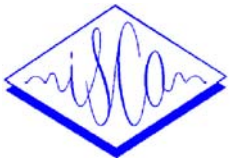

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

# FDVQ Based Keyword Spotter Which Incorporates A Semi-Supervised Learning for Primary Processing

*Chakib Tadj* [†], *Pierre Dumouchel* [†‡]

[†]École de Technologie Supérieure
1100 rue Notre Dame Ouest
Montréal (Qc) - H3C 1K3 - Canada

[‡]Centre de Recherche Informatique de Montréal
1801, avenue McGill College, bureau 800
Montréal (Qc) - H3A 2N4 - Canada

*Franck Poirier* [◇]

[◇]Institut Universitaire Professionnalisé
8, rue Montaigne BP 1104
Vannes - 56014 - France

## ABSTRACT

In this paper, we present a novel hybrid keyword spotting system that combines supervised and semi-supervised competitive learning algorithms. The first stage is a S-SOM (Semi-supervised Self-Organizing Map) module which is specifically designed for discrimination between keywords (KWs) and non-keywords (NKWs). The second stage is an FDVQ (Fuzzy Dynamic Vector Quantization) module which consists of discriminating between KWs detected by the first stage processing. The experiment on Switchboard database has show an improvement of about 6% on the accuracy of the system comparing to our best keyword-spotter 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 Models (HMM) [9, 10]. Several hybrid models were proposed to improve these systems [1, 6, 16].

An other research area has demonstrated the power of the competitive learning in pattern recognition and speech recognition. Such algorithms are well known as Learning Vector Quantization (LVQ) [5], Dynamic Vector Quantization [8], and the Fuzzy Learning Vector Quantization (FLVQ) [3]. In our research, we have first proposed a new adaptive learning rules based on spatial geometry considerations [12]. The adaptive learning rules use a membership function defined on the nearest neighbors [13]. 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 "wordspot" 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 [2, 11]. As the FDVQ was not designed to represent the acoustic garbage (or filler) models, our standard FDVQ based keyword-spotting system [14] 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 SOM module which is specifically designed for this task [15]. A particular advantage of using SOM 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. To use the power of the supervised learning, we are interested in this paper in adding supervision to the SOM in order to improve the discrimination between KWs and NKWs.

## 3. SYSTEM ARCHITECTURE

In our system, the architecture proposed is based on a two stage process. In the first one, the SOM stage with supervision 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. The approach described in this paper uses an whole-word keyword-spotter, where the keywords occurrences in the training data are used to construct an overall models. Knowledge of the phonetic structure of the keyword is not needed.

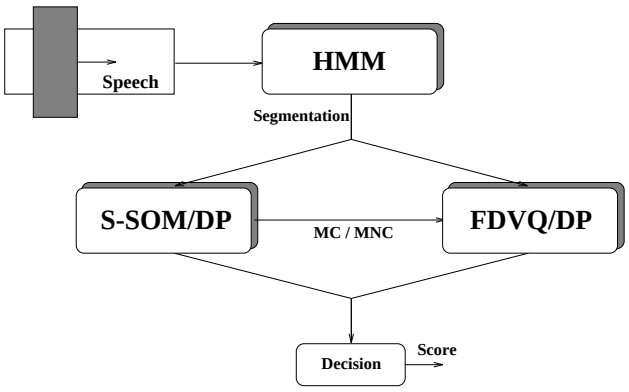


Figure 1: System Architecture

The process in which the S-SOM is formed is a semi-supervised learning process. It is used to find clusters in the input data, and to identify an unknown data vector within one of the clusters. This process combines the advantage of a topological representation in the map space and the discriminating power of supervised learning [7]. Concretely, the semi-supervised learning phase consists of two different periods: (a) adaptation of the weight vectors by the standard SOM learning rule, (b) adaptation of the weight vectors by the LVQ rule. The first period correspond to the initial formation of the map. The second one corresponds to the final convergence of the map. The training process is formed in order to discriminate between regions from the map space and to separate NKWs and garbage models from the KWs. 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 and after adaptation and to reorganize their distribution in an optimal way, according to the example presented to the network. To solve the problem relative to the time variability of the speech units, the system integrates for each stage processing a Dynamic Programming (DP) module which performs a non-linear adaptive learning rules. DP models are used by the system because they 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.

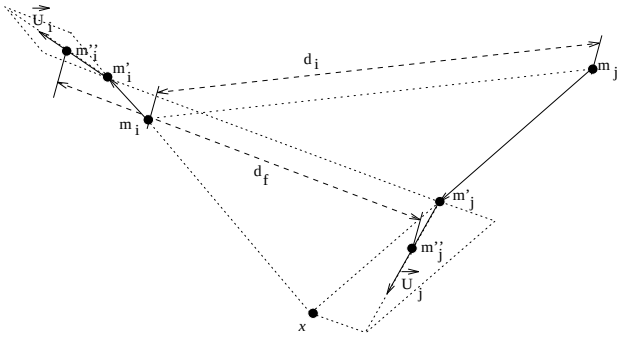


Figure 2: *Descriptive representation of the* FDVQ. *Orientation of* $m_i''$ *and* $m_j''$ *are given by the hyperplanes defined by* ($\overrightarrow{m_i m_i'}$, $\overrightarrow{m_j' m_i'}$) *and* ($\overrightarrow{m_i' m_j'}$, $\overrightarrow{m_j m_j'}$) *respectively.* $m_i'$ *and* $m_j'$ *are the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

## 4. EXPERIMENT RESULTS

### 4.1. Switchboard Database

The proposed approach is evaluated using the conversational speech. The most important characteristics of this kind of speech signal are: (a) fundamental differences in speaking style between read and conversational speech and (b) non-grammatical speech events.

The corpus contains training data for word spotting on the Switchboard credit card conversations. Thirty five conversations are included. The keyword spotting task is to detect a vocabulary of keywords and their variants from the utterances corresponding to the individual speakers in the stored conversations [4]. From this corpus, 20 keyword and their variants were chosen to be spotted (e.g. bank banks bankruptcy citibank bankrupt banked). Table 1 summarizes all the variants used in Switchboard database.

| Keyword | # Repetitions | Variants |
|---|---|---|
| *account* | 37 | accounts accounting |
| *amr_exprs* | 49 | — |
| *balance* | 41 | balances balancing |
| *bank* | 56 | banks bankruptcy citibank bankrupt banked |
| *card* | 622 | cards mastercard mastercards card's americard mastercard's |
| *cash* | 100 | cashing cashed |
| *charge* | 125 | charged charges surcharge |
| *check* | 114 | checks checking checkbook |
| | 114 | paycheck checked |
| *credit* | 455 | credited |
| *credit card* | 358 | credit cards |
| *discover* | 29 | discovered discovers |
| *dollar* | 96 | dollars |
| *hundred* | 40 | — |
| *interest* | 104 | interesting interested |
| *limit* | 32 | limits limited limiting |
| *money* | 112 | money's |
| *month* | 122 | months monthly month's |
| *percent* | 54 | percentage |
| *twenty* | 17 | twenties |
| *visa* | 76 | visas visa's |

Table 1: *Keywords and their variants in Switchboard database*

The proposed task is to spot these occurrences while minimizing the number of false detections. We use 50% of the database for training and 50% for testing the generalization ability of the system.

### 4.2. Results

Table 2 shows the performance of the system proposed in section 3. The approach proposed has improved (a) detection of words poorly represented in the database as *twenty* and *hundred* and (b) discrimination between keywords which were frequently confused as in the case of *cash* and *card* and between *credit* and *card*. Figure 3 shows the comparison results, in the test mode, for the systems *S-SOM/FDVQ*, *SOM/FDVQ*, *FDVQ/DP*, the *HMM* based word spotter and the *MS-TDNN* one, proposed in [16].

The results show an improvement of about 15% on the accuracy of the system comparing to our standard system [14] and about 6% on our best keyword-spotter one [15]. The performances obtained show a high efficiency in both garbage rejection and keyword accuracy comparing to both MS-TDNN and HMM systems.

| Keyword | Recognition (%) |
|---|---|
| *Account* | 78 |
| *American_express* | 92 |
| *Balance* | 79 |
| *Bank* | 62 |
| *Card* | 58 |
| *Cash* | 89 |
| *Charge* | 72 |
| *Check* | 62 |
| *Credit_Card* | 79 |
| *Credit* | 76 |
| *Discover* | 85 |
| *Dollar* | 79 |
| *Hundred* | 97 |
| *Interest* | 78 |
| *Limit* | 89 |
| *Money* | 71 |
| *Month* | 65 |
| *Percent* | 86 |
| *Twenty* | 69 |
| *Visa* | 82 |

Table 2: *Performance of the proposed system S-SOM/FDVQ. Generalisation test*

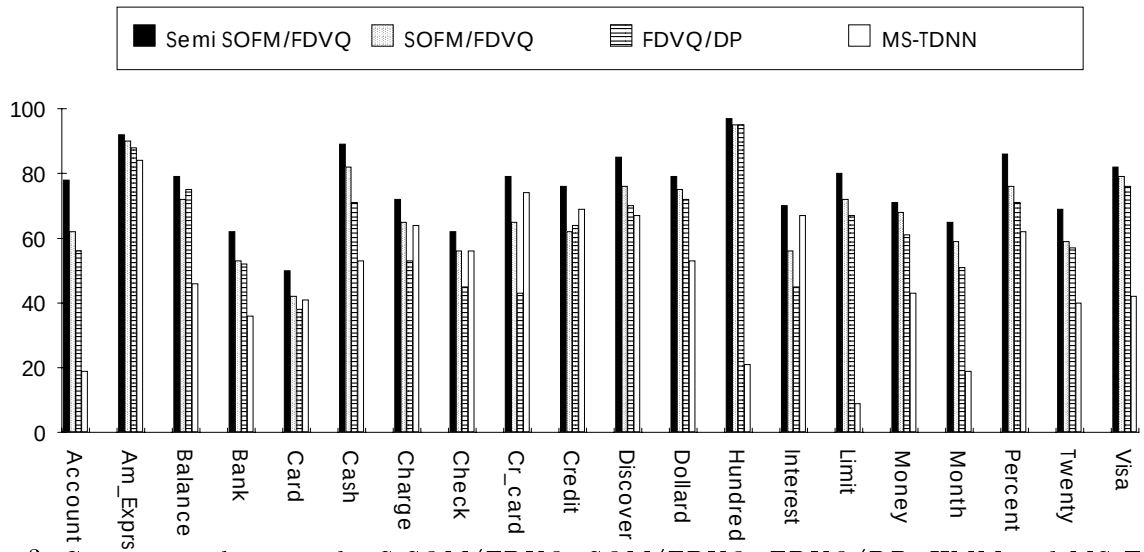


Figure 3: *Comparison between the S-SOM/FDVQ, SOM/FDVQ, FDVQ/DP, HMM and MS-TDNN. Generalisation test*

## 5. CONCLUSIONS

In this paper, a modular architecture based on vector quantization is presented, to perform the keyword spotting task. Integrating semi-supervised and supervised learning modules as well as a dynamic programming one has shown a good performance and an improvement of 6% comparing to our best keyword-spotter. An interesting and suggested future work is the integration of an other module to replace the S-SOM one. A kind of dynamic module which can be able to add/annihilate models (or formal neurons) during the training phase. This module will allow the system to not fix the number of neurons a priori, and thus, leads to a better topological map.

## 6. REFERENCES


[1] J. Alvarez-Cercadillo and A. H.-Gomez, "Grammar Learning and Word Spotting Using Recurrent Neural Networks", *EuroSpeech 93*, pp. 1277-1280, 1993.

[2] S. Austin and all, "Speech Recognition Using Neural Nets", *ICASSP 92*, Vol. 1, pp. 625-628, 1992.

[3] F. L. Chung and T. Lee, "Fuzzy Competitive Learning", *Neural Networks*, Vol. 7, No. 3, pp 539-551, 1994.

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

[5] T. Kohonen, "The Self Organizing Map", *IEEE*, Vol. 78, No. 9, 1990.

[6] E. McDermott and S. Katagiri, "Prototype-Based MCE/GPD Training for Word Spotting and connected Word Recognition", *ICASSP 93*, Vol. 2, pp. 291-294, 1993.

[7] S. Midenet and A. Grumbach, "Learning Associations by Self-Organization: The LASSO Model", *Neurocomputing*, Vol. 6, pp. 343-361, 1994.

[8] F. Poirier, A. Ferrieux, "DVQ : Dynamic Vector Quantization - An Incremental LVQ", *ICANN-91*, Vol. 2, pp. 1333-1336, 1991.

[9] J. R. Rohlicek and all, "Phonetics Training and Language Modeling for Word Spotting", *ICASSP 93*, Vol. 2, pp. 459-462, 1993.

[10] R. C. Rose and D. B. Paul, "A Hidden Markov Model Based Keyword Recognition System", *ICASSP 90*, Vol. 1, pp. 129-132, 1990.

[11] R. C. Rose, "Discriminant Word Spotting Techniques for rejection non Vocabulary Utterances in Unconstrained Speech", *ICASSP 92*, Vol.2, pp.105-108,1992.

[12] C. Tadj and F. Poirier, "Improved DVQ Algorithm for Speech Recognition", *EuroSpeech 93*, Volume 2, pp. 1009-1012, 1993.

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

[14] C. Tadj and F. Poirier, "A Two Pass Classifier for Utterance Rejection in Word-Spotting", *ICEEE 94 workshop*.

[15] C. Tadj and F. Poirier, "Word Spotting Using Supervised/Unsupervised competitive Leaning", *ICASSP 95*, Vol. 1, pp. 301-304, 1995.

[16] T. Zeppenfeld, R. Houghton and A. Waibel, "Improving the MS-TDNN for Word Spotting", *ICASSP 93*, Volume 2, pp. 475-478, 1993.