A Knowledge-Based System for Speaker-Independent Recognition of Letters

H. Méloni, A. Betari, P. Gilles

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

I - INTRODUCTION

The major difficulties for speaker-independent recognition of letters in French result from the acoustic likeness of some words ("M" and "N", "B" and "D", "F" and "S", "J" and "G", etc.) and from the lack of information to solve ambiguities. This particular vocabulary requires the encoding of the main informations taken into account for the acoustic and phonetic decoding of continuous speech.

The techniques more frequently used to build up such systems are based on pattern matching. There are various pattern-coding and pattern-matching methods (Vector Quantization, Hidden Markov Modelling, Neural Networks, Dynamic Time Warping, etc.) [Aldelfeld 80], [Burton 85], [Jelinek 85], [Huang 88], [Bulot 89], but most of them do not use explicitly coded knowledge.

We are proposing a system based on a set of acoustic, phonetic, phonologic and lexical knowledge, represented by rules in Prolog II [Méloni 86]. The informations represented are various and they allow the localization and the identification of acoustic and phonetic phenomena (patterns, grouping of patterns, events, properties, cues, features, phonemes, syllables, etc.). A score is associated to each phenomenon described when the latter is identified. Rules and control use these scores in many ways to assign valuations to complex phenomena and to sort the most probable hypotheses of recognition.

II - SIGNAL PARAMETRIZATION

II.1 - Standard parameters

The signal is digitized at a 12.8 kHz frequency, and coded in a 16 digit word, then characterized at 10 ms intervals by its total energy, its zero-crossing density and its spectral energy in 24 channels distributed according to a Mel scale (fig. 1). The spectra can be processed by means of different methods such as FFT, LPC, Cepstrum, Vocoder, cochlear simulation [Caelen 79]. In the environment where knowledge is encoded and processed (Prolog II), the parameters are retrievable by means of predefined rules which link the numeric data and the symbols used in the rules.

II.2 - Provisional parameters

The phenomena useful to localize or identify phonetic units are best characterized by means of parameters measuring differences in energy for particular spectral bands. Such attributes are obtained by means of predefined rules which make it possible to carry out simple operations with standard parameters or newly made parameters (sum and difference in the energy in two frequency bands, localization of spectral peaks and valleys, time-difference, instability, etc.). The provisional parameters are defined in Prolog II by means of rules using predefined predicates. Thus, the clause:

\[ \text{Peak}_{0}(z,n-par) \rightarrow \text{Pic}(z,<0.5600>,n-par) \]

defines Peak0 as the parameter measuring the value of the first spectral peak in the frequency band <0.5600>. We have defined more than a hundred provisional parameters which represent the position and the amplitude of spectral peaks in given bands, the energy density in spectral bands, the differences of energy density in bands or peak amplitudes, and more complex parameters such as the position of the first two formants, the rise up of particular peaks, or general rules calculating various types of instability for any attribute.

III - PATTERN RECOGNITION

III.1 - Description of patterns
The localization of acoustic events can be processed with great precision by means of patterns on parameters [Bulot 88]. The environment in which we describe the knowledge makes it possible to define and identify simple pattern schemes (peaks, valleys, plain segments) on the curves representing the time evolution of any parameter. A pattern scheme is described by means of a rule using one of the evaluable predicates parametrized according to the group of patterns to be identified [Meloni 86]. Thus the clause:

\texttt{PeakO-Dpz(Dpz,z0,z) -> colline(Dpz,z0,10,15,35,35,5,z);}

defines a certain type of peak on the parameter measuring the zero crossing density. The characteristics defining the scheme and the deletion constraints of the rule \texttt{colline} are controlled by the values of the variables.

In the application to speech recognition of the French alphabet, we defined 29 pattern schemes, most of which are pertinent for a number of parameters.

III.2 - Pattern handling

One of the principal use of patterns is the spotting of intervals in speech signal. The description of acoustic and phonetic units is obtained by grouping several of these patterns associated as regards their position in time.

Our environment makes it possible to define relations and operations in time intervals which are used to compare the positions of two forms (coincidence, inclusion, adjacency, etc.) or to create new intervals (intersection, merging, complement, etc.). Fig. 2 shows that for an emission of the letter "B" (/b e/), there is always a peak of the parameter \texttt{Ebf1-Esp} (difference of the energy density between very low frequencies and average frequency). This peak characterizes the buzz during the occlusion of /b/, and at the burst of /b/ it is immediately followed by a narrow peak of the parameter \texttt{Emf-Ebf} (difference between the energy in the middle of the spectrum and the low frequency energy).

IV - ENCODED INFORMATIONS

Only informations pertinent in a particular context of speaker-independent recognition of emitted letters are encoded and treated. The description of the possible emissions of the 26 words required about 500 rules defining various acoustic and phonetic phenomena such as patterns, pattern associations, events, macro-classes, properties, cues, features, phonemes, groups of phonemes. To each rule describing a relevant phenomenon is associated a value calculated by means of the weighted scores of the elements the rule uses.

IV.1 - Word descriptions

Each word is described by means of minimal elements which make it possible to differentiate them from other words. Most of the time we only keep macro-classes marking the occurrence of a certain type of phoneme. Such is the case, for instance, for the description of the letter "X" (\(\bar{b}k\)) in which the occlusion between the phonemes /b/ and /s/ is enough to point out the word without ambiguity. A letter (here "Q") is described in Prolog by means of the clause:

\texttt{letter("Q",z0,z,<4,v1>,<3,v2>,v3,v4) ->
vocalic-interval(z0,z2)
vowel-UU(z2,v1)
KK-before(z0,z1,v2)
silence-before(z1,v3)
silence-after(z0,v4)
union(z1,z0,z);}

We calculate the score of the letter "Q" by means of the list of scores \(v_i\) weighted by \(a_i\). This rule scheme makes it possible to describe the letters "B", "C", "D", "G", "J", "K", "P", "T" and "V". For the sequence "F", "H", "L", "M", "N", "R" and "S", the consonant is to be found after the initial vocalic unit. In order to take into account arbitrary emissions a mute /\(\bar{d}\)/ - particularly by southern French people - we search the optional occurrence of this unit before the silence. The letters "A", "E", "I", "O" and "U" can only be expressed by means of a vowel with a silence before and after. To differentiate them from similar words such as "K", "P" or "T", we defined a non-event characterizing the non-occurrence of a burst. The letters "W", "X", "Y" and "Z" contain more units more or less precisely described.

IV.2 - Description of phonemes

Phonemes are described by means of macro-classes, acoustic and phonetic cues, pseudo-phonetic features, events, groups of patterns and properties. In Prolog a phoneme is represented by clauses such as:
First the vowels are localized by means of macro-classes describing four types of vocalic units: Except for schwa (directly defined by pattern associations on given parameters) the discrimination is made by means of pseudo-phonetic features the redundancy of which makes it possible, first to take into account the various emissions, then to characterize more accurately phonemes in quite similar words (formed with the vowels /U/, /y/ and /w/).

For the consonants, several macro-classes ("silence", "occlusive", "fricative", "consonant", "inter-vocalic-consonant", "final-consonant", "nasal consonant", "vocalic-consonant", "non-nasal-consonant") make it possible to characterize them summarily in the various contexts in which they can be found. The fricative consonants are defined by means of the following features: voiceless, voiced, low, high-pitched, compact, separated and flat. For the occlusives the features are in a larger number and often they are attributed to one phoneme. They are associated as well to the occlusion as to the burst and they are usually organized into a hierarchy. Moreover, some features are used to differentiate the importance of bursts and make it possible to differentiate words such as "A" and "K". The two nasal consonants /m/ and /n/ are differentiated from the others by the feature "nasal" and from each other by the features "low" and "high-pitched" (respectively associated to /m/ and /n/). The phonemes /l/ and /l/ both classified as "vocalic-consonant" and "non-nasal-consonant" are also characterized by the features "compact" and "separated" which are attributed respectively to /l/ and /l/.

IV.3 - Description of the macro-classes

We call pseudo-phonetic macro-class a symbol which represents a coherent group of phonemes characterized by the same set of acoustic and phonetic phenomena. We defined a dozen of labels identifying more or less important non-separated classes.

\[
\text{vowel-El}(z_0, v_1, v_2, <z_3>, v_4) \rightarrow \\
\text{half-closed-vowel}(z_0, v_1) \\
\text{high-pitched-vowel}(z_0, v_2) \\
\text{non-sharpened-vowel}(z_0, v_3) \\
\text{non-separated-vowel}(z_0, v_4) ;
\]

The features we use have been mentioned in the description of the phonemes. They are often quite different from those used by phoneticians to describe French sounds [Rossi 1977]. These features are classified according to a hierarchy and the acoustic phenomena defining them (cues and properties) depends on the phonemes they characterize. For instance, the feature "high-pitched" describing the dental nasal consonant /n/ is completely different from the feature describing the dental fricative consonant /s/.

Each feature is represented by a dozen of rules classified according to an approximate measure of the reliability of the results. Each possible representation of the features is directly encoded in a rule by means of associations of patterns and/or properties on various parameters. For example, one of the features most difficult to define is the feature "high-pitched" differentiating the nasal consonant /m/ and the nasal consonant /n/ for the emissions of the words "M" and "N". The different observed and encoded phenomena are essentially the time-evolution of the spectral energy around the second formant.

IV.4 - Description of features

Acoustic and phonetic events characterize definite stages of phonemes (burst, release, etc.). They usually make it possible the localization of phonetic units by means of properties reliable as regards pattern recognition. Such is the case when we want to localize vocalic nuclei, constractive consonants, occlusives consonants, bursts, intervals of stable formants, etc.

Acoustic and phonetic events are usually characterized by coincidences in patterns for various parameters (fig. 3).
IV.6 - Description of cues and properties

Cues and properties point out acoustic traces of the articulation of sounds. Consequently we have to characterize stable phenomena (buzz, resonance, etc.) as well as the transitions between phonemes. These elements are seldom represented by rules but appear in clauses as a specific relations between parameters (comparison of their average value in time intervals, limits of their variations, various pattern associations, etc.).

V - RECOGNITION CONTROL AND RESULTS

It is possible to consider a number of different ways to control the recognition process. The aim of some of them is to improve the efficiency of the recognition process, the aim of others is to refine and improve the results in particularly difficult cases when words are very similar.

As our aim is the testing of our technique of representation of acoustic and phonetic knowledge, we used the less efficient strategy which consists in looking for every word whenever a vocalic nucleus is detected. The valuation of all the identified phenomena and their weight in the description of more complex symbols make it possible to give more importance to the characteristics which seem the most relevant in the context of our system. This is another way of encoding the process of recognition control and calculating likely results in a non-deterministic way.

The results were evaluated for ten speakers (4 women, 6 men) who spoke the French alphabet several times in natural circumstances including noise. The rate of recognition was as follows: 93% of words correctly recognized in first position, 4% of words identified in second position, and 3% of words not recognized.

These results are appreciably better than those obtained by means of connexionist technique which do not explain all the knowledge [Bulot 1989]. Performances are improved principally in some particular cases where words are very similar (chiefly "M" and "N"). The dynamic use of definite contextual informations can explain the difference between the performances of the two systems.

The principal errors appear about the following pairs of words ("M" and "N", "B" and "D", "J" and "G", "A" and "K"). However this work was completed in three months' time, therefore we think that some of the rules could be perfected and consequently the performances would be improved.

VII - CONCLUSION

The work we have just presented shows the possibilities of representing in an explicit way acoustic and phonetic knowledge and their treatment in an environment based on the language Prolog II. These techniques show that in difficult cases they are the more able to characterize the subtle elements which make it possible to differentiate sequences of similar sounds.

However to score a phenomenon in a given context, we think that the various methods of pattern recognition are more efficient than rules to give the valued symbols making up larger units. The models of such phenomena require input data which can be quite different according to context (specific parameters, time evolution within a variable range, etc.). This knowledge needs to be explicitly used to encode and choose, with the utmost accuracy, the representative elements of the phenomena to be identified.

This application allowed us to assert that any speech recognition systems should use both explicit representation techniques for some informations, and pattern recognition for non-evident other ones.

REFERENCES

ALDEFELD B., RABINER L.R., ROSENBERG A.E.
Automated directory listing retrieval system based on isolated word recognition ; Proc. IEEE, Vol. 68, pp. 1364-1379, 1980

BULOT R., MELONI H.

BULOT R., NOCERA P.
A system for speech Recognition using both connexionist methods and knowledge ; European Conference on Speech Recognition, Paris novembre 1989

BURTON D.
Applying matrix quantisation to isolated word recognition ; Proc. IEEE ICASSP, pp. 29-32, 1985

CAELEN J.
Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonétique ; Thèse d'Etat Sciences, 1979

HUANG S., GRAY R.M.
Spellmode recognition based on vector quantization ; Speech Communication Vol.7 n° 1, North-Holland, pp. 41-53, 1988

JELINEK F.
The development of an experimental discrete dictation recognizer ; Proc. IEEE, Vol. 73, pp. 1616-1624, 1985

MELONI H., BULOT R.
A knowledge based system for acoustic and phonetic decoding of continuous speech ; Congrès International d'Intelligence Artificielle de Marseille, 1986

ROSSI M.
Les traits acoustiques ; La Linguistique, 13, pp 63-82, 1977