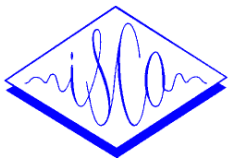

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

BOTTOM-UP ACOUSTIC-PHONETIC DECODING FOR THE SELECTION OF WORD COHORTS FROM A LARGE VOCABULARY

H. MÉLONI, F. BÉCHET, P. GILLES

Laboratoire d'Informatique, Faculté des sciences,
33, rue Louis Pasteur, 84000 AVIGNON FRANCE

ABSTRACT

Within the framework of the analytical recognition of the words of a large vocabulary, we propose a system permitting the selection of limited cohorts of lexical items from a lattice of valued phonetic units.
Before the system can work, it has to learn all the phonetic units of a given speaker by storing the spectra of each phoneme. During the construction of the phonetic lattice, the units are simultaneously localized and identified by means of various types of spectral distances adjusted according to the phonemes, the context and certain characteristics of the speaker's voice. A few factors dependent on the phonemes as well as various types of spectral stability make it possible to determine the kind of sound and the areas in which the scores will be calculated.

In the first stage the system interprets the information available in the lattice and then it accesses the lexis through a series of filters thus reducing the number of possible words. The solutions are given in the form of a cohort of words that are assigned a value according to the scores and the rate of temporal covering-up of the phonemes identified.

1. INTRODUCTION

The recognition system functions in three stages : bottom-up Acoustic-Phonetic Decoding [6], word cohort selecting, and top-down sharp discrimination of the cohort item [7]. Here we are presenting the main characteristics of the first two stages of the system whose specifications are as follows :
- a large incremental vocabulary (more than 20,000 words),
- limited machine learning (a few reference spectra for each speaker),
- analytical recognition of isolated and equally probable words,
- results given in the form of ordered cohorts.

The simplicity of the method used makes it easily adaptable to the lexis and to the speaker ; this flexibility is the main feature of our system compared with others using far more sophisticated techniques [9], [3], [2], [11].
The data processed at this stage do not take into account the effects of the coarticulation of the phonemes, which are examined in the top-down speaker-independent analytical stage [7].
The general structure of this part of the system is shown in figure 1.

2. GENERAL STRUCTURE OF THE APD SYSTEM

The speech signal is digitized on 16 bits at a frequency of 12.8 kHz, then preaccentuated and characterized every 10 ms by its global energy, the fundamental frequency, zero crossing density and the spectral energies in 24 channels distributed according to a Mel scale.

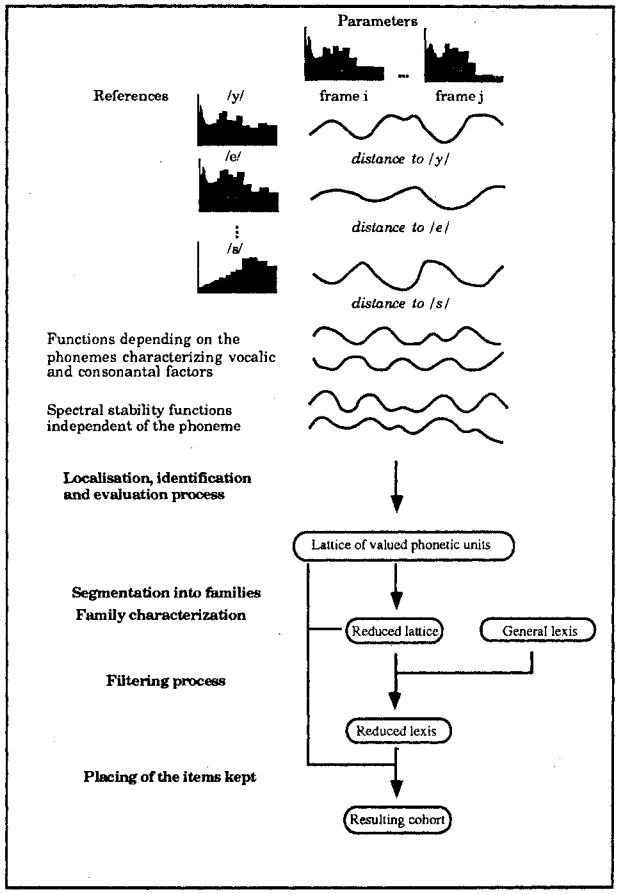

Figure 1 : general structure of the system

2.1 Spectral references

The bottom-up APD system uses a set of spectral references corresponding to the stable phases of the phonemes (vowel and constrictive duration, plosive occlusion) and to certain interesting transient areas (burst of voiceless occlusives). The aim here is not to give a precise description of the sounds -

which in any case are more or less modified by the context - but to offer an average pattern close to the ideal utterance of a phoneme by a speaker.
The acquisition of references is carried out for each speaker from a very limited set of phrases in which the phonemes appear in contexts allowing few distortions. Therefore this adaptation stage is quick and imposes few constraints.

2.2 General strategy

Different types of distances to the references are systematically calculated and constitute the essential parameters to localise and identify the phonemes. When a part of the signal includes a distance that is inferior to a relatively wide threshold, this interval is considered a likely candidate.
A very limited set of simple rules uses the hypotheses deducted from the distances, as well as some additional parameters (spectral instability, vocal or consonnant type, etc.) to propose valued phonetic units by means of a distance that is independent of the phoneme and taken at a particularly stable point of the signal.

2.3 Distances

A number of techniques have been used in many contexts [6], [4], [5], [8] in order to calculate the distances to reference vectors representing phonetic units. Numerous factors determine the choice of an optimal method, such as signal quality, the type of parametric representation (MFCC coefficients, spectra, etc.), the number of vectors used in the calculation, the taking into acount of information concerning the phonetic unit to be identified and/or its context, etc.
In spite of the spectral references being adapted to the speaker, the acoustic variation of phonemes - resulting mainly from coarticulation phenomenons - remains enormous and does not allow the very accurate identification of phonetic units. However the distances must highlight the most relevant information in various contexts. In view of the chosen system of parametric representation, our task consisted in solving the following problems :
- quick and easy adaptation to the speaker,
- adjustment of spectral energy levels,
- optimum consideration of variations in position and amplitude of the spectral maxima (formants),
- integration of contextual information available in the signal.

2.4 Building up the lattice

Figure 1 illustrates the general strategy of the system to build up the lattice of phonemes. The previously defined shapes found on the curves that measure the distances to the phonemes (valleys, areas where the values are inferior to a threshold) make up the starting elements which will lead to the selection of units.
In most bottom-up APD systems a stage of segmentation into pseudo-phonetic macro-categories is implemented before any identification is carried out. The limitations of this technique have led us to associate the localisation and the identification processes ; the segmentation comes in exclusively to validate and define the choices obtained through the distance system.

The phonetic lattice is made up by storing the units valued according to the following method :
- locating the intervals whose distance to the phoneme is inferior to a threshold chosen according to the phonetic unit and to the results expected (the areas being calculated by means of pattern recognition tools),
- selecting the intervals corresponding to the phoneme dealt with from the viewpoint of vocalic and consonantal factors (the initial area is limited to its intersection with the area defined by the factor),
- calculating the score assigned to the phoneme by evaluating the total distance on the most stable frame in the interval associated with the phoneme.
The lattices thus obtained generally contain all the phonemes uttered in the phrases although a fairly great distance may exist with the references. The only units that can be missing correspond to situations in which the consonantal quality is difficult to highlight (it may be the case, in particular, of sequences of consonants in which only the most closed one is likely to be localised ; e.g. /t/ in /trw/).
Given the simplicity of the system (reduced machine learning, no context to be taken into account) the scores obtained for the phonemes actually uttered are generally satisfactory ; the unit expected frequently appears at the top of the list especially if the context has caused few distorsions or if the syllable is stressed.

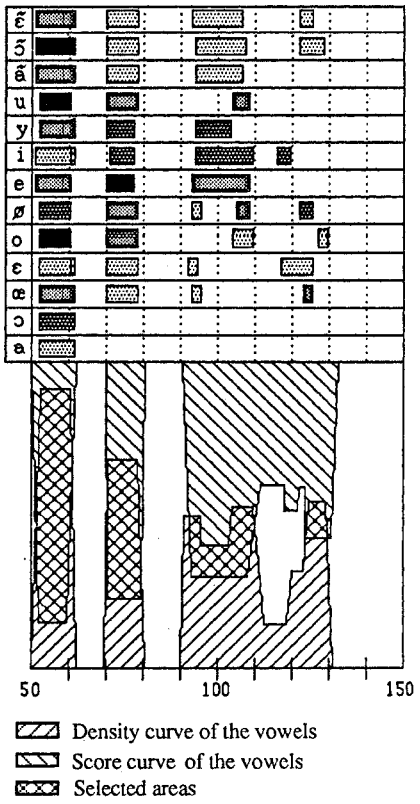

Figure 2 : intersection of the density and score curves of the vowels in the word "aubépine".

3. SELECTING THE COHORTS OF WORDS

3.1 Segmentation into families of phonemes

This segmentation consists in grouping the phonemes of the lattice in sets that we shall call *families* ; these families constitute the most likely stages along the route of admissible solutions. The principle is that each family should match one phoneme uttered by the speaker.
So as to obtain these families the system will look for the

different areas where a certain type of phoneme is most likely to be found (vowel, nasal consonant, voiceless occlusive, fricative, etc.). For each type, calculations are made of :
- the density curve of the phoneme in the lattice,
- the score curve, keeping for each frame of the signal the best score obtained by the phonemes including it.
Those areas represent the intersection between the hills of the density curve and the valleys of the score curve. Figure 2 shows the lattice of the vowels in the word "aubépine" and the overlapping of the two curves. The vowel areas obtained in this example are : <51,62>, <70,79>, <93,111>, <116,128>.

Once calculated, the areas are grouped so as to keep only two distinct types : vowel and consonant. The families are made up of all the phonemes of the same type whose most stable frame belongs to the previously found area.

3.2 Characterizing the families of phonemes

This stage will permit representing each family by a number of categories of phonemes. The consonants are grouped in classes according to their articulation (for example the class of voiceless fricatives { /f/, /s/, /ʃ/}). On the other hand each vowel represents a distinct category containing the phonemes that are close to this vowel in various contexts. This classification permits taking into account the distortions due to coarticulation effects.
In one family each class is evaluated from the best scores of the phonemes constituting it. A family is made up of the five best classes matching the corresponding area. Figure 3 shows the vowel classes kept in the lattice of the word "aubépine".

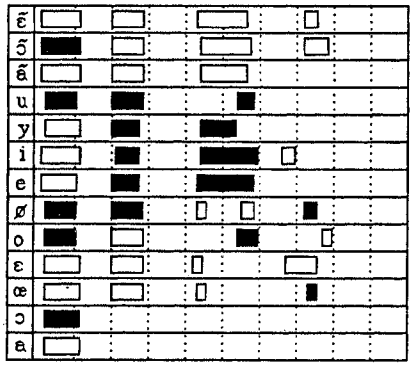

Figure 3 : in black the classes kept in the vowel lattice for the word "aubépine".

After the search for information is completed, we have a reduced lattice, segmented into families in which each family comprises five vowels or five groups of consonants. The number of families corresponding to vowels is generally equal to the number of vowels in a word. However in certain situations (stressed syllables, nasal vowels, optional utterance of /ə/) additional vocalic families may appear.

3.3 Exploration of the lexis

The lexis used for the tests is that of BDLEX [10] consisting of about 22,000 words. This lexis has been broken down into about a hundred word files grouping the items according to some characteristics (number of vowels, first vowel in words, etc.).
The system will select the cohort of words through a series of filters progressively reducing the number of possible choices.

3.3.1 Reduction of the number of candidate words

The first filter consists in choosing among the possible word files those whose schematic description matches the characteristic elements extracted from the lattice.
The second filter uses the reduced lattice previously calculated. The phonetic breakdown of each word selected is limited to the vowels and the intervocalic consonants of the word. In a group of consonants, only the most closed one is kept ; each is represented by the class it belongs to. The words kept after this filtering stage are put in a tree where they are encoded according to their phonetic representation just obtained. This second filtering operation consists in going simultaneously over both the tree and the reduced lattice while checking at every stage the compatibility between the phonemes of the phonetic representations and those of the families.

3.3.2 Selecting the cohort of words

The general lexis having been reduced to a reasonable number of items by the different filters, the system now calculates a plausibility score for the words selected. The evaluation is carried out as follows :
a) The system chooses the best possible correspondence (according to the criterion defined further down) between the vowels and the intervocalic consonants of the lattice and the words selected. If one of the necessarily present consonants of a word is not in the lattice, the word is rejected.
b) Then it looks for the phonetic units that are not necessarily present in the lattice. If the phoneme has not been identified in the lattice, it is looked for in a top-down way in the temporal interval where it is expected by a simple calculation of the distance between the signal and the reference.
c) Each word is evaluated according to the scores of the phonemes identified and to the rate of temporal covering up of the speech signal. The resulting cohort is made up of the words with the best scores.

4. RESULTS

The tests carried out aimed at validating and measuring the capability of the system to select limited cohorts of words from a large vocabulary by using the units obtained by the bottom up APD system. We have chosen at random 100 utterances (properly described by the phonetic lattice) out of the 22,000 words of the lexis .
We are giving below the percentage of words actually identified among the first n candidates. The phonemic composition of the words (number of phonemes, distribution of the phonemes in the word, coarticulation, etc.) strongly influences the placing of the words in the cohorts.

Placing	Percentage
1st position	31%
first 5	53 %
first 10	62 %
first 20	72 %
first 50	87 %
first 100	94 %
first 150	99 %

Table 1 : percentage of the words identified among the first n candidates.

5. CONCLUSION

The first results of the word cohort selection system from a bottom-up acoustic-phonetic decoding process are encouraging and can prove quite usable for words of several syllables. However the bottom up ADP technique used does not make it possible to take coarticulation effects into acount. Therefore the recognition system will have to be complemented by a top-down checking stage so as to discriminate the various elements of a cohort with precision and whatever the speaker. This method has already produced noteworthy results for the top-down identification of voiceless occlusives [7].

BIBLIOGRAPHY

[1] APPLEBAUM T.H., HANSON A.H., WAKITA H., (1987), *Weighted distance measures in vector quantization based speech recognizers*"" ; Proceedings ICASSP, pp. 1155-1158.

[2] BILLI R., ARMAN G., CERICOLA D., MASSIA G., MOLLO MJ., TAFINI F., VARESE G., VITORELI V. *A PC-based very large vocabulary isolated word speech recognition system* ; European Conference on Speech Recognition, Paris novembre 89.

[3] FERRETTI M., SCARCI S. *Large vocabulary speech recognition with speaker-adapted codebook and HMM parameters* ; European Conference on Speech Recognition, Paris novembre 89.

[4] GRAY A.H. Jr., MARKEL J.D., (1976), *Distance Measures for Speech Processing* ; IEEE Trans. Acoust. Speech and Signal Proc., Vol. ASSP 24, n°5.

[5] ITAKURA F., UMESAKI T., (1987), *Distance measure for speech recognition based on the smoothed group delay spectrum* ; Actes du 7ème Proc. ICASSP 87 (Dallas, TX).

[6] MÉLONI H., GILLES P. *Décodage acoustico-phonétique ascendant* ; Revue Traitement du Signal, N° 2, 1991.

[7] MÉLONI H., GILLES P. *Représentation de connaissances indépendantes du locuteur pour la reconnaissance de mots acoustiquement proches* ; XIIème Congrès International des Sciences Phonétiques, 19-24 Août 1991, Aix-en-Provence.

[8] NOCERINO N., SOONG F.K., RABINER L.R., KLATT D.H., (1985), *Comparative study of several distortion measures for speech recognition* ; Proceedings ICASSP, pp. 25-28.

[9] PARDO JM., HASAN H. *Large vocabulary speaker-independent isolated-word speech recognition using hidden Markov models : status report and planned research* ; European Conference on Speech Recognition, Paris novembre 89.

[10] PERENNOU G. *Le projet BDLEX de base de données et de connaissances lexicales et phonologiques* ; Journées nationales du GRECO PRC-CHM, tome 3, pp 1205-1214, Paris, 24-25 novembre 1988.

[11] SCIARRA D., SCAGLIOLA C. *Two-step recognition of large vocabulary isolated words based on diphone spotting* ; European Conference on Speech Recognition,Paris novembre 89.