International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:6, 2007

A New Vector Quantization front-end Process for Discrete HMM Speech Recognition System

M. Debyeche, J.P Haton, and A. Houacine

***Abstract*—**The paper presents a complete discrete statistical framework, based on a novel vector quantization (VQ) front-end process. This new VQ approach performs an optimal distribution of VQ codebook components on HMM states. This technique that we named the distributed vector quantization (DVQ) of hidden Markov models, succeeds in unifying acoustic micro-structure and phonetic macro-structure, when the estimation of HMM parameters is performed. The DVQ technique is implemented through two variants. The first variant uses the K-means algorithm (K-means-DVQ) to optimize the VQ, while the second variant exploits the benefits of the classification behavior of neural networks (NN-DVQ) for the same purpose. The proposed variants are compared with the HMM-based baseline system by experiments of specific Arabic consonants recognition. The results show that the distributed vector quantization technique increase the performance of the discrete HMM system.

***Keywords*—**Hidden Markov Model, Vector Quantization, Neural Network, Speech Recognition, Arabic Language

I. Introduction

AUTOMATIC Speech Recognition (ASR) can be viewed as a successive transformations of the acoustic micro-structure of the speech signal into its implicit phonetic macro-structure. The main objective of any ASR system is to realize the mapping between the two structures. The hidden Markov model (HMM) is actually the most used approach to the ASR. Several types of HMMs as discrete, continuous and semi continuous HMMs [1], [2] have been developed and applied to the ASR. The discrete HMM (DHMM) is attractive in terms of algorithmic complexity; that is why, it has been investigated in several studies [3], [4], [5], [6]. Recently, in the context of the prodigious growth of network applications, discrete HMM-based speech recognition systems that use a Vector Quantization (QV) front-end process constitute a very useful and inexpensive solutions [7], [8].

Manuscript received October 5, 2005.

M. Debyeche, Laboratory of Signal Processing and Speech communication (LCPTS), faculty of Electronics and Computer Sciences, (USTHB), P.O. Box 32 El-Alia, Bab-Ezzouar, Algeirs, Algeria.(phone:+213 72 63 86 44; fax:++213 21 24 71 87; e-mail: mdebyeche@caramail.com and mdebyeche@usthb.dz).

J.P Haton, LORIA/INRIA-Lorraine, 615 rue du jardin botanique, P.O. Box 101, F-54600, Villers-lès-Nancy, France (e-mail: jph@loria.fr).

A. Houacine, Laboratory of Signal Processing and Speech communication (LCPTS), faculty of Electronics and Computer Sciences, (USTHB), P.O. Box 32 El-Alia, Bab-Ezzouar, Algeirs, Algeria.(e-mail: ahouacine@usthb.dz).

In this scenario, it is highly desirable to perform compression of acoustic features, but it is crucial that the VQ involved in the front-end stage does not introduce noise that degrades the recognition accuracy. This is the dilemma. In fact, discrete HMM inherently suffers from some problems due to the Vector Quantization (VQ) process. The lack of sufficient training data involved by the VQ causes poor HMM parameter estimation, and this inevitably leads to a degradation of recognition performance. This paper is dedicated for improving accuracy issues of discrete HMM-based ASR systems. It proposes a complete discrete statistical framework, based on the use of a novel VQ-based front-end process. This new approach performs an optimal distribution of VQ codebooks on HMM states. This technique, which has been named the distributed vector quantization (DVQ) of hidden Markov models, succeeds in unifying acoustic micro-structure and phonetic macro-structure, when the parameter estimation of HMM is performed. The DVQ technique is implemented through two variants. The first variant uses the K-means algorithm (K-means-DVQ) to optimize the VQ, while the second variant exploits the benefits of the classification behavior of neural networks (NN-DVQ) for the same purpose. The evaluation is done by focusing on specific Arabic consonants: emphatic and back consonants. The characterization of these consonants has captured the interest of many researchers, since they are specific to the Arabic language [9]. The paper is structured as follows: after the first, introductory section, we present in the second section the wellknown statistical paradigm used for speech recognition represented by the HMM. In section 3 we depicts the framework of distributed vector quantization. Section 4 reports the comparative results of trials that aim to evaluate the proposed techniques by focusing on some specific Arabic phonemes. Finally, we summarize our major findings in section 5.

II. VQ/HMM System for Word Recognition

To illustrate an application of HMMs for speech recognition, we present in Fig.1 our implementation of an isolated word recognition system based on discrete hidden Markov models. We have a vocabulary of L words to be recognized, and each word is to be modeled by a disctinct HMM. The training sets consist of K utterances of each word, pronounced by one or more speakers. In order to obtain a word recognizer, we performed the following steps:

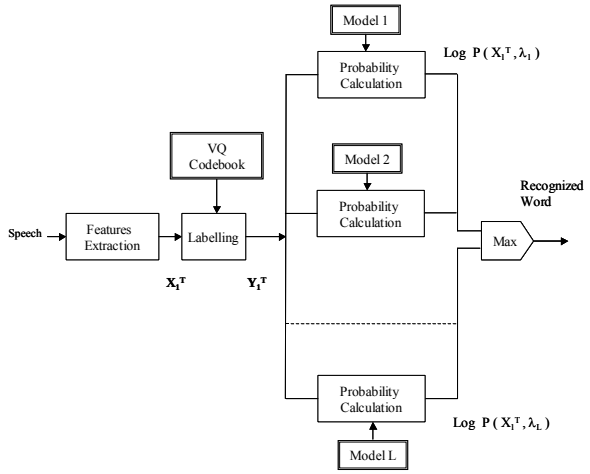

Fig. 1 Schematic diagram of the VQ/HMM isolated word recognition system

A. Features Extraction

The digitized speech signal is pre-emphasized by a first-order digital filter in order to spectrally flatten the signal $\hat{S} = S(n) - \mu S(n-1)$, with μ = 0.96. The signal is fragmented into frames by using a 25.6 ms Hamming window with 10 ms shifting. For each frame, the *mel frequency cepstral coefficients* (MFCCs) [10], their corresponding first and second derivatives, named respectively ΔMFCC, ΔΔMFCC and the *energy* E are computed. Each frame is thus represented by an acoustic vector x_t as follows:

$$x_t = \{MFCC(m), \Delta MFCC(m), \Delta\Delta MFCC(m), E\} \quad (1)$$

The first and second order derivatives of cepstral coefficients were approximated respectively by equations (1) and (2) given as follows:

$$\Delta MFCC_l(m) = \sum_{k=-K}^{K} k\,(MFCC_{l-k}(m)) \quad (2)$$

$$\Delta\Delta MFCC_l(m) = \Delta MFCC_{l+1}(m) - \Delta MFCC_{l-1}(m) \quad (3)$$

where *k* and *l* are frame indexes, and *m* the MFCC component.

B. VQ Codebook

In discrete HMM system, the continuous feature space is subdivided by a vector quantizer into J non-overlapping subsets and each subset is represented with a codeword m_j ($1 \le i \le J$). The set of available codewords is termed the codebook. The VQ codebook is constructed by an unsupervised cluster algorithm, the LBG (Linde–Buzo-Gray) algorithm [11].

C. Re-estimation of HMM

For each word of the vocabulary, we built a HMM, that is, we estimated the model parameters that optimize the likelihood for the training set of observation sequences. There are many criteria that can be used to this problem. We have used for this problem the Baum-Welch algorithm [12] developed by Baum which is one of the most successful optimization methods.

D. Recognition

For each unknown word to be recognized, we calculated the model likelihood for all possible models, and selected the model with the highest likelihood. The probability calculation was performed using the Viterbi algorithm [13], more precisely the logarithm of the maximum likelihood.
The system developed can be applied not only for word recognition, but also for recognition of other speech segments. Experiments performed with this system will be discussed in Section 5.

III. Distrubuted VQ/HMM System

The main weak point of VQ/HMM, in the field of ASR, resides in the fact that they inherently suffer from some problems linked to the quantization error induced by the limited number of clusters of input vectors, and the lack of sufficient training data that causes poor estimation of HMM parameters. In order to limit the effect of this insufficiency, we propose the use of a new technique (DVQ) based on the principle of optimally distributing the codebook components, issued from a vector quantization, over the HMM states. This approach will allow a model parameter initialization based on the expected unification of acoustic and phonetic sources. Two hybrid implementations of this approach are presented: the K-means DVQ and the neural network NN-DVQ. The synoptic of a DVQ-based system dedicated to isolated word or phoneme recognition, is given in Fig. 2.

A. Overview of the DVQ approach

For recognition systems that use HMMs, it is important to be able to estimate probability distributions of the computed feature vectors preferably over a high multi-dimensional space. To reach this goal, it is often easier to start by quantizing each feature vector to one of a relatively small number of template vectors. In fact, DVQ aims to make probability distributions estimation over this finite set of templates more effective by tying this set of templates to a corresponding HMM state. The training phase of DVQHMM is illustrated in Fig. 3.

The observation sequence labelling, and evaluation, are done simultaneously. For each sequence X_1^T, we have the sequence $\{Y_1^{T(j)},\ 1 \le t \le T \text{ and } 1 \le j \le N\}$. To compute the probability $P(X_1^T,\ \lambda)$, probability of generating the sequence by the

model, we use the modified logarithm of the maximum likelihood :

$$\alpha(t,j)=\max_j\left[\alpha(t-1,j)+Loga_{ji}\right]+Logb_i\ (Y_t^{(i)}) \qquad (4)$$

with $1 \le t \le T$ and $1 \le i, j \le N$

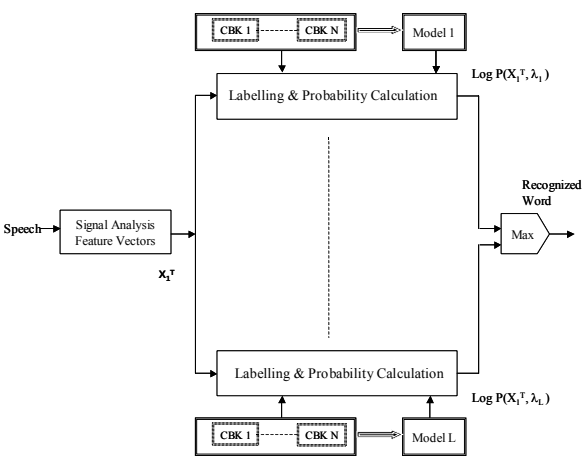

Fig. 2 Overview of DVQ approach applied to a word recognition process

Two hybrid techniques are used to optimally distribute code vectors over the HMM states: K-means- and neural networks-based techniques.

B. The Hybrid K-means DVQ

In the K-means DVQ variant, we use the K-means algorithm [14] to generate the codebook. From the codebook distribution, the model parameters are re-estimated. Different steps are required in order to generate codebooks that are optimally distributed over HMM states:

1. **Take** several realizations of utterances, spoken several times by many speakers.
2. **Determine** the optimal state sequence of each utterance (Viterbi).
3. **Put** the whole observations belonging to each state from all versions of the spoken words into separate cells. Each cell contains the population of a given state.
4. **Apply** VQ to split the population of each cell into M classes within each state.
5. **Re-estimate** the discrete output probability by using the following formula:

$$b_{jk}=\frac{N_k}{N_j} \qquad (5)$$

with $1 \le j \le N$ and $1 \le k \le Lj$

where: N : the number of HMM states, N_k : the number of prototype in the class k and, N_j : total number of prototypes in state j

6. **Refine** model parameters using standard re-estimation formulas.

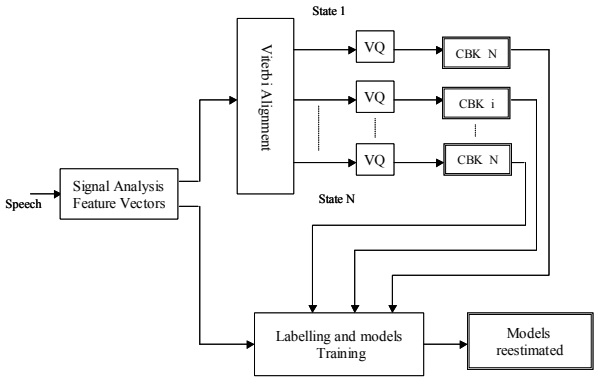

Fig. 3 Description of the training phase involved in the DVQ-HMM-based system

The K-means algorithm is based on the minimization of a distortion criterion. Thus, the phonetic classification of the acoustic vectors is not taken into account during the design of codebooks, and therefore, this information is missing in the acoustic processor of a discrete HMM system. In order to take in account the phonetic information contained in the input vectors, the use of a neural network-based configuration to generate the codebook is proposed to perform the DVQ.

C. The Hybrid Neural Network DVQ

In this method, the standard LBG algorithm is replaced by a neural network VQ algorithm trained on unsupervised mode using the principles of the mutual information theory. Before describing the neural network that is used, and its training algorithm, definitions related to the mutual information theory are briefly recalled in what follows. The mutual information (M.I) is a measure of the information content that one variable contains with respect to another random variable. This means a reduction in the uncertainty of one random variable, i.e. Y, due the knowledge of another variable, i.e. W as described by equation (6):

$$F=M.I(Y,W)=H(Y)-H(Y|W) \qquad (6)$$

$H(Y)$ represents the entropy $Y = \{ y_1, y_2, \ldots, y_M \}$. It is given by equation (6):

$$H(Y)=-\sum_{m=1}^{M}P(y_m)\log P(y_m) \qquad (7)$$

H(Y|W) represents the conditional entropy:

$$H(Y|W) = -\sum_{n=1}^{N} P(W_n) \sum_{m=1}^{M} P(y_m|W_n) \log P(y_m|W_n) \quad (8)$$

This conditional entropy is interpreted as the average incertitude on symbols y_m when symbols W_n are observed. The mutual information *M.I(Y, W)* can then be expressed as follows:

$$F = M.I(Y,W) = \sum_{n=1}^{N} \sum_{m=1}^{M} P(y_m|W_n) \log \frac{P(y_m|W_n)}{P(y_m)P(W_n)} \quad (9)$$

The topology of the network used as vector quantizer is given in Fig. 4. It is a network with two layers. The input layer contains *D* neurons. *D* is the number of components of the feature vector $\mathbf{X}=\{x_1, x_2, ..., x_D\}$ and the output layer with *M* neurons. *M* is equal to the desired codebook size.

The same steps 1 to 3 of the K-means DVQ training algorithm are used. The difference begins from step 4 described below.

At the beginning of the learning procedure, the weights of the neural network are initialized. Each presentation of a feature vector *X(k), with k = 1,...,K,* will result into the activation for each of the *M* neurons in the output layer, denoted $Z_m(k)$, *m = 1, ... , M.* The Euclidean distance, between the weights and the input values has been used for computation of the activation. This distance is calculated as follows:

$$Z_m(k) = \left\| g_m - X \right\| = \sum_{d=1}^{D} (g_{dm} - x_d)^2 \quad (10)$$

For each presentation *k*, the activation of the neuron of output layer with the smallest distance is set to 1.0, and all other activations are set to 0.0.

The conditional probabilities $P(y_m|W_n)$ of the label *m* in the label stream *Y* resulting from the presentation of all feature vectors of word W_n can be computed as follows:

$$P(y_m|W_n) = \frac{1}{L_w} \sum_{l=1}^{L_n} Z_m(l) \quad (11)$$

The probabilities $P(y_m)$ of the label *m* in the label stream *Y*, resulting from the presentation of all features vector *K,* with $K = \sum_{n=1}^{N} L_n$, where *N* is the number of vocabulary words (phonemes), can be computed as follows:

$$P(y_m) = \frac{1}{K} \sum_{k=1}^{K} Z_m(k) \quad (12)$$

These probabilities and the probability $P(W_n)$ which is the *a priori* probability of word W_n can now be used for the computation of mutual information (M.I) by using equation (9).

The training procedure iteratively modifies the weights g_{dm*}, where *m** denotes the label with the largest frequency by using:

$$g_{dm*}(j) = g_{dm*}(j-1) + \Delta g \quad (13)$$

The computation of the change in activation for label *m** is computed according to:

$$\Delta Z_{m*}(k) = \Delta g \, (\Delta g + 2(g_{dm*} - x_d(k)) \quad (14)$$

Equations (11) and (12) permit the computation of both the label stream, and the change in the probabilities of labels. The resulting changes in the mutual information *(ΔF)* can also be computed. If this change is positive, the modification of weight according to equation (14) is accepted. If not, the procedure is repeated with the negative value of Δg. If this does not lead to a positive value of *(ΔF)*, the weight remains unchanged and the next weight is modified in the same way. The network training is stopped once all weights have been visited. At the end of the learning procedure, the weights of connections of the cells represent the prototypes of the codebook and the probabilities $P(y_m)$ represent the discrete output probabilities initials b_{jk}.

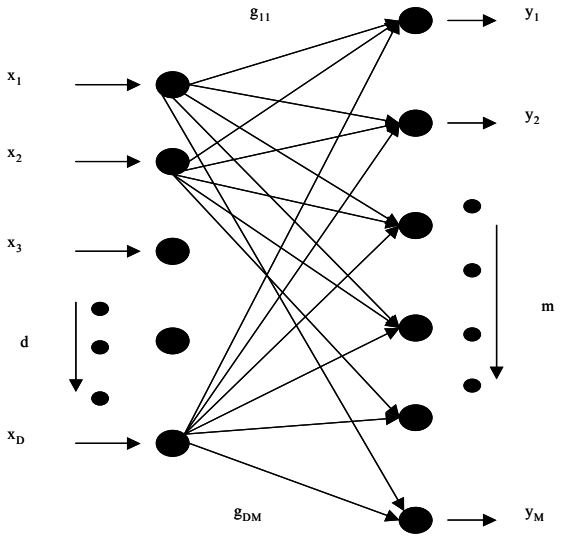

Fig. 4 Topology of the neural network used in the hybrid NN-DVQ configuration

IV. Experimental Results

Various sets of experiments have been carried out in order to assess the improvement involved by the proposed approach. This latter witch consist on the distributed VQ with its two hybrid implementations, namely the K-means-DVQ and the NN-DVQ is compared to the conventional VQ/HMM technique. In the case of the Arabic language, one important

issue that needs to be addressed is the characterization of the particular phonemes such as back consonants, and how the proposed techniques deal with this type of complex phonemes. The objective is to determine the key issue pertaining to Arabic speech recognition by identifying precisely the root of the recognition drawback.

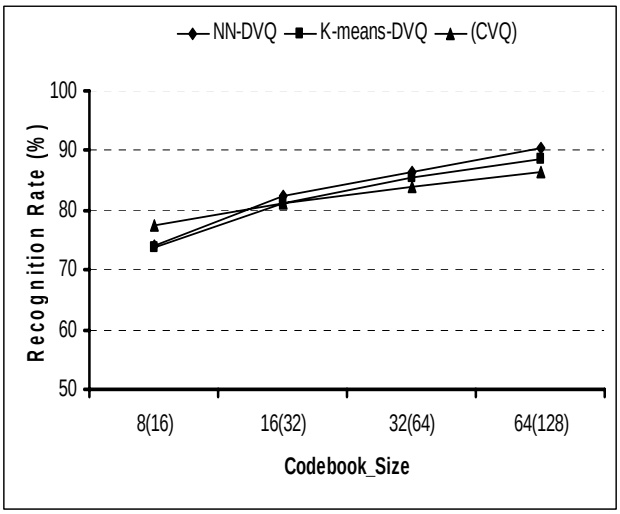

Fig. 5 Comparative phoneme recognition rate in multi-speakers mode using conventional VQ (CVQ), K-means-distributed vector quantization (K-means-DVQ) and neural networks distributed vector quantization (NN-DVQ)

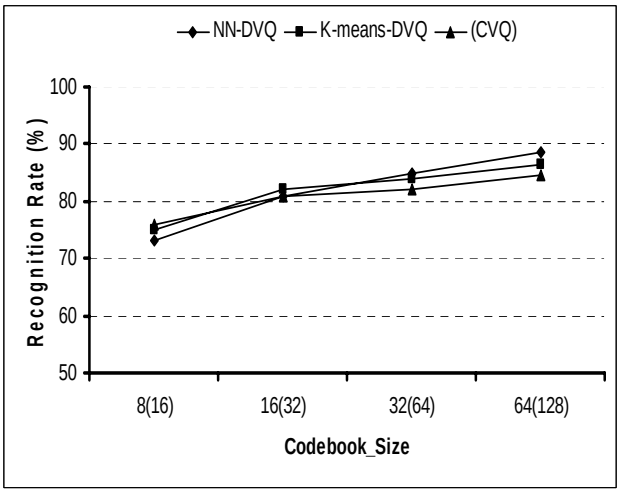

Fig. 6 Comparative phoneme recognition rate in speaker independent mode using conventional VQ (CVQ), K-means distributed vector quantization (K-means-DVQ) and neural networks distributed vector quantization (NN-DVQ)

A. Speech Material

Standard Arabic is distinct from Indo-European languages because of its consonantal nature. It is characterized by the presence of back consonants. These consonants are characterized by having particular vertical places of articulation. In fact, the point of articulation is situated in the rear of the vocal tract. There are four back consonants in Arabic: two glottal /?/ and /h/ classified respectively as plosive and fricative and two pharyngeal /h/ and /ε/ classified as unvoiced fricative and sonorant. The emphasis aspect is a phonetic feature that characterizes the consonants in the Semitic language. There are also four emphatic consonants in the Arabic language: two plosive consonants /t/ and /d/ and two fricative consonants /s/ and /∂/. Designers of systems dedicated to the Arabic language have unanimously observed that emphasis, germination, and vowel lengthening[1] constitute the main root of failure. It is the reason why we focus our experiments on these consonants.

Two test sets of data were used throughout all experiments. The first set is composed of words containing these consonants in different phonetic contexts. These words were pronounced by 80 Algerian speakers and repeated five times. The second set composed of sentences, was also used, since isolated words do not take into account the co-articulation phenomenon.

B. Evaluation of Implemented Systems

In order to evaluate the DVQ, a set of experiments in both multi-speakers and speaker-independent mode has been carried out. The acoustic vector used is a 33-dimensional vector $\{MFCC(11), \Delta MFCC(11), \Delta\Delta MFFC(11)\}$. In the NN-based front-end technique, the value of Δg retained for the modification of network weights, is 0.05. Various sizes of codebooks varying from 8 to 128 were used in a comparison between the conventional (CVQ), the K-means-DVQ and NN-DVQ. Fig. 5 and Fig. 6 show the comparative results in multi-speakers mode and speaker-independent mode, respectively. As expected, in the speaker independent mode, more errors have been observed. We can see that the difference between the two modes is not noticeable, which confirms that the DVQ scheme leads to more robustness of the recognition process. We must note that NN-DVQ exhibits better performance at all codebook sizes.

V. Conclusion

In this paper several strategies were proposed to improve discrete HMM-based automatic speech recognition systems dedicated to the Arabic language. A new approach of discrete HMM (distributed vector quantization: DVQ), based on the tying between Markovian states and the conventional vector quantization, was presented. Two implementation schemes of this approach, namely the K-means DVQ and the NN-DVQ, were tested in both multi-speakers and speaker-independent modes. Results suggest that this new approach with the NN-DVQ variant is more effective in terms of error reduction and of the decoding speed of the discrete HMM. We currently

[1] In Arabic, the vowel duration is semantically relevant.

attempt to apply this new variant to a multiple codebook large vocabulary speech recognition system.. An important challenge is somehow to adapt the vector quantization when it is incorporated into the training optimization process, in such a way that it takes into account the diversity of human-language particularities. In the near term, such integration will no doubt result in massive increases in computation, but will certainly constitute a very promising way towards the design of multilingual speech recognition systems.

REFERENCES

[1] X.D. Huang, H.W. Hon, M.Y. Hwang, and K.F. Lee, "A comparative study of discrete, semi continuous, and continuous hidden Markov models," Computer Speech and Language, vol. 7, pp. 359–368, 1993.
[2] N. Morgan and H. Bourlard, "Continuous speech recognition," IEEE Signal Processing Magazine, vol. 12, no. 3, 1995.
[3] J.C Segura, A.J. Rubio, A.M. Peinado, P. Garcia, and R. Roman, "Multiple VQ hidden Markov modeling for speech recognition," Speech Communication, vol. 14, pp. 163–170, 1994.
[4] Q. Huo and C. Chan, "Contextual vector quantization for speech recognition with discrete hidden Markov model," Pattern recognition, vol. 28 no. 4, pp. 513–517, 1995.
[5] V. Digalakis, S. Tsakalidis, C. Harizakis, and L. Neumeyer, "Efficient speech recognition using sub vector quantization and discrete-mixture HMMs," Computer Speech and Language, vol. 14, pp. 33–46, 2000.
[6] F. Lefevre, "Non parametric probability estimation for HMM-based automatic speech recognition," Computer Speech and Language, vol. 17, pp. 113–136, 2003.
[7] A. Bernard and A. Alwan, "Low-bit-rat distributed speech recognition for packet-based and wireless communication," IEEE Trans. on Speech and Audio Processing, vol. 10 no. 8, pp. 570–580, 2002.
[8] R. Ethman, D.A. Subramaniam, and B.D. Rao, "Improved quantization structure using generalized HMM modeling with application to wideband speech coding," presented at IEEE Int. Conf. on Audio Speech and Signal Processing, Montreal, pp. 161–164, 2004.
[9] M.A Elkhouli, "Hearing distinction of speech sound," Arabic Linguistic and computer science, publication of Tunis university, pp. 267–295, 1989.
[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 28 no. 4, pp. 357–366, 1980.
[11] Y. Linde, A. Buzo, and R.M Gray, "An algorithm for vector quantizer," IEEE Trans. on Communication, vol. 28, no.1, 1980.
[12] L.R Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceeding of the IEEE Trans. Speech Process, vol. 77, no. 2, pp. 257–285, 1989.
[13] P. Hedelin and J. Skoglound, "Vector quantization based on Gaussian mixture models," IEEE Trans. on Speech and Audi Processing, vol. 8, no. 4, pp. 385–401, 2000.
[14] A. Likas, N. Vlassis, and J.J. Verbeck, "The global K-means clustering algorithm," Pattern Recognition, vol. 36, no. 2, pp. 451–461, 2003.