-labial nasal consonant [m] in various vocalic and consonantal contexts is read by two speakers (male and female). A tracing is made from each frame of the video films (one frame every 20 ms). The movements of the central part of the velum and the lateral walls of the velo-pharynx are calculated by reference to their position when the subject is breathing in quietly. For the movement of the lips only the inter-labial aperture is taken into consideration. These values are transposed onto graphs with time as abscissa (see fig. 1, three top curves). The extreme positions of the articulators are labelled as follows: F (closed) and O (open) for the lips (with DF further indicating beginning of labial occlusion). As to the velum, H indicates the highest position and B the lowest. Finally, the movements of the lateral walls are labelled E (widened) and R (narrowed). A minute tempspeech segmentation of the speech signal corresponding to the pictures is achieved by means of the 'Signaix' speech signal processing system (ref 2 & 3).

2. Description of the acoustic-phonetic data.

The signal undergoes numeration again at the CERFIA laboratories and is analysed on a filter bed. At the end of this process, J. Caelen's cues (ref 4) - mellow/strident, open/close, stretched/compact, grave/acute, - are calculated. The energy and spectrum curves are also available. A function of pre-segmentation, established from the cues, makes it possible to split the spectrum up into homogeneous infra-phonemic units. These units are then labelled according to their macro-class, phoneme, acoustic phase and modality (ref 5).

3. Synchronisation.

The correlation of the physiological data transposed at the end of each 20 ms frame with the 16 ms signal blocks is achieved by means of 4 ms samples (lowest common divisor). The upward trend of a click serves as starting point on the time axis and enables the synchronisation of a video frame with a spectral sample. In order to avoid any variation between acoustic and physiological data, we had to take into consideration technical problems resulting from the transfer of tempspeech information from analogical analyses in a numerical system.

RESULTS

They are an attempt at summing up the present state of our knowledge following observations made on phonic sequences containing the nasal bi-labial consonant [m] in a symmetrical [a] vocalic context, as in fig. 1: [isamassil].

1. Evolution of physiological parameters.

The movements of the velum and the velo-pharyngeal lateral walls are similar although the lateral walls have a lesser amplitude and a greater inertia. The velum falls rapidly on the first vowel and rises more slowly on the second and longest. The inter-labial aperture is greater in the case of the second vowel than in the first ('opening' effect of stress).

2. Comparison with acoustic segmentation cues.

It should be made with caution - we have, on the one hand, a complete set of acoustic cues and, on the other hand, a limited number of physiological cues (we have no data on tongue movements). The lip movement curve (inter-labial aperture) is correlated to the energy and compactness curves. The curves of two cues, namely 'compact' and 'open', follow the fall of the velum. The 'grave' cue marks the limit of the nasal bi-labial consonant [m]. In the final part of the second [a], the resurgence of the 'grave' cue may signal the end of the nasalisation of the vowel together with sufficient rising of the velum (which persists nonetheless with lesser amplitude on the subsequent consonant).

3. Comparison of segmentations.

The movements of the lips provide the best segmentation cues. The different trends of the curves during rise and fall of velum and their points of inflexion in relation to concomitant changes in the acoustic cues are currently the subject of statistical treatment.

CONCLUSION

The initial results of this confrontation between part of the segmentations from articulatory and acoustic cues appear to be promising. They prompt us to pursue further research incorporating more physiological data (ref 1). An acoustic-articulatory data base thus established should be of great help in the understanding of the fundamental phenomena of acoustic-phonetical decoding. Applied to language technology, especially as far as multi-speaker recognition is concerned, it should be extremely productive.

BIBLIOGRAPHY

1. B.Teston & D. Autesserre, "Description d'un dispositif d'enregistrement simultané des organes articulatoires", XV^e J.E.P., Aix, 1986, 65-68.
2. R.Esperer, "Signaux: a speech signal processing system software using unix", T.I.P.A. 10, 1985-1986, 335-357.
3. D.Autesserre & M.Rossi, "Propositions pour une segmentation et un étiquetage hiérarchisés; application à la base de données acoustiques du GRECO Communication Parlée", XIV^e J.E.P., Paris, 1981, 147-151.
4. J.Caelen & G.Caelen, "Indices et propriétés dans le projet ARIAL II. Processus d'encodage et de décodage phonétiques", Toulouse, 1981, 128-143.
5. C.Barrera, J.Caelen, G.Caelen-Haumont, J.F.Malet, N.Vigouroux, "Towards an automatic labelling system", XIth Int. Cong. of Phon. Sciences, Tallinn, 1987.

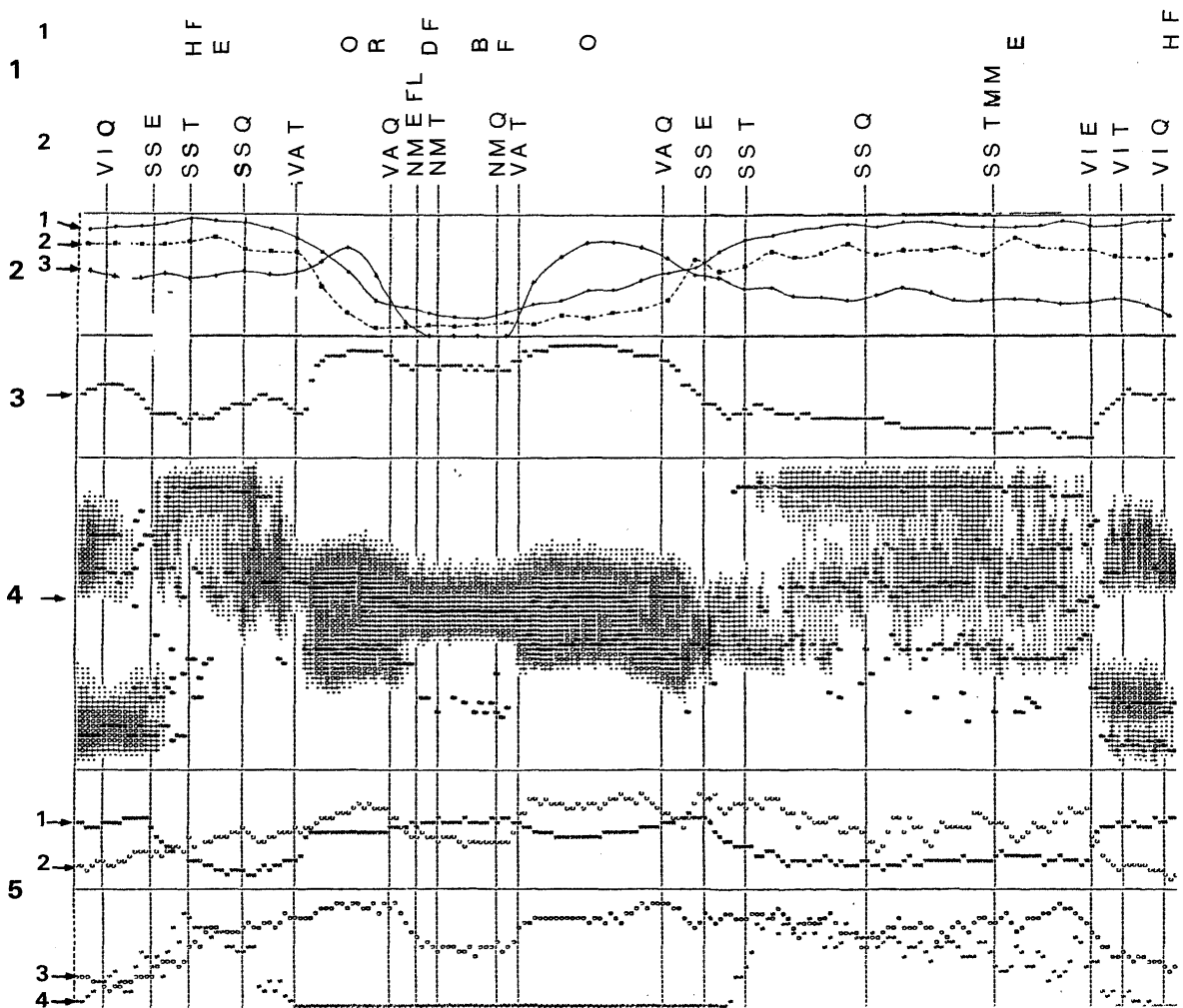


Fig. 1 Correlation of physiological and acoustic cues for the phonic sequence [isamassil]

1. labels
 - 1.1. physiological
 - 1.2. acoustic
2. physiological cues
 - 2.1. movements of the velum
 - 2.2. movements of the velo-pharyngeal lateral walls
 - 2.3. inter-labial aperture
3. energy (dB)
4. spectrum (Hz)
5. selected acoustic cues
 - 5.1. grave (g)
 - 5.2. compact (c)
 - 5.3. open (o)
 - 5.4. strident (s)