EXPLICIT KNOWLEDGE AND NEURAL NETWORKS
FOR SPEECH RECOGNITION

Bulot R., Nocéra P.

G.I.A. Luminy Science Faculty, Marseille, France

Abstract

In this paper we present a simple speech recognition application in which an explicit knowledge base system is coupled to a connectionist machine. A set of Prolog rules describes the structural aspects of the words in the vocabulary, in terms of acoustico-phonetic events. These events are jointly evaluated by a multi-layer network in order to identify the corresponding word. The system provides good results, notably with unknown speakers.

I. Introduction

Many Automatic Speech Recognition systems include an Acoustico-Phonetic Decoding (APD) phase, based on the processing of knowledge provided by a phonetic expert [De Mori 82], [Carbonnel 84], [Gillet 84], [Caelen 86], [Stern 86], [Zue 86] [Bulot 87, 88]. These techniques are particularly useful for the validation and confrontation of the models proposed by the expert. In addition they enable the rapid integration of new knowledge. However, there are two major problems with the "expert system" approach: the first, relevant to speech in particular, is that there is no ideal distance for measurement of acoustico-phonetic phenomena, which introduces a certain amount of uncertainty when characterising these events. The second problem applies to automatic theorem proving in general, and concerns the structural side of speech: the information used is not completely reliable or is tainted with imprecision, making it difficult to evaluate the real value of their combinations.

Among the methods based essentially on stochastic models, a return to connectionism is taking place in the domain of APD [Bourlard 87] [Robinson 88] [Nocéra 88] due to new learning techniques (gradient back propagation, simulated anneal etc.). This is because neural networks have particularly interesting 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, because a speech signal carries a very large quantity of information which cannot be considered directly by a network with a fixed and finite number of inputs, (in particular because of the time axis distortion between two different executions of the same word).

In this paper we present the combined use of a network and a rule base in a restricted application; the multi-speaker recognition of letters in the French alphabet. Although it appears to be very simple, in fact the problem presented by the recognition of 26 words representing the pronunciation of letters is considered as difficult by phonetic experts. The main difficulties stem from the acoustic proximity of certain words (B and D, J and G, M and N, F and S, etc.) and from the absence of syntactic and semantic information to resolve ambiguities. In the system to be presented below, the main phases in the execution of a word are described in an explicit manner by a set of Prolog rules and the identification of the corresponding acoustico-phonetic events is performed by a multi-layer neural network.

II. Description of the environment

We use a Prolog II environment in which it is possible to define and process complex acoustic and phonetic knowledge (parameters, patterns, relations etc.) used to designate the limits of many characteristic phenomena in French language sounds. The flexibility and precision of the available tools make it possible to describe acoustic and phonetic events as well as the specific contexts in which they are relevant.
II. 1 Parameterizing the signal

The speech signal is characterized at 10 ms intervals by its global energy (ErO), the density of zero passages (Dpz) and the spectral energies in 24 channels distributed according to a Mel scale.

In the environment (Prolog II) where the knowledge is represented and processed, these parameters are recorded in an array indexed according to time, which can be accessed by numerous evaluable predicates that provide the links between the numeric data and the symbols manipulated in the rules.

The phenomena which are of interest for the localization and identification of phonetic units are best characterized by parameters which measure differences in energy between certain spectral zones. These characteristics are obtained by means of predefined predicates which perform simple operations on the standard parameters and the newly created parameters (sum and difference of energies located in 2 frequency bands, identification of spectral minimums and maximums, instability, etc.) and which obey Prolog's general backtracking process. Each new parameter created in a given time zone is recorded temporarily in the array described above. There is no limit to the number of new characteristics, which can undergo the same processes as the initial parameters.

II. 2 Pattern recognition

The localization of acoustico-phonetic events can be performed very precisely by means of patterns on sufficiently representative parameters. The environment in which we describe the knowledge enables the definition and identification of schemes of simple patterns (peaks, valleys, monotonous segments) on the curves representing the evolution in time of any parameter. We have defined a certain number of relations and operations on time intervals to compare the positions of 2 patterns (coincidence, adjacency, inclusion, etc.), or to create new zones (union, intersection, complementarity, etc.). The relative positions of several patterns, in addition to some of their associations, constitute decisive information for the description of acoustic and phonetic events, phonemes, syllables and words in the vocabulary.

II. 3 Example of rules

In order to illustrate some of the possibilities of the environment, we give below a series of rules which describe a way of localizing the spectrum that best characterizes a voiced occlusion at the beginning of a letter (such as B, D, W). One way of detecting the presence of a buzz is to evaluate energy concentration in very low frequencies (0 - 300 Hz). The construction of the parameter which measures this phenomenon is described by:

\[ \text{Buzz}(z\text{-one}, \text{par-num}) \rightarrow \]
\[ \text{energy-density}(z\text{-one}, <0, 300>, \text{par-num}) \]
\[ \text{maximum-energy}(z\text{-one}, <500, 6000>, \text{par-num}) \]
\[ \text{less}(\text{par-num}, \text{par-num}', \text{par-num}) ; \]

The scheme of the minimum pattern characterizing a voiced occlusion is defined by:

\[ \text{small-energy-dif-peak}(p, z0, z) \rightarrow \]
\[ \text{construct-parameter}(p, z0, \text{parameter-num}) \]
\[ \text{peak}(\text{parameter-num}, z0, 0, 2, 5, 5, z) ; \]

and the position \( t \) of the spectrum is obtained by the following rule:

\[ \text{voiced-occlusion}(z0, z, t) \rightarrow \]
\[ \text{small-energy-dif-peak}(\text{Buzz}, z0, z) \]
\[ \text{less-than-or-eq}(3, \text{length}(z)) \]
\[ \text{less-than}(\text{average}(\text{ErO}, z), 40) \]
\[ \text{middle}(z, t) ; \]

Associations of patterns on parameters

The context for the application of the above rule is limited to the 100 ms (z0) preceding a vocalic nucleus (also detected by rules), and thus depends on higher levels and on the recognition strategy.

\[ \text{previous-segment}(z\text{-vocl}, t) \rightarrow \]
\[ \text{zone-before}(z\text{-vocl}, 0, 10, z0) \]
\[ \text{or}(\text{fricative-nucleus}(z0, z, t), \]
\[ \text{voiced-occlusion}(z0, z, t)) ; \]
\[ \text{previous-segment}(i, j, t) \rightarrow \text{less}(j, 4, t) ; \]

buzz
dpz
erO

/ a / / be / / se /

/ a / / be / / se /
letter(a, b, c, d) ->
  search-zone(z0)
  voc-nucleus(z0, z, c)
  transition-to-voc(z, c, b)
  previous-segment(z, b, a)
  next-segment(z, d);

III Choosing the most relevant information

The time taken to pronounce a letter varies greatly from one speaker to another and on the average represents half a second of signal, i.e. more than 1,000 values after parameterization. However, because the alphabet represents a limited vocabulary, we can identify the words by means of a restricted set of criteria, based on their decomposition into principal phonetic events. The combinations of these events define the categories in the table of properties below.

| A, E, I, O, U | silence | vowel | silence |
| B, D, K, P, Q, T | occl | burst | vowel | silence |
| C, G, J, V, Z | fricative | vowel | silence |
| F, H, S, X | silence | vowel | fricative |
| M, N | silence | vowel | nasal |
| L, R | silence | vowel | voc. cons. |
| W, Y | multi-syllabic special cases |

Our recognition strategy leads us to search first for the vocalic segment of the letter, which subsequently acts as the anchor nucleus in the search for more subtle phenomena (see section II.3). The letters are identified with the same neural network, whose input nodes receive 4 spectra characterising the most distinct events. In order to ensure a certain coherence of segmentation between the different categories of letter, the spectra are chosen in the following way:

The first spectrum is chosen at the first detected event:
- a fricative nucleus before the vowel,
- voicing in the occlusion before the vowel,
- 40 ms before the "beginning" of the vowel.

The second spectrum is chosen at:
- the explosion if it is detected,
- the point where the energy increase is greatest at the beginning of the vowel.

The third spectrum is taken at the middle of the vowel.

The last spectrum is chosen at:
- the fricative nucleus after the vowel if it exists,

- the end of the vowel.

These different types of events are described in several ways by rules which use parameters and schemes of specific patterns (see section II.3). These rules provide the position of each localized phenomenon in the signal as well as the spectrum which best characterises it. Recognition of the letters Y and W does not often require the receptron, since the presence of two or three vocalic nuclei conforming to a few crude criteria enables simple identification. Nevertheless, the detection of a / ve / or / i / syllable (i.e. the output node associated with V or I in the neural network has a value close to 1) causes a search for the presence of vocalic nuclei before or after this syllable respectively, in order to conclude, possibly, that the letter W or Y is present.

Although the series of detected events makes it possible to restrict the set of solutions (for example, the series of silence-vowel-fricative macro-categories limits the set of solutions to {F, H, S, X}), we have decided not to take into account this information, by letting the neural network choose the 26 letters. This is because the localization rules are not completely reliable, and could lead to a fatal pre-classification of the pronounced word. In this way, an error in the choice of a spectrum can be compensated by the relevance of the other three. For example, the occlusive burst in B, D, P is not always detected, but the spectrum chosen at the point of greatest increase in energy nevertheless contains a part of the desired information. Identification of these words is all the less penalised because this type of example occurred during the learning phase, and enabled several classifications of the same letter to be established by the neural network (non-related classification). This is also the case for the letter R, for which the various executions may appear as rolled, strongly marked, or fricative, depending on the speaker.
IV Description of the neural network

The Neural network used to identify the letters of the alphabet is of the Multi-Level type, and contains three layers where each layer unit is linked to all the units of the next layer. It comprises:
- a layer of input units consisting of 96 units corresponding to the 4 spectra, each of which is divided into 24 channels,
- a layer of 50 hidden nodes, enabling non-linearity of the classification,
- a layer of 26 response units, each of which corresponds to a letter of the alphabet.

The connections between the units were initialised at random when the network was constructed. These connections were adjusted with the gradient back propagation algorithm [Fogelman 86]. The state of each output varies between 0 and 1, and the closer the pattern to be identified is to that of a node, the closer the value of this node will be to 1. In addition, it is possible to vary the acceptance threshold of a response, below which the neural network will not give any results.

IV Results and Conclusion

We recorded three male speakers who constituted the learning corpus, and the recognition performances were measured on three other speakers. The scores obtained on the speakers unknown to the system were 86%. The choice of spectra to characterise a letter remains rather simplistic and partly explains the mistakes often occurring in the identification of L, M and N. The recognition of the other letters is good, especially in difficult cases such as B, D, P, T.

Although this new environment loses some of its flexibility because of the learning phase required by the network, classification becomes very rapid on a corpus such as the alphabet. Indeed, the explicit description of the different phases in the 26 words only requires approximately a hundred rules, and identification of the spectra by the neural network is virtually instantaneous. Of course, this application remains very limited, but it enables us to envisage the use of such techniques for continuous speech as well as extensive vocabularies.

Bibliography


[Fogelman 86] Fogelman Soulie F., Le Cun Y. Modèle connexionniste de l'apprentissage. Intellectica 86 n° spécial "Apprentissage et machine".


