A knowledge-based model for speaker independent acoustic-phonetic decoding

Alain Ghio, Mario Rossi

A KNOWLEDGE-BASED MODEL FOR SPEAKER-INDEPENDENT, ACOUSTIC-PHONETIC DECODING

A. Ghio & M. Rossi
e-mail:phonetic@univ-aix.fr
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 examine 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 in that the speech signal activates parallel distributed processes of decoding. The modules, conceptually different, are :
- an automatic segmentation module.
- a first analytic recognition based on oriented graphs with state transitions.
- a second analytic recognition module based on phonetic rules.
- a global recognition based on metric methods.
Finally, scrutiny of all the parallel results and access to a dictionary allow the inference rules to propose ranked word candidates.

1. GENERAL PRESENTATION OF THE SYSTEM

The object of this study is automatic speech recognition and concerns more precisely speaker-independent acoustic-phonetic decoding. We examine to what extent a knowledge-based model can recognise segmental structure without feedback from semantic information and without stochastic modelling [1, 2]. The system proposed is inspired, in a functional way, by some features of human cognitive processing [3]. The sequence of operations can be characterised as data driven. The speech signal first arrives at the low level analysis demons [4] 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 e.g. phonological rules, access to a dictionary, etc. Finally, after a top-down verification, a decision process selects the alternative that has the strongest evidence in its favour.

First of all, we present the three different modules of the bottom-up decoding: automatic segmentation, analytic recognition and global recognition. Secondly, we develop the main ideas used in the high-level processes, especially in the access to a dictionary and the supervisor. Preliminary results are presented in a third part.

2. THE DIFFERENT MODULES OF THE BOTTOM-UP DECODING

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

2.1. Automatic segmentation

The automatic segmentation module (Fig.1) called SAPHO (Segmentation by Acoustic-Phonetic Knowledge) has already been presented in its first version [5]. It is not an unguided method in which boundaries are generally placed a priori on the regions of spectral instability. In the SAPHO algorithm, the idea is to identify the phonetic forms beginning progressively with the most evident ones and finishing with the most subtle. The emergence of segments is not immediate and requires several steps. The new version of SAPHO is based on the extraction of robust acoustic parameters, the identification of cues and the application of a set of rules.

Overall energy, number of zero crossing and some spectral features are extracted for each temporal frame (10 ms). Four basic cues are deduced from these parameters : silence/signal, voiced/unvoiced, strong/weak energy, strong/weak friction. These cues permit a first labelling of frames in three rough classes: [silence], [transition+consonantism] and [vocalism]. Then, a temporal tracking permit the construction of phonetic forms by grouping different sets of frames. In a third step, the algorithm analyses each phonetic form more precisely and tries to specify its nature. If it is clear, segments are categorised by one of the six macro-classes which are: vowels, vocalic consonants, voiced or unvoiced fricatives, voiced or unvoiced stops. If the algorithm does not have enough information at its disposal, it makes no decision: the segment label stays fuzzy. The information about the nature of the segment and its context permits a posteriori the precise location of boundaries, which are placed progressively between the different units. Continuums of vocalic phonemes, e.g. in the word "analyse", are segmented if the algorithm can find evident discontinuities.

Finally, the automatic segmentation provides very important information on the temporal distribution of phonetic units (syntagmatic axis) and their possible nature (paradigmatic axis).

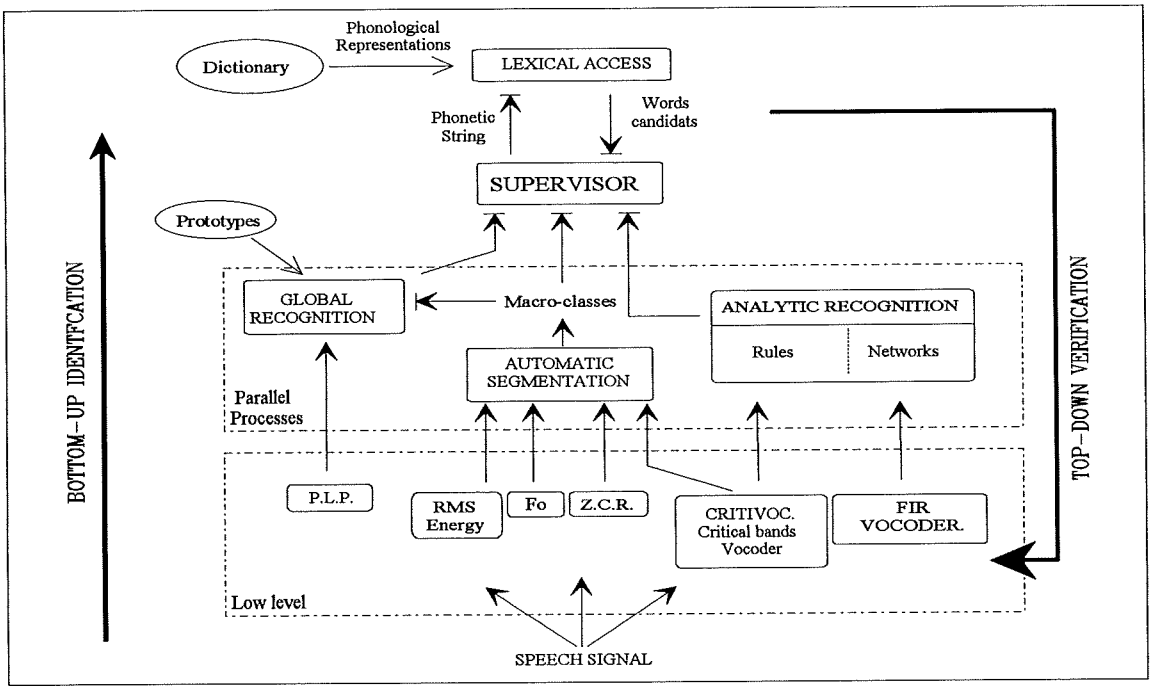

Figure 1 : Functional diagram of the system

2.2. Analytic recognition

The first analytic recognition module (Fig.1) uses networks which are oriented graphs with state transitions. Unlike Markov chains, this model does not use probability and works on the paradigmatic axis rather than the syntagmatic one. It models 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 (Fig.1) is based on phonetic rules. Acoustic parameters, different from the previous ones, are extracted for every temporal frame (10 ms) by modelling some psycho-acoustic phenomena e.g. weighted sound level, critical bands, etc. For the decoding of vowels, the distribution of energies calculated in the critical bands permit the deduction of two phonetic features : open/close and grave/acute. This analysis allows the system to classify the vowels in four classes : open/acute (/œ/, /ɛ/, /a/[1], /ɛ̃/), open/grave (/œ/, /ɔ/, /a/[2], /ɑ̃/, /ɔ̃/), closed/acute (/y/, /i/, /e/, /ø/) and closed/grave (/u/, /o/). In a second step, the algorithm locates the relevant peaks in the spectral distribution of the critical bands. Finally, for each frame, one or two vowel candidates are proposed by the module taking into account the phonetic class and the place of the peaks. Results are filtered to eliminate the isolated appearance of a candidate. The decoding of consonants has been abandoned in this module because such a task needs precise contextual information which is not available at this step of the decoding.

2.3. Global recognition

The global recognition module (Fig.1) is based on metric methods. Decoding units are CV groups. We have selected these units because their number is limited compared to the great number of words in a large vocabulary. For French, we have used 10 vowels /a,i,u, o+ɔ,e+ɛ,y,œ+ø,ɔ̃,ɑ̃,ɛ̃+œ̃/ and 16 consonants /ptkbdgfsʃv zʒmnlʀ/ which give 160 different CV combinations. Secondly, most coarticulation phenomena take place within such CV groups.

The first step of the global recognition module is feature extraction done by a Perceptually based Linear Prediction analysis [6]. Then, a Data Time Warping algorithm is used in order to compare stimuli to references. To account for variability, 10 prototypes extracted from 10 different speakers (5 male and 5 female) are stored for each CV combination, which gives 1600 prototypes.

The output of the comparison is a list of classified CV candidates. The analysis of cues relative to the best candidates allows the construction of the solution using a vote procedure. For example, if among the 10 best consonant candidates, 9 are voiced, 8 acute, 1 compact, 0 continuant, 1 nasal, 1 vocalic, the module proposes [d] as solution for the consonant, because [d] is voiced, acute, non compact, non continuant, non nasal and non vocalic.

[1] in velar context
[2] non velar context

3. THE HIGH-LEVEL MODULES

3.1. Access to a dictionary

Some methods of automatic spelling-correction compute a distance between a reference-word and a test-word; they rely on a series of operations that model errors of insertion, deletion and substitution. It is possible to realise these operations using dynamic programming [7]. The module for lexical access (Fig.1) is inspired by this method.

In our case, distance is not computed between graphemes but between a decoded phonetic string and the phonetic representations of words stored in a dictionary (Fig.2). Dynamic programming efficiently integrates in a single algorithm all the phenomena of insertion, deletion and substitution that can appear in bottom-up decoding. The comparison between test and reference words 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), a more sophisticated measure is required in the case of phonetic strings. 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 permits a 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/.

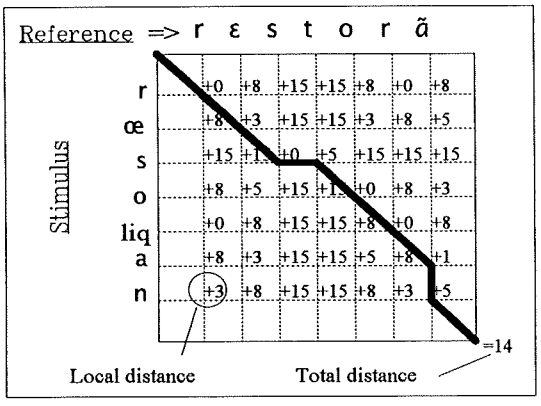

Figure 2 : Calculation of distance between 2 phonetic strings.

This method could be useful in evaluating, in a precise lexicon, the degree of difficulty of decoding: if, globally, a great distance was calculated between each word of the lexicon and the rest of the dictionary, it could indicate an easy task of decoding.

This method also requires phonological rules that 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], [pøtit], [ptitø] or [pøtitø] and should have 4 phonetic entries in the dictionary.

3.2. Supervisor

The supervisor process (Fig.1) is in its preliminary version. In a first step, it collects the different results of the bottom-up decoding (Table 1).

Table 1: Bottom-up decoding with the word "dictée" (French word for "dictation") pronounced /dicte/

columns : t=frame number, *Segm.*=automatic segmentation, *Ana.*=analytic decoding, *Global*=global decoding
abbr : SIL=silence, OCV=voiced stop, VOY=vowel, CSN=unvoiced consonant, OCN=unvoiced stop, è=/ɛ/, Ê=/ɛ/ or /e/, Ë=/ɛ̃/

t	*Segm.*	*Ana.*	*Global*	*t*	*Segm.*	*Ana.*	*Global*
1	SIL	.	-	33	OCN	.	-
2	SIL	.	-	34	OCN	.	-
3	SIL	.	-	35	OCN	.	-
4	.?.	.	-	36	OCN	.	-
5	OCV	.	bdvz	37	OCN	.	ptkfs
6	OCV	.	bdvz	38	OCN	.	ptkfs
7	OCV	.	bdvz	39	OCN	.	ptkfs
8	OCV	.	bdvz	40	OCN	.	ptkfs
9	OCV	.	bdvz	41	VOY		ÊË
10	OCV	.	bdvz	42	VOY	èy	ÊË
11	OCV	.	bdvz	43	VOY	èy	ÊË
12	OCV	i	bdvz	44	VOY	èy	ÊË
13	VOY	i	i	45	VOY	èy	ÊË
14	VOY	i	i	46	VOY	èy	ÊË
15	VOY	ie	i	47	VOY	èy	-
16	VOY	ie	i	48	VOY	è	-
17	VOY	ie	i	49	VOY	è	-
18	VOY	ie	-	50	VOY	è	-
19	VOY	ie	-	51	VOY	.	-
20	VOY	e	-	52	VOY	.	-
21	VOY	.	-	53	VOY	.	-
22	VOY	.	-	54	VOY	.	-
23	.?.	.	-	55	VOY	.	-
24	CSN	.	-	56	VOY	.	-
25	CSN	.	-	57	VOY	.	-
26	CSN	.	-	58	VOY	.	-
27	CSN	.	-	59	VOY	.	-
28	CSN	.	-	60	VOY	.	-
29	CSN	.	-	61	SIL	.	-
30	OCN	.	-	62	SIL	.	-
31	OCN	.	-	63	SIL	.	-
32	OCN	.	-	64	SIL	.	-

The syntagmatic information, i.e. the distribution of phonemes, is provided by the automatic segmentation (Table 1, column *Segm.*). This module also provides the macro-classes for the consonants. The precise identification of vowels is given by the analytic and global recognition modules (Table 1, columns *Ana.* and *Global*). For the time being, combining all these knowledge sources, the supervisor builds different phonetic string candidates using a very simple strategy (Table 2). We plan to improve these methods of decision, especially in case where an inconsistency is detected.

Table 2: Phonetic string candidates

[OCV][i][CSN][OCN][ɛ,e]
[OCV][i][CSN][OCN][ɛ̃]
[OCV][i][CSN][OCN][y]
[OCV][e][CSN][OCN][ɛ,e]
[OCV][e][CSN][OCN][ɛ̃]
[OCV][e][CSN][OCN][y]

In the case of isolated words, access to a dictionary allows the supervisor to provide a ranked list of word candidates (Table 3). This operation is done by comparing each phonetic string decoded with all the entries of a dictionary as described in § 3.1.

Table 3: Lexical access

columns: 1 = position, 2 = decoded phonetic string, 3 = orthographic form of the word in the dictionary, 4 = phonetic form of the word in the dictionary, 5 = distance between the phonetic string of the decoded word and the phonetic string of the stored word
abbr : Ê=/ɛ/ or /e/

pos	decod.phonetic string	orthogr.	phonetic	dist
1	[OCV]i[CSN][OCN]Ê	dictée	diktÊ	0
2	[OCV]i[CSN][OCN]y	discute	diskyt	2
3	[OCV]e[CSN][OCN]Ê	goûter	gutÊ	5
4	[OCV]i[CSN][OCN]Ê	quitter	kitÊ	5
5	[OCV]i[CSN][OCN]Ê	bonté	bõtÊ	5
6	[OCV]e[CSN][OCN]y	dessus	døsy	5
7	[OCV]Ê[CSN][OCN]ɛ̃	veston	vÊstõ	7
8	[OCV]i[CSN][OCN]Ê	discret	diskrÊ	8
9	[OCV]Ê[CSN][OCN]ɛ̃	latins	latɛ̃	8
10	[OCV]e[CSN][OCN]Ê	poster	postÊ	10

In the example presented, the appropriate word "dictée" has been placed in first position (Table 3). A top-down verification process, which is to be developed, would eliminate candidates such as "goûter" pronounced /gute/ or "bonté" pronounced /bõte/ (Table 3) because in these words, the first vowel is grave, which is in contradiction with the stimulus where the first vowel is definitely acute.

4. RESULTS

In a first test, each module has been evaluated independently using 10 French speakers (5 male, 5 female) recorded in the corpus SYL of the French database BDSONS. Stimuli were CVCV non-words such as "titi", "rara", "sussu",... that represents the combination of 2 * 10 vowels * 16 consonants * 10 speakers = 3200 tokens. Results have been published [8].

In a second test, we have evaluated the system with all the modules working together. A corpus[3] of 500 French isolated words recorded by 6 speakers (5 males, 1 female) has been used. For this operation, we have constructed a dictionary containing 500 orthographic entries corresponding to the 500 utterances. 'Standard' pronunciations of these words have been found in the BDLEX[4] lexicon. The recognition is completely speaker-independent. The supervisor has a very simple strategy and no top-down verification is done. 6 speakers * 500 words = 3000 utterances have been tested. For each test, the result consists in locating, in the ranked list of words-candidates (ex: Table 3), the place of the tested-word. In 47% of the case (Fig.3), the tested-word was in first position. In 70 % of the case, it was in the first ten places. In 86% of the case, it was in the first fifty positions.

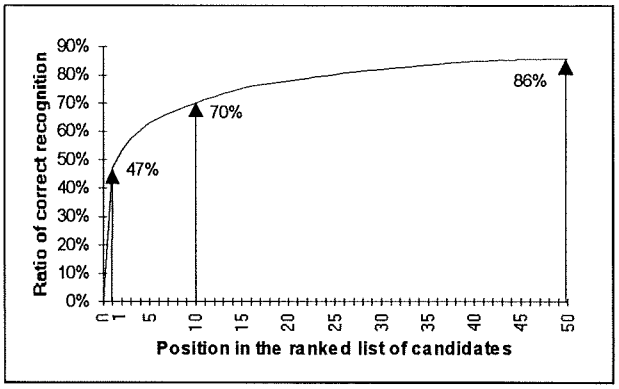

Figure 3 : Preliminary results of recognition using a corpus of 500 words with six speakers.

These results are encouraging and can be easily improved. Bad results come from the partial inadequacy between the phonetic strings proposed by the supervisor and the dictionary, especially in the case of semi-vowels such as /jwɥ/.

CONCLUSION

We have presented the different parts of a system based on parallel distributed processes for speaker-independent acoustic-phonetic decoding. All these knowledge-sources collaborate in a single system. The bottom-up decoding is working efficiently. Our aim now is to improve the inference rules of the supervisor and to integrate a top-down verification step to eliminate the candidates which are clearly incorrect.

REFERENCES

[1] Zue V.W., Lamel L.F. (1986), "An expert spectrogram reader: a knowledge-based approach to speech recognition.", *Proc. ICASSP*, Vol.2, pp. 1197-1200

[2] Carbonnel N., Haton J.P., Fohr D., Lonchamp F., Pierrel J.M. (1986), "APHODEX, design and implementation of an acoustic-phonetic decoding expert system", *Proc. ICASSP*, Vol.2, pp. 1201-1204.

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

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

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

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

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

[8] Ghio A., Rossi M. (1995), "Parallel distributed processes for speaker-independent acoustic-phonetic decoding", *Proc. 13th Int. Cong. Phon. Sc.*, Stockholm.

[3] we are grateful to the *Laboratoire d'informatique de l'Université d'Avignon, France* for generously offering us this corpus.

[4] BDLEX is a French lexical and phonological database. This project is headed by G. Pérennou, IRIT, Toulouse, France