KEYWORD SPOTTING USING SUPERVISED/UNSUPERVISED COMPETITIVE LEARNING

Chakib Tadj and Franck Poirier

Télécom Paris - Signal Départment
46 Rue Barrault - 75634 Paris Cedex 13, FRANCE
tadj@sig.enst.fr

ABSTRACT

In this paper, we present a novel hybrid keyword spotting system that combines supervised and unsupervised competitive learning algorithms. The first stage is a *SOFM* (Self-Organizing Feature Maps) module which is specifically designed for discriminating between keywords (*KWs*) and non-keywords (*NKWs*). The second stage is a *FDVQ* (Fuzzy Dynamic Vector Quantization) module which consists of discriminating between *KWs* detected by the first stage processing. The results show an improvement of about 9% on the accuracy of the system comparing to our standard one.

Key Words: Word Spotting, Fuzzy Supervised Competitive Learning, Incremental Learning, Non-Linear Adaptive Learning Rules.

1. INTRODUCTION

Word spotting systems for continuous, speaker independent speech recognition are becoming more and more popular because of the many advantages they afford over more conventional large scale speech recognition systems. Several systems with different architectures have been proposed, most of them being based on the statistical Hidden Markov Model (*HMM*) [10, 9]. Several hybrid models were proposed to improve these systems [6, 1, 16].

Recent works have demonstrated the power of the competitive learning algorithms such as Learning Vector Quantization (*LVQ*) [4] and the Dynamic Vector Quantization (*DVQ*), an incremental version of the *LVQ* [7]. In our research, we have first proposed a new adaptive learning rules based on a spatial geometry considerations [12]. More recently, we have introduced the *FDVQ*, a supervised competitive learning algorithm based on fuzzy knowledge [13]. The adaptive learning rule uses a membership function defined on the nearest neighbors. For each input vector, the membership values are computed to adapt, create or annihilate units of the network.

2. MOTIVATIONS

In an application such as telephony-based automatic speech recognition, the recognizer must be able to word-spot valid utterances and reject non-valid ones. This means that word-spotting and rejection are related in that good word-spotting capability necessarily implies good rejection performance.

Several methods for non-keywords rejection have been proposed in the context of word-spotting for conversational speech monitoring [11, 2]. As the *FDVQ* was not designed to represent the acoustic garbage (or filler) models, our standard *FDVQ* based keyword spotter system [15] was based on some threshold considerations to reject the *NKWs* and garbage models. This implies that the discrimination capability of the algorithm must be very high according to the complexity of the problem. This conduct us to introduce on upstream a *SOFM* module which is specifically designed for this task. A particular advantage of using *SOFM* representation of acoustic garbage models, allows acoustical garbage models to assimilate information over many different speaker and word contexts. A strong collaboration between these two modules is done to improve the performance on both garbage rejection and keyword accuracy.

3. SYSTEM ARCHITECTURE

In our system, the architecture proposed is based on a two stage process: In the first one, the *SOFM* stage is specifically designed for discrimination between keywords and non-keywords. The non-keyword models can be entire words or smaller units. In the second processing, the system makes use of the strong generalization ability strength of the *FDVQ*. In its original formulation, the *FDVQ* discriminative training framework was developed to minimize the recognition errors by adaptation of the units. In this work, it consists of discriminating between *KWs* detected by the first stage processing. A paradigm of the basic network architecture is shown in figure 1. An *HMM* module is introduced on upstream to realize the speech signal detection, where each model is a left-to-right continuous density *HMM*. The approach described in this paper uses a whole-word keyword spotter, where the keywords occurrences in the training data are used to construct an overall models for the words. Knowledge of the phonetic structure of the keyword is not needed.

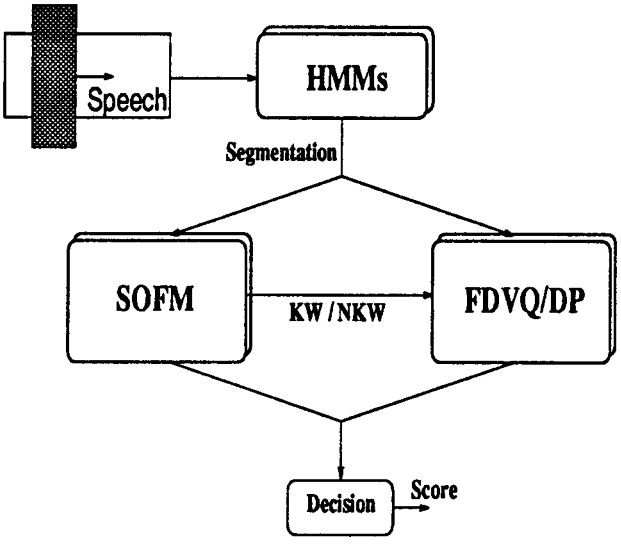

Figure 1: *System Architecture.*

3.1. SOFM Module

Unlike the other vector quantizer, *SOFM* organizes the codebooks in a topological structure called map. This allows to approximate the distribution of the training vectors in an orderly fashion.

The process in which the *SOFM* is formed is an unsupervised learning process. It is used to find clusters in the input data, and to identify an unknown data vector within one of the clusters.

Concretely, the training process is performed in order to discriminate between regions from the map space and to separate *NKWs* and garbage models from the *KWs*.

3.2. Review of the FDVQ

The principle of the *FDVQ* adaptive learning rule is performed according to some spatial considerations to optimize the references adaptation as described in figure 2. The advantage of this adaptation rule is to preserve the inter-class distance between the references before adaptation (d_i) and after (d_f) and to reorganize in an optimal way their distribution according to the example x presented to the network [13]. The corresponding adaptive learning rules are:

$$\begin{aligned} \overrightarrow{m_k m'_k} &= sgn * \delta * \overrightarrow{m_k x}, \\ \overrightarrow{m'_k m''_k} &= \alpha_k * \vec{U}_k, \qquad k = i, j \end{aligned} \tag{1}$$

Figure 3 gives a geometrical description of the unit vector $\vec{U}_j$. The vectors $\vec{U}_{j'j''}$ and $\vec{U}_{i'j'}$ are the director vectors of $\overrightarrow{m'_j x}$ and $\overrightarrow{m'_i m'_j}$ respectively.

The learning rate α_k is determined to preserve the inter-class distances between references. m'_k is the reference after the first adaptation and m''_k, the reference after the second adaptation. $\vec{U}_k$, $k = i, j$, is the medium hyperplane of the normalized hyperplanes $\vec{U}_{kk'}$ *and* $\vec{U}_{k'l'}$, defined by :

$$\vec{U}_{kk'} = sgn * \frac{\overrightarrow{m'_k x}}{\| \overrightarrow{m'_k x} \|}, \quad k = i, j \tag{2}$$

$$\vec{U}_{k'l'} = sgn * \frac{\overrightarrow{m'_k m'_l}}{\| \overrightarrow{m'_k m'_l} \|}, \quad k, l, = i, j, \quad k \neq l \tag{3}$$

3.3. Dynamic Programming module

To solve the problem relative to the time variability of the speech units, the system integrate for each stage processing a Dynamic Programming (*DP*) module which perform a **non-linear adaptive learning rules** [14]. *DP* models are used by the system because they

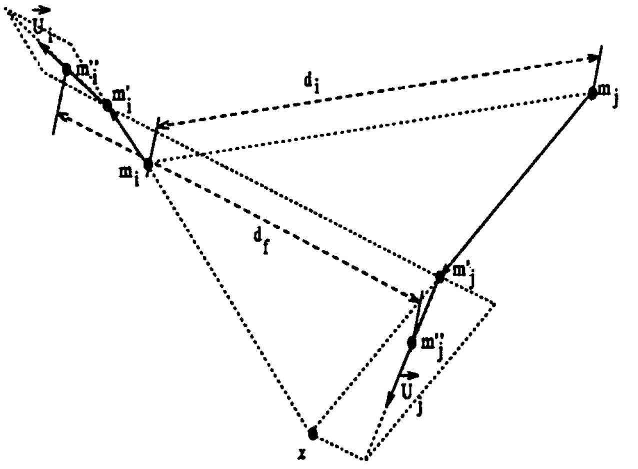

Figure 2: *Descriptive representation of the* FDVQ. *Orientation of m''_i and m''_j are given by the hyperplanes defined by ($\overrightarrow{m_i m'_i}$, $\overrightarrow{m'_j m'_i}$) and ($\overrightarrow{m'_i m'_j}$, $\overrightarrow{m_j m'_j}$) respectively. m'_i and m'_j are the references after the first adaptation. m''_i and m''_j the references after the second adaptation.*

have the useful capability of time warping input speech patterns. This allows the system to map the different variable length occurrences of the same keyword to a single output unit.

4. EXPERIMENT

We have trained and tested our system on the extensive Switchboard corpus of telephone conversations. The system proposed is compared to our standard *FDVQ* based keyword spotter system and to the *MSTDNN* one [16].

4.1. Switchboard Database

The Switchboard speech corpus consists of informal conversations between speakers on their use of credit cards. The word spotting task is to detect a vocabulary of 20 keywords and their variants from the utterances corresponding to the individual speakers in the stored conversations [3].

4.2. Performance Measure

To evaluate the performance of our system, we used the receiver operating characteristic (*ROC*) curve which allows a trade-off between detection probability and false-alarm rates [8]. We are interested in the false-alarm rates below 10 false alarms per-keyword per hour ($fa/(kw * hr)$); therefore, we characterize the result of an experiment by the average value of the *ROC* curve over the range of 0 to 10 $fa/(kw * hr)$.

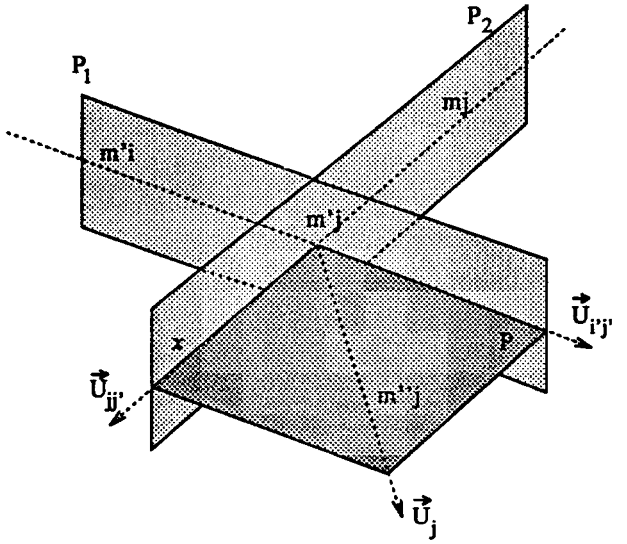

Figure 3: *Descriptive Representation of $\vec{U}_j$.*

4.3. Results

The figure 4 shows a comparison between the proposed architecture with both our standard *FDVQ* keyword spotting and the *MS-TDNN*.

The results show an improvement of about 9% on the accuracy of the system comparing to our standard system. The performances obtained show a high efficiency in both garbage rejection and keyword accuracy.

5. CONCLUSION

In this paper, we present a novel hybrid keyword spotting system that combines supervised and unsupervised competitive learning algorithms. The results show an improvement of this architecture for both learning and testing data.

As the SOFM module is very important in the sense that it permits to discriminate between *KWs* and *NKWs*, we are more investigated in implementing a Self-Organizing Map with supervision which will cooperate with the *FDVQ* module to reduce the false alarm rates.

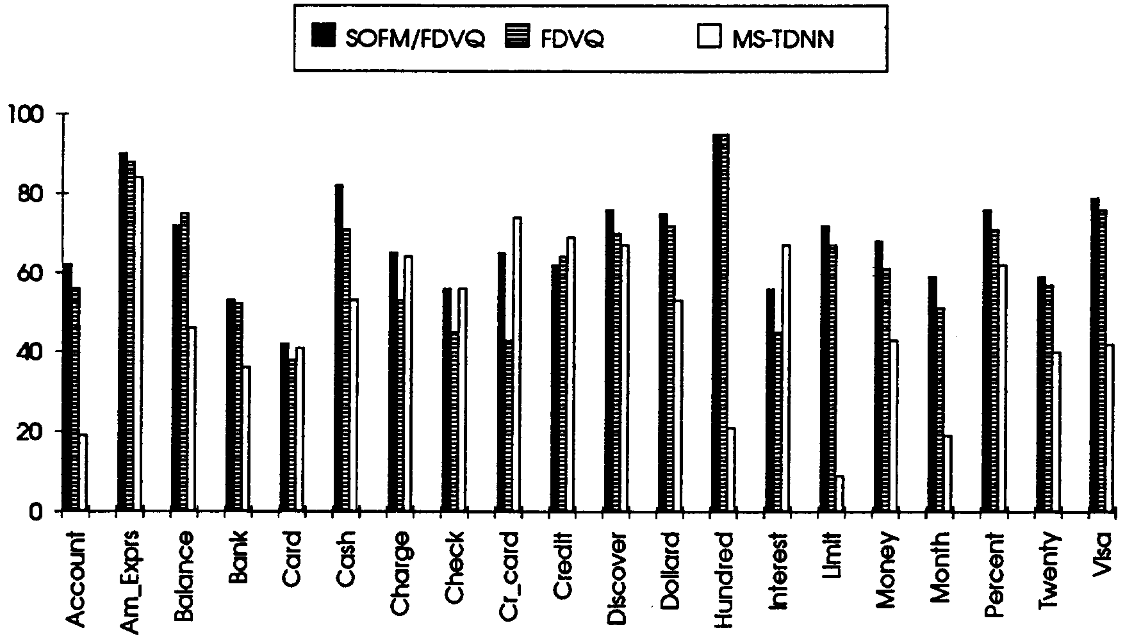

Figure 4:*Comparison between SOFM/FDVQ, FDVQ and the MS-TDNN*

6. REFERENCES

[1] J. Alvarez-Cercadillo and A. Hernandez-Gomez, "Grammar Learning and Word-Spotting Using Recurrent Neural Networks", *EuroSpeech93*, Vol. 3, pp. 1277-1280, Sept., 1993.

[2] S. Austin, G. Zavaliagkos, J. Makhoul and R. Schwartz, "Speech Recognition Using Neural Nets", *ICASSP92*, Vol. 1, pp. 625-628, 1992.

[3] J. Godfrey, E. C. Holliman and L. McDaniel, "Switchboard: Telephone Speech Corpus for Research and Development", *ICASSP92*, Vol. 1, pp. 517-520, 1992.

[4] T. Kohonen, "The Self-Organizing Map", *Proc. of IEEE*, Vol. 78, No. 9, Sept., 1990.

[5] T. Komori and S. Katagiri, "An Optimal Learning Method for Minimizing Spotting Errors", *ICASSP 93*, Volume 2, pp. 271-274, April 27-30, 1993.

[6] D. P. Morgan, C. L. Scofield, T. M. Lorenzo, E. C. Real and D. P. Loconto, "A Keyword-Spotter Which Incorporates Neural Networks for Secondary Processing", *ICASSP90*, Vol. 1, pp. 113-116, April, 1990.

[7] F. Poirier, A. Ferrieux, "*DVQ* : Dynamic Vector Quantization - An Incremental *LVQ*",*ICANN91*, Vol. 2, pp. 1333-1336, June, 1991.

[8] J. R. Rohlicek, W. Russel, S. Roucos and H. Gish, "Continuous HMM for Speaker Independent Word Spotting", *ICASSP89*, Volume 1, pp. 627-630, May 23-26, 1989.

[9] J. R. Rohlicek, P. Jeanrenaud, K. Ng, H. Gish, B. Musicus and M. Siu "Phonetic Training and Language Modeling for Word-Spotting", *ICASSP93*, Vol. 2, pp. 459-462, April, 1993.

[10] R. C. Rose and D. B. Paul, "A Hidden Markov Model Based Keyword Recognition System", *ICASSP90*, Vol. 1, pp. 129-132, April, 1990.

[11] R. C. Rose, "Discriminant Word-Spotting Techniques for Rejecting non Vocabulary Utterances in Unconstrained Speech", *ICASSP92*, Vol. 2, pp. 105-108, Mar., 1992.

[12] C. Tadj and F. Poirier, "Improved *DVQ* Algorithm for Speech Recognition: A New Adaptive Learning Rule With Neurons Annihilation", *EuroSpeech93*, Vol. 2, pp. 1009-1012, Sept., 1993.

[13] C. Tadj and F. Poirier, "On a Fuzzy *DVQ* Algorithm for Speech Recognition", *NATO-ASI93*, pp. 215-219, July 93.

[14] C. Tadj and F. Poirier, "Fuzzy DVQ Algorithm Integrating Time Alignment for Continuous Speech Recognition", *Progress and Prospects of Speech Research and Technology - CRIM/FORWISS94 Workshop*, to appear.

[15] C. Tadj and F. Poirier, "A Two Pass Classifier for Utterance Rejection in Word-Spotting", *ICEEE94 Workshop*, to appear.

[16] T. Zeppenfeld, R. Houghton and A. Waibel, "Improving the MS-TDNN for Word-Spotting", *ICASSP93*, Vol. 2, pp. 475-478, April, 1993.