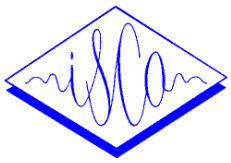


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

RULE-DRIVEN NEURAL NETWORKS FOR ACOUSTIC-PHONETIC DECODING

Rémy Bulot, Henri Meloni, Pascal Nocera

Groupe d'Intelligence Artificielle, Luminy Science Faculty,
case 901, 163 av. de Luminy, 13288 Marseille, France

ABSTRACT

We are presently developing an Acoustico-Phonetic Decoding system which uses a Prolog II rule base combined with various neural networks. The rules organize the learning process by choosing relevant examples from a data base of sound. Their essential role in recognition is to describe the structure of the sounds ; they select the input data for the networks from the signal, and interpret their output according to the context. Different strategies were used for localizing and identifying vowels, fricatives and occlusives (depending upon the acoustic features of each macro-class); several network architectures were tested in parallel.

I. INTRODUCTION

Many Automatic Speech Recognition systems include an Acoustico-Phonetic Decoding phase (APD) based on the processing of explicit knowledge provided by phonetic experts. This type of technique is particularly useful for the confrontation and validation of models proposed by the experts, and in addition they enable rapid integration of new knowledge. However, there are two major problems with this "expert system" approach. The first problem, which applies especially to speech, is that there is no ideal distance from which to measure acoustico-phonetic phenomena, thus introducing a degree of uncertainty into the characterization of these events. The second problem which applies to automatic theorem proving in general, and affects the structural aspects of speech, is that the data used is not completely reliable or is imprecise, making it difficult to evaluate the real value of their combinations.

Connectionist methods seem particularly useful in APD because of their capacities for classification and generalization (i.e. they are capable of responding correctly to a degraded stimulus). However, their use is still at the prototype stage in limited applications. The most serious difficulties are to be found in the distorsion of the time axis between two executions of the same sound, and in the selection of relevant information (especially in continuous speech).

In this paper we present the use of a Prolog II rule base combined with networks, in an APD system which is intended to process single speaker continuous speech (we have preferred to envisage adaptation to the speaker rather than moving towards a multi-speaker system [1]). The rules describe the structural aspect of the different sounds of French and use several networks to measure the relevance of numerous acoustico-phonetic phenomena.

II. ENVIRONMENT FOR SPEECH RECOGNITION

We use a Prolog II environment adapted for Speech processing [2] [3]. The flexibility and precision of the available tools, in particular

- multiple parameters of the same portion of signal,
- shape recognition (diagrams of peaks, of valleys, ...),
- definition of constraints,

make it possible to describe acoustic and phonetic events as well as the specific contexts in which they are relevant. The system includes two data bases of sound containing sentences with most "consonant+ vowel" contexts : one is used for learning and one is used to provide recognition scores.

This environment has been enriched by a certain number of external predicates which define and

manage neural networks. These are multi-layer networks with no cycle and whose node values are contained in the interval [0, 1]. The values of the connections are adjusted by the gradient back-propagation algorithm [4] [5]. These networks are essentially used to recognize shapes in the signal and are rule-driven in the following manner :

- in the learning phase, the rules select the relevant examples in a previously labelled sound base,
- in the recognition phase, the rules interpret the responses of the different networks to classify and identify the sounds.

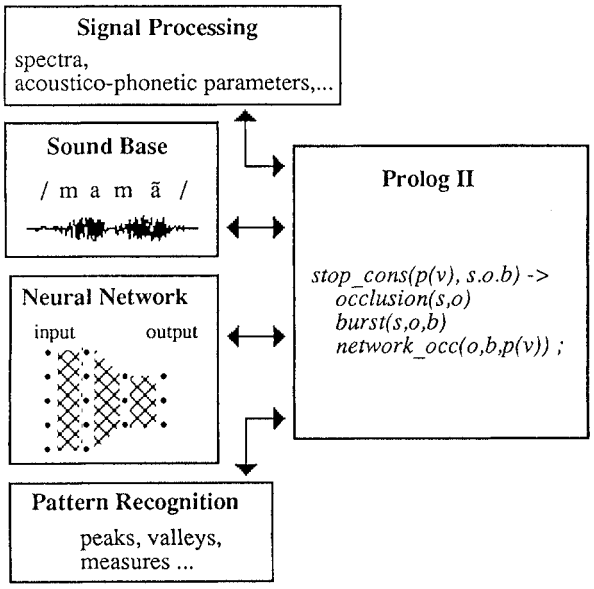


Figure 1 : tools used under Prolog.

III. CHARACTERIZATION OF FRICATIVES

The aim is to use a network to characterize the signal in hundredths of seconds so that we can localize and identify the fricatives. Because the sound spectrum in this macro-class is relatively stable, we have restricted input data to a single spectrum. Our network consists of 24 input cells (representing a spectrum according to a Mel scale on 24 channels), and 5 output cells for / f s ʃ z ʒ / (we renounced to characterize /v/ in this manner). After experimentation with several types of network, we have selected an architecture with just one hidden layer. Learning was performed on the "stable" part of the fricatives in the data base of sound, and on a certain number of counter-examples taken from silences, occlusions, nasal consonants (/m/, /n/) and vowels (there was a danger that the spectra in transitions, bursts and /r/ consonant might cause contradictory learning).

In recognition, the system inputs to the network all the spectra of the statement in succession, and records the response curves of the five output cells (figure 2). These curves are then analyzed by the APD rules which give the signal a non-deterministic label depending on the context. However, these rules are currently being developped ; the first tries with 200 fricatives in continuous speech are already encouraging :

pho.	/ f /	/ s /	/ ʃ /	/ v /	/ z /	/ ʒ /	¬fric
/ f /	84	5	0	0	0	0	11
/ s /	0	94	0	0	0	0	6
/ ʃ /	0	0	100	0	0	0	0
/ z /	0	0	0	0	93	0	7
/ ʒ /	0	0	0	0	0	100	0

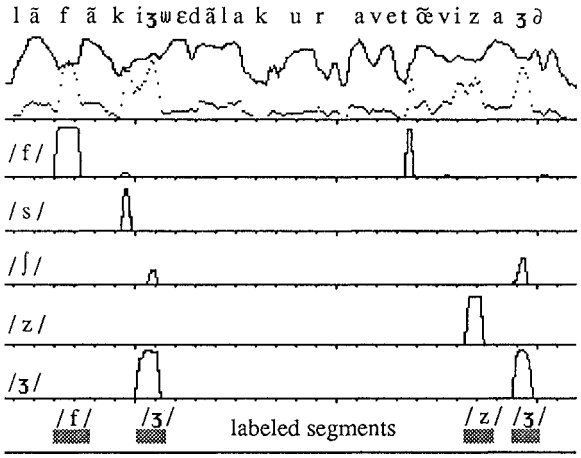


Figure 2 : response curves of the output cells and segments obtened with rules.

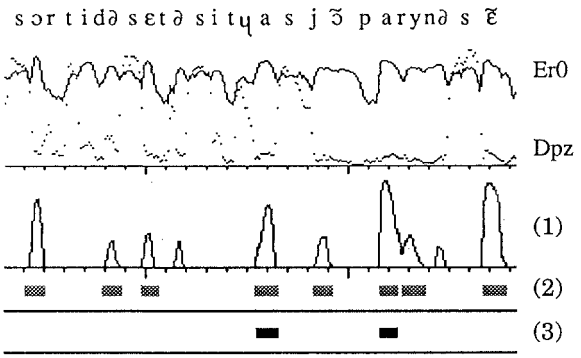
IV. RECOGNITION OF ORAL VOWELS APPLICATION WITH /a/

For the case of the vowels, we prefer that the influence of the left context and the right context are directly taken into account by the neural network. Here, the approach is completely different, because rules preceed the network and they propose to it particular places with which the caracterisation

should be made. We present an example with a specialized network for /a/. The caracterisation of the other oral vowels are being developped with the same method.

The localisation of /a/ is made with a spectral distance and a reference spectrum which is chosen in a neutral context. A rule construct a temporal parameter of "likeness" (figure 3) on which we research all the hills (i.e. all the segments which can be the vowel /a/). The spectral distance is voluntarily not too selective to accept all the realisations of /a/ in most deformed contexts ; on the other hand, a lot of nasal vowels (/ɔ̃/, /ã/, /œ̃/, /ɛ̃/) or half-open vowels (/ɔ/, /œ/, /ɛ/) are detected as a possible /a/. Three spectra are chosen in the middle and at the limits of the segment (where the sound diverge from /a/) ; these extremities are sensible to represent acoustic traces of deformations resulting of the phenomena of coarticulation. These three spectra are proposed by the rules to the neural network for the final decision ; only segments with a score of more than 0.5 are memorized. In order to have the same situation in the learning phase and in the recognition phase, examples for learning have be selected in the data base of sound with the same localisation rules.

The rate of recognition for /a/ is 88% with the modul "rules+network". The confusions represent 10% of other vowels (nasal or half-open) ; 40% of these confusions are made with the sound /ã/.



(1) : likeness curve for /a/,
(2) : possible segments /a/ detected by rules,
(3) : segments selected by the network.

Figure 3 : three steps for the recognition of /a/.

V. IDENTIFICATION OF STOP CONSONANTS

In contrast to the two previous macro-classes, execution of occlusives is split into two phases: an occlusive part and a transient part which is the burst. This 2nd phase, which is difficult to detect in the signal, is particularly sensitive to coarticulation phenomena.

We have a set of rules (approximately one hundred) which enable us to detect in the signal the segments likely to correspond to occlusives (figure 4). In addition to the localization, these rules provide a "relevant" spectrum for each of the probable occlusive and burst phases. In order to take into account the right-hand context (the most influential), we also use a third spectrum chosen 30 ms after the burst.

```
stop_consonant(seg, pho(val)) ->
    stop_segment(seg, occ)
    burst(seg, occ, burst)
    add(burst, 3, ctxt)
    ident_occ(<occ, burst, ctxt>, pho, val) ;
```

Figure 4 : stop consonants are detected in the sentence and enumered by bactracking .

This data is given to a network with a similar architecture than the precedent (figure 5) and the result is an ordered list of scored candidates (we just select the stop consonants with a score more of than 0.5).

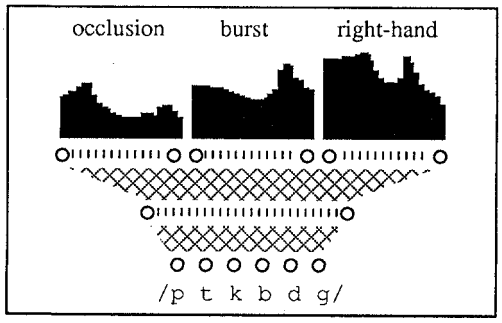


Figure 5 : architecture of the network for the identification of stop consonants.

The mistakes which occured in recognition with the system "rules+neural network", can be summarized

the matrix of confusion below. Sometimes, it is possible that the consonnant which must be recognized, is only in a second position in the list of the candidates, but nevertheless with a significant score and not too distant from the first candidat. This situation seem a satisfactory result. So, we are giving a supplementary rate (colon 1&2) which represent the number of occurrences when the stop consonant is in first or second position with a significant score (more than 0.5). The line "¬occ." represents the classification of the segments detected by the rules and which are not stop consonant.

pho.	/p/	/t/	/k/	/b/	/d/	/g/	¬occ	1&2
/p/	71	8	8	0	4	4	5	86
/t/	7	83	6	0	3	0	1	91
/k/	10	14	67	0	0	0	9	78
/b/	2	0	0	78	2	5	13	87
/d/	0	1	2	4	85	3	5	92
/g/	3	0	10	3	10	65	9	74
¬occ	3	1	0	7	7	1	81	96

Figure 6 : matrix of confusion for occlusives.

VI. CONCLUSION

By using a Prolog II rule base, it is possible to choose the relevant information and validate the responses by intervening in the networks upstream and downstream. This applies to both learning and recognition. This approach enables networks to be freed from the representation of knowledge which is already mastered and which they might not be able to model. Connectionist methods provide the additional flexibility required by automatic theorem proving, in order to represent the ambiguous nature of speech. However, we tried numerous configuration of networks to determine the optimal number of nodes in the hidden layer and to allow a good generalization of the phenomena studie.

Contrary to other research in this domain [6], we prefer to globally recognize these phenomena with relevant spectra without use of an intermediate step for the identification of acoustic, phonetic or articulatory cues. Indeed, we think that a decomposition with cues is interesting for the qualittive level, but penalising for the quantitative level (for example, in the valuation of the conjunction of scored cues which determine each phonemes). However, we consider the characterization of macro-class where the identification of phonemes is bad. This partial information (like the detection of nasal consonant without more precision) must permit to limit the ambiguity in numerous cases, especially in connection with lexical access.

BIBLIOGRAPHIE

[1] M. Rossi, *De la Quiddité des variables*, Variabilité et spécificité du locuteur, séminaire du 20 et 21 juin 89 au CIRM, Luminy, Marseille.

[2] Bulot R., Méloni H., *Reconnaissance de formes et localisation d'événements acoustiques et phonétiques.* Journal d'Acoustique, Vol 1, n°3 sept 88.

[3] Méloni H., Bulot R., *Processing acoustic and phonetic knowledge in Prolog ;* T. O'Shea and V. Sgurev Editors, Elsevier Science Publishers B. V. North-Holland, IICR, 1988, pp 177-185

[4] F. Fogelman Soulié, P. Gallinari, Y. Le Cun, S. Thiria. *Automata Networks and Artificial Intelligence.* dans "Automata networks in computer science", de F. Fogelman Soulié, Y. Robert, M. Tchuente Eds, Manchester Univ. Press, (1987) , p 133-186.

[5] D. E. Rumelhart,J. L. Mc. Clelland and The PDP Research Group. Parallel Distributed processing : Explorations in the microstructure of cognition Vol 1: Foundations.Chapter 8 : *Learning Internal Representations by Back-Propagation* MIT Press. (1986)

[6] Y. Bengio, R. Cardin, R. De Mori, E. Merlo, *Programmable exécution of multi-layered networks for automatic speech recognition,* Communication of the ACM, feb 89, Vol 32, n° 2.