PARALLEL DISTRIBUTED PROCESSES FOR SPEAKER-INDEPENDENT ACOUSTIC-PHONETIC DECODING

A. Ghio and M. Rossi
Institut de phonétique d'Aix-en-Provence
Laboratoire "Parole et Langage" URA 261, CNRS
29, Av.R.Schuman, 13621 Aix-en-Provence, FRANCE

ABSTRACT

We aim at examining to what extent a knowledge-based model can recognise segmental structure without feedback from semantic information and without stochastic modelling. The system proposed is inspired by some features of human cognitive processing: the speech signal activates parallel distributed processes of decoding. A supervisor module takes the final decision after the access to a dictionary and a top-down verification.

GENERAL PRESENTATION OF THE SYSTEM

The field of this study is automatic speech recognition and concerns more precisely speaker-independent acoustic-phonetic decoding. We aim at examining to what extent a knowledge-based model can recognise segmental structure without feedback from semantic information and without stochastic modelling.

The system proposed is inspired, in a functional way, by some features of human cognitive processing [1, 2]. The sequence of operations can be characterised as data driven. The speech signal first arrives at the low level analysis demons which then activate parallel distributed processes of decoding (Fig.1). The modules of this multi-analysis and multi-expert system are conceptually different. They consequently do not give the same output. Their results, then, are sent to the cognitive demons, who act upon them using high level information (phonological rules, access to a dictionary...). Finally, after a top-down verification, a decision process selects the alternative that has the strongest evidence.

First of all, we present the different modules of the bottom-up decoding i.e. the automatic segmentation, the global and the analytic recognition. Secondly, we develop the main ideas used in the high levels processes, especially in the access to a dictionary and the supervisor. Partial results are presented in a third part, just before a conclusion.

THE DIFFERENT MODULES OF THE BOTTOM-UP DECODING

The bottom-up decoding is composed by different parallel distributed processes.

The automatic segmentation

According to the general outlines of the Level Building procedure [3], an automatic segmentation module SAPHO (Segmentation by Acoustic-PHOnetic knowledge) supplies a hierarchical set of acoustic properties and segments, and phonetic properties and segments which fit the phonetic parsing of the acoustic wave [4]. It is not an unguided method, which are generally based on an instability function. In the SAPHO algorithm, energy + zero cross ratio parameter + some spectral features permit the location and the rough identification of segments. Only the nature of the segment authorises a posteriori the precise location of the boundaries.

The output is the labelling of temporal frames in macro-classes (vowel, stop, fricative, vocalic consonants, silence...) which permit an oriented analysis.

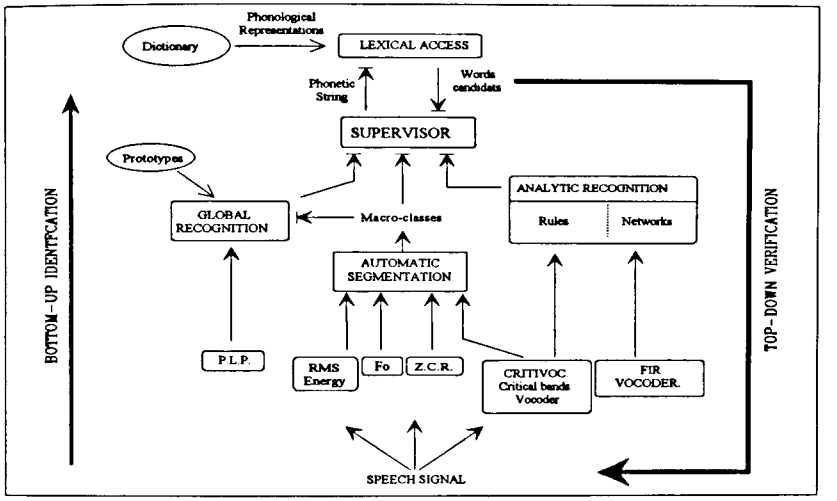

Figure 1 : Functional diagram of SYMULDEPHO

The global recognition

The global recognition module [5] is based on metric methods. Decoding units are CV groups. First, the feature extraction is done by a Perceptually based Linear Prediction analysis [6, 7]. Then, the system uses a Data Time Warping algorithm in order to compare stimuli to references. The output is a list of classified CV candidates. On account of the variability of speech, the system does not keep only the best. The analysis of cues relative to the ten best candidates allows the construction of the best prototype using a vote procedure (ex: if among the 10 best consonant candidates, 9 are voiced, 8 acute, 1 compact, 0 continuant, 1 nasal, 1 vocalic, the module propose [d] as solution)

The analytic recognition

The first analytic recognition module uses networks which are oriented graphs with state transitions. The differences with the Markov chains are the lack of probability and the fact that they work on the paradigmatic axis and not the syntagmatic one. They are supposed to model the allophones of vowels and not the abstract units. Each network is specialised for the recognition of one vowel. All are activated at each temporal frame. If a path is found along a network, an output appears at the end. The result is a table containing the allophone candidates. The analysis of the temporal distribution of phonemes leads to strong hypotheses on the vowel identification and its context.

The second analytic module is based on phonetic rules. Acoustic cues, different from the previous ones, are extracted every temporal frame (centi-second) by modelling some psycho-acoustic phenomena (weighted sound level, critical bands). The analysis of phonetic features allows the system to identify the phonemes using rules. A temporal tracking, which takes into account the contextual variability of segments, provides information by analysing acoustic transitions of coarticulation in triphones. For example, the sequence [...eeœœaaaaaaa...] is typically the result of the decoding of /a/ in the left context of /k, g/.

THE HIGH-LEVEL MODULES

Each parallel distributed process of decoding provides a set of phonetic units which are sent to the high-level modules.

The supervisor

The supervisor-process is not finished and we would like to improve the decision methods. Time being, the different results of the bottom-up decoding are received by the supervisor which then makes up a list of possible phonetic strings.

In the case of isolated words, the access to a dictionary allows the supervisor to classify the phonetic strings by associating them to words-candidates (cf. next section). The top-down verification, which will be described in a next paper, authorises the selection of the alternative that has the strongest evidence.

The access to a dictionary

Some methods of automatic spelling-correction compute a distance between a reference-word and a test-word; it relies on a series of operations that model errors of insertion, deletion and substitution [8]. It is possible to realise these operations using dynamic programming [9]. The module of lexical access is inspired by this method.

In our case, distance is not computed between graphemes but between the decoded phonetic string proposed by the supervisor and the phonetic representations of words stored in a dictionary (Fig.2). The dynamic programming is efficient to integrate in a single algorithm all the phenomena of insertion, deletion and substitution which appears in the bottom-up decoding.

The comparison requires the computation of a local distance between the sub-units of the strings (Fig.2). Whereas in the case of orthographic strings, the local distance is basic (0 if graphemes are the same, 1 if they are different), the case of phonetic strings requires a more sophisticated measure. Actually, the difference between /i/ and /e/ is less important than the confusion of /i/ with /p/. This is the reason why we have introduced a matrix of cost-confusion, which indicates the difference between each phoneme. It also authorises the non-precise definition of a phoneme in the stimulus string. For example, on Figure 2, the 5th unit of the stimulus has been decoded as 'liquid' which is the macro-class of /l/ and /r/.

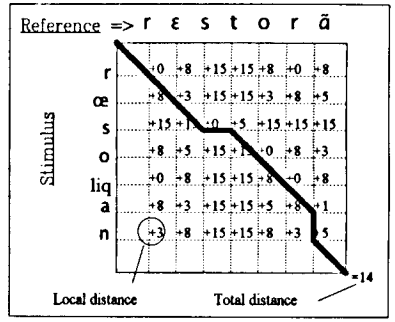

Figure 2 : Calculation of distance between 2 phonetic strings.

It is interesting to mention that such a method could be useful to evaluate, in a precise lexicon, the degree of difficulty of decoding: a great distance between words could indicate an easy task of decoding.

We also mention that phonological rules are necessary to forecast all the variations of a word's pronunciation. For example, the word "petite" (French word for "small"), whose phonological transcription is /pətitə/, can be pronounced as [ptit], [pətit], [ptitə] or [pətitə] and should have 4 phonetic entries in the dictionary.

RESULTS

In a first step, each module has been tested independently using 10 French speakers (5 male, 5 female) recorded in the corpus SYL of the French database BD-SONS. Stimuli were CVCV non-words as "titi", "rara", "sussu"... that represents the combination of 2 * 10 vowels * 16 consonants * 10 speakers = 3200 tests.

Results of the global recognition

For each tested stimulus, the module can provide different candidates. We distinguish (Table 1): the case where the good phoneme is in the list of candidates (column *'Correct'*) and the case where it is not (column *'Error'*). The column *'Num cand'* furnishes the average number of candidates proposed by the module, which can be compared with the number of potential candidates (column *'Num max'*). For the vowels, the identification process classifies the candidates: the column *'1st rank'* (Table 1b) is the case where the good vowel is in the first position, the column *'other rank'* is the case where the good vowel is in the second, the third, ... position.

Table1 : Results of the global recognition

(a)	Correct	Error	Num cand	Num max
Consonants	86 %	14 %	2.89	17

	Correct				
(b)	1st rank	other rank	Error	Num cand	Num max
Vowels	65 %	25 %	10 %	2.18	10
	90 %				

Results of the analytic recognition

Here, we are only presenting the tests relative to the identification of vowels.

In the module of analytic recognition by rules, the system proposes 2.35 candidates on average, among the ten vowels [a,i,u,o,e,y,ø,ɑ̃,ɛ̃,ɔ̃]. In 92.4% of the cases, the good vowel is among the candidates.

The module of analytic recognition by networks has been tested using a corpus of 42 words pronounced by 5 speakers. On about 450 vowels, the result was correct in 92% of the cases.

CONCLUSION

We have presented the different parts of a system based on parallel distributed processes for speaker independent acoustic-phonetic decoding. Each module seems efficient in the recognition task.

Our aim now is to integrate all these knowledge-sources to collaborate in a single system. We are testing it with a large corpus of 500 French words pronounced by seven speakers. The results will be published in a next paper.

REFERENCES

[1] Lindsay P., Norman D. (1977), *Human information processing - An introduction to psychology,* Academic Press, New York, USA.

[2] Edelman G.M (1992), *Bright air, Brilliant Fires : on the matter of mind,* Basic books, New York, USA.

[3] Meyers C.S., Rabiner L.R. (1981), "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition", *IEEE ASSP*, vol.29, pp. 284-297.

[4] Rossi M. (1990), "Automatic speech segmentation:why and what segments ?", *Revue Traitement du Signal*, GRETSI, vol.7, n°4, pp. 315-326

[5] Ghio A., Rossi M. (1994), "Reconnaissance globale et analytique dans SYMULDEPHO, un système multi-locuteurs de décodage acoustico-phonétique", *Proceedings of workshop on 'Automatic Speech Recognition'* Nancy,France, GDR-PRC Man-Machine Communication

[6] Yong G., Mason J.S. (1987), "A comparison beetween vocal tract and auditory feature analysis in ASR.", *Proceedings of Eurospeech*, Edinburgh, pp.132-135

[7] Hermansky H. (1990), "Perceptual linear predictive (PLP) analysis of speech.", *J.Acous.Soc.Am.*, vol.87, *pp.* 1738-1752.

[8] Wagner R.A, Fisher M.J. (1974), "The string-to-string correction problem", *Journal of the ACM*, 21,1.

[9] Véronis J.(1994), "Distance entre chaînes: extension aux erreurs phonographiques", *Travaux de l'Institut de Phonétique d'Aix*, vol.15,pp.219-233.