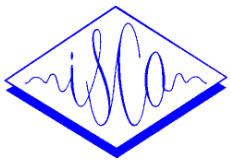

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

KNOWLEDGE-BASED SEGMENTATION AND FEATURE MAPS FOR SPEECH RECOGNITION

Franck POIRIER

Télécom Paris
46 rue Barrault
75013 Paris Cedex 13 France

ABSTRACT

A system for multi-speaker letter recognition in the French alphabet is presented. The main characterictic of this system is to combine a knowledge-based segmentation with a connexionist classifier. The segmentation module detects, automatically for each letter, discriminant acoustic patterns. For the classification module, two different approachs are tested. One is unsupervised using Self-organizing Feature Maps ; while the other requires supervised training data using Linear Vector Quantization classifier. A problem called bord effect appears with the first classifier and mainly explains the recognition rate. Best results are obtained with a new version of the second classifier which improves the recognition rate of 10%.

1-INTRODUCTION

Acoustic-phonetic decoding is mainly a pattern matching problem. Many methods have been used such as Dynamic Time Warping, Vector Quantization, Hidden Markov Modeling and more recently Neural Net classifiers.

It is well known that the speech waveform has a great variability and is very redundant. The intra-speaker and inter-speaker variabilities of a particular phoneme in a particular context are very important (speech features, duration). On the other hand, Neural Networks are very efficient for noisy or variable pattern classification, and generalization. However, they are especially fitted to process static patterns with fixed length and not too much quantity of information. So, Neural Networks are not adapted to process directly the speech waveform without earlier acoustic event extraction [1].

The speech database contains multi-speaker isolated pronunciations of the French alphabet. The recognition task of spoken letters by several speakers is difficult. The main reasons are the acoustic closeness of some letters (A and K, J and G, L and R, M and N, P and T ...), the short duration of these words, the variety of phonemes (25 phonemes on 33 in French), the lack of syntactic information. For such a specific vocabulary, it is better to use acoustic and phonetic knowledge for the segmentation process. For Neural Network, the segmentation is done in order to reduce the amount of data and to retain discriminant input.

2- SPEECH DATABASE AND FEATURE EXTRACTION

The corpus is composed of the French alphabet plus the phoneme /e/. The speech database is a subset of BDSONS with 28 speakers (14 male and 14 female) who pronounced 4 times the 27 words of the corpus. The speech database contains 3024 letters, 18 speakers (9 male and 9 female) are used for training (1944 letters) and 10 other speakers (5 male and female) are used for the test phase (1080 letters).

The corpus does not contain the nasal vowels but contains all the consonants execpt the phoneme /ɲ/.

The speech waveform is sampled at 8 kHz on 16 bits. Every 10ms, using 20ms overlapping Hamming window, an 8-dimension MFCC vector is computed.

3- THE SEGMENTATION

The segmentation is based on the typology of the 27 letters shown in Fig. 1. The segmentation uses phonetic knowledge in order to describe each letter by only three discrinant events such as beginning or end of vowel, frication before or after the vowel, plosion, voiced occlusion, vowel nucleus. Each event is associated with one frame of 20ms.

Two complementary parameters, the global energy (E) and the zero crossing density (Z) are used to localize these events. On both parameters, characteristic patterns (primitives) are searched such as ascendant, descendant or monotonous segments, peaks or valleys. E and Z are smoothed in order to suppress their local unstability and to enhance their structure. Rules on these primitives allow to describe each letter by three events (three frames).

For example, the letter 'C' is described in the following form :
frication + vowel nucleus + transient vocalic
- the letter 'B' is described by :
voiced occlusion + vowel nucleus + transient vocalic
- the letter 'S' is described by :
beginning of vowel + vowel nucleus + frication

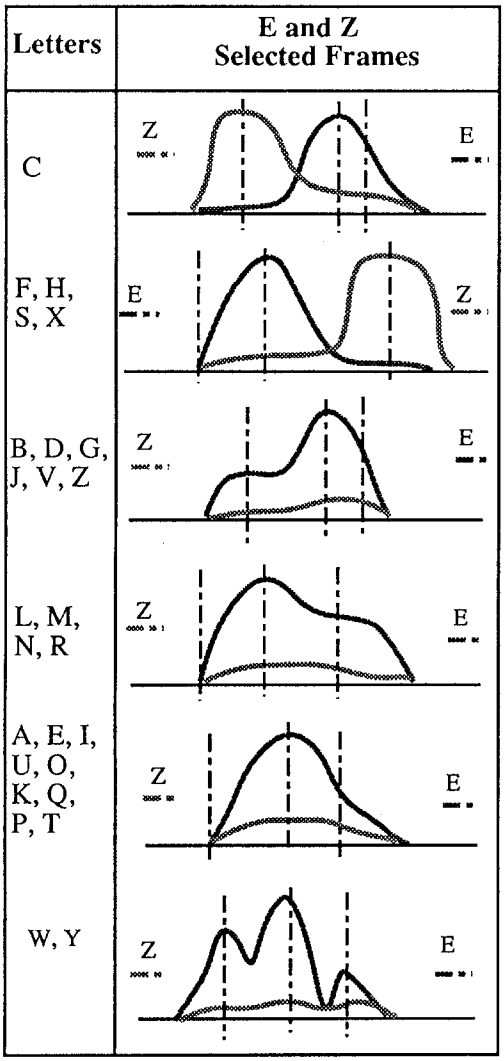

Fig. 1. Typology of the Letters and Selected Frames

4- NEURAL NETWORK ARCHITECTURE

4-1 Self-Organisation and unsupervised learning

Self-Organising Feature Map (SOFM) was originally proposed by Kohonen [2] in order to represent hight dimensional data in a topological map.

SOFM is an exemplar classifier that performs classification based on the identity of training examples and shows a great capability to organize input features. Each node or unit of the SOFM receives the same input. SOFM trains rapidly and requires less memory than k-Nearest Neighbors classifier.

The topology is defined by means of a neighbourhood on each unit that decreases over time to ensure asymptotic stability and by lateral interaction (Mexican hat function). The closest unit to the input and all the units that belong to its neighbourhood are moving in the direction of the input frame.

In fact, the units may be regarded as the image of inputs. The units are some kind of vector quantization of inputs.The map forms the projection image of the higher-dimensional speech frame distribution.

The structure of the SOFM is done by the hyper-cube $[1, 2,..., r]^p$ of r^p units. Each unit is characterized by its internal state, a weight vector (or reference vector) of the same dimension as the input vector.

Let $v_{i1i2...ip}$ be r^p reference vectors, where $i_k \in [1...r]$. At each step, for an input vector x, all the units are adapted, the adaptation rule is the following :

$$v_{i1i2...ip}(t+1) = v_{i1i2...ip}(t) + \alpha . (x - v_{i1i2...ip}(t)) \quad (1)$$

where

- $\alpha = \phi(\rho) . \varepsilon(t)$
- ρ is the distance between $v_{i1i2...ip}$ and v (the closest unit to x), $\phi(\rho)$ decreases over the distance ρ
- ε is the adaptation gain, $\varepsilon(t)$ decreases over time

So, every unit becomes sensitized to a particular input vector. The final result of the adaptation is that the distribution of the weight vectors tends to approximate the distribution of the input vectors.

4-2 Learning Vector Quantization (LVQ)

LVQ is similar in structure to the feature map classifier but LVQ is a supervised Nearest Neighbour classifier. So, training data must be labelelled. This vector quantization method gives a codebook of k reference vectors without any topology.

LVQ adjusts slighty the reference vectors in a direction towards the input vector x that attempts to improve performance.

Two versions of LVQ (LVQ1 and LVQ2) are pictured in Fig. 2.

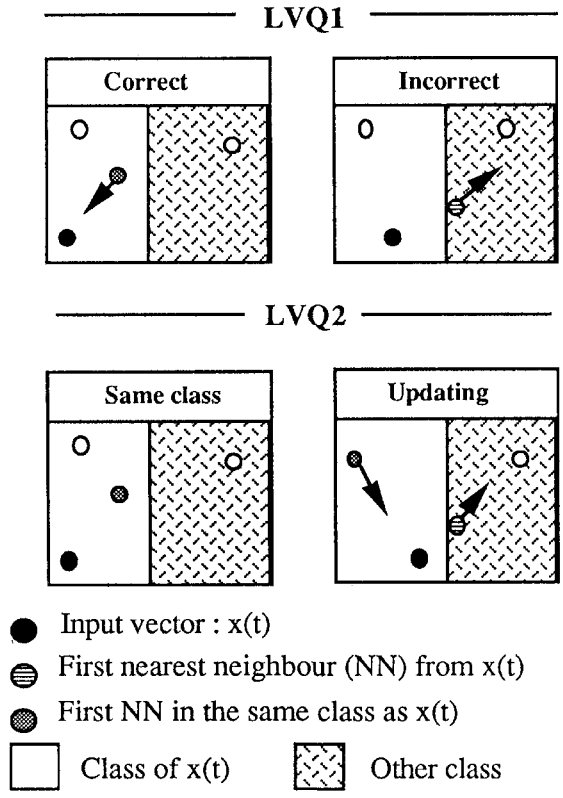

Fig. 2. LVQ1 & LVQ2

5- EXPERIMENTS AND RECOGNITION RESULTS

5-1 Results using SOFM

The three discriminant frames (8-dimension MFCC vector) are the inputs of 3 SOFM. Each map is organised as a two dimensional lattice of 16*16 units. At each step all nodes are adapted by a center adaptation rule [3] that is a function of both time and distance from the node that is closest to the input. After the adaptation phase (1), the calibration phase of the SOFM consists of the estimation of a letter probability vector for each unit.

The system is shown in Fig. 3. A stationary map is used to decode the vowel segment (2^{nd} frame) and two transient maps are used to decode the other segments (1^{st} and 3^{nd} frames).

A rule decision system transforms the outputs of the 3 maps into an unique letter label. These rules take into account 3 types of information :

- phonetic confusions (for example, confusion between 'b' and 'd'),
- letter probability vector computed on the training set,
- letter discriminant weight of each frame (for example, the stationary map is not discrimant between the both letter /k/ and /a/).

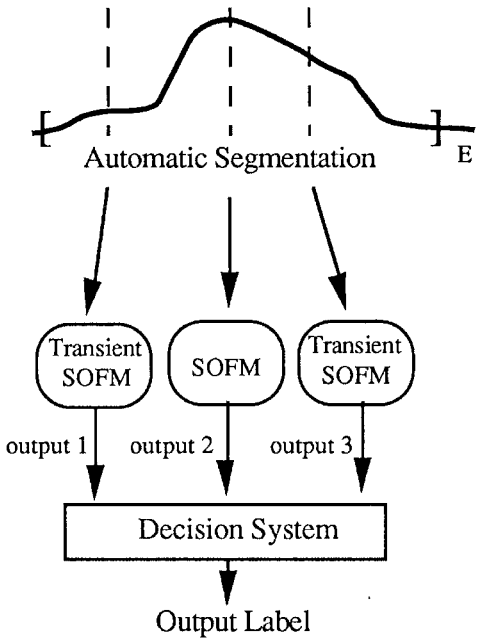

Fig. 3. System Architecture

The network is trained by 20000 iterations. Fig. 5 shown the recognition rate for male and female speakers on the training and test sets for all the experiments. On the training set (972 letters) the result is 77% and on the test set (540 letters) the result is 66% for male speakers and 70% for female.

The initialization of the map does not have any effect on the final result because the map is very mobile at the beginning . The representation of the different classes (letters) is quite continuous but with an important border effect. So, the most distant classes from the

center of the set are badly represented. It is the reason why LVQ is tested for letter recognition.

5-2 Results using LVQ

The first version of the LVQ gives bad asymptotic results (of the order of 60%). So, another version called LVQ2 [4] is used. The recognition rate is better than the SOFM one. On the training set, the recognition rate is 98% but it decreases, on the test set, at 71% for male speakers and 67% for female. LVQ2 creates too precise frontiers which induces bad generalization. In order to decrease the precision of the frontiers, the training set must be smoothed (some letters are removed from the set). In this case, the recognition rate is 72% on the test set for both male and female speakers.

In order to improve the ability of generalization without smoothing procedure, a new version of the algorithm called LVQ3 [5] is proposed. For LVQ2, the number of reference vectors is constant. Number and position is decided at the initialization. In a way, the frontier complexity is fixed and can not depend on the intrinsic complexity of the training set and more precisely on the complexity of each class. During the initialization of LVQ3, only one reference vector per class is chosen. During the adaptation phase, if the closest reference from the input x does not belong to the same class than x and if the closest reference in the class of x is too far from x then a new reference is created. This rule is shown in Fig. 4. Moreover LVQ3 trains faster than the other versions of LVQ.

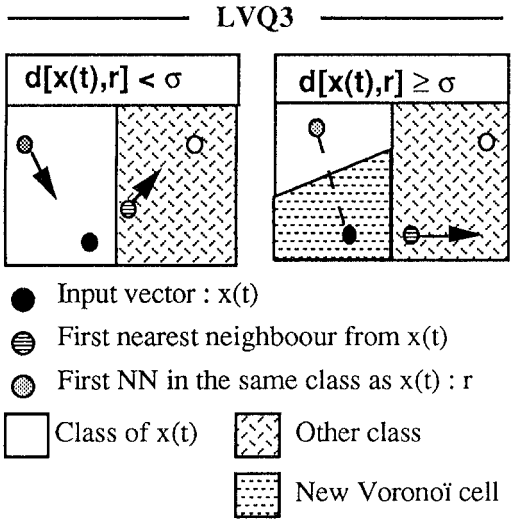

Fig. 4. LVQ3 Principle

For LVQ3, the smoothing procedure has no effect and for male speakers the recognition rate is 92% on the training set and 77% on the test set.

6- CONCLUSION

Knowledge-based segmentation is well fitted for specialized corpus like letters and allows to classify their by connexionist methods. SOFM does not need complete labelled data set and keep the topologic relations on the inputs. LVQ is a supervised vector quantizer. LVQ3 is well fitted to complex data like speech data classification (various density functions). The results are promising for multi-speaker experiments, a recognition score of 77% is achieved on unknown male speakers. Further researches will be focused on the border effect reduction for SOFM and the adaptation of such methods for continuous speech.

Method	Speakers	Smooth	Learning Data	Test Data
SOFM	female	no	77	70
	male	no	77	67
LVQ2	female	no	98	67
	male	no	98	71
	female	yes	95	72
	male	yes	91	72
LVQ3	female	no	93	74
	male	no	92	77

Fig.5. Recognition rates

References :

[1] F. Poirier, "Reconnaissance multi-locuteur de voyelles par un réseau connexionniste auto-organisateur". 18ème JEP, Montréal, 1990.

[2] T. Kohonen, "Self-Organization and Associative Memory". Springer Verlag, Series in Information Sciences, 1988.

[3] P. Brauer, P. Knagenhjelm, "Infrastructure in Kohonen maps". IEEE, S12.13,1989.

[4] T. Kohonen & al, "Statistical Pattern Recognition with Neural Networks : Benchmarking Studies". IEEE Proc. of ICNN, Vol. 1, july 1988.

[5] F. Poirier, "Réseaux auto-organisateur, LVQ2 et nouvelle version LVQ3 pour la classification de segments de parole". ENST Technical Report, 1990.