ELSEVIER

Speech Communication 14 (1994) 163–170

SPEECH COMMUNICATION

Multiple VQ hidden Markov modelling for speech recognition

J.C. Segura [1,*], A.J. Rubio [1], A.M. Peinado [1], P. García [1], R. Román [2]

[1] *Dept. Electrónica y Tecnología de Computadores, Univ. Granada, 18071 Granada, Spain*
[2] *Dept. Física Aplicada, Univ. Granada, 18071 Granada, Spain*

(Received 3 December 1992; revised 12 December 1993)

Abstract

In this paper a new variant of HMM, named Multiple VQ HMM (MVQHMM), is presented. Its main characteristic is the use of a separate codebook for each model. Procedures for training and probability evaluation of these models are described. The evaluation procedure combines the quantization distortions of the vector sequences with the discrete HMM generation probabilities. Comparative results on an isolated word recognition system are shown, between MVQHMM and discrete and semi-continuous HMM. These results show that using separate codebooks and including the quantization distortion in the decision criterion improve the performance of the system. Furthermore, the multiple VQ hidden Markov models seem to be more robust than the discrete and semi-continuous ones in relation to the inter-speaker variability of the recognition system.

Zusammenfassung

In diesem Artikel wird eine neue Variante der HMM vorgestellt, Multiple VQ HMM (MVQHMM) genannt. Ihr Hauptmerkmal liegt in der Benutzung von separaten Codebooks für jedes Modell. Es werden Verfahren beschrieben, die zum Training und zur Bewertung der Wahrscheinlichkeit bestimmt sind. Das Bewertungsverfahren verbindet die infolge der Vektorquantisierung auftretenden Fehler mit den diskreten Emissionswahrscheinlichkeiten. Die Ergebnisse der MVQHMM und der diskreten und semi-continious HMM, angewendet auf ein Erkennungssystem von Einzelwörtern, werden verglichen. Die Ergebnisse zeigen, daß die Benutzung von separaten Codebooks und die Berücksichtigung der Quantisierungsfehler in dem Entscheidungskriterien die Leistung des Systems verbessern. Außerdem scheinen die Multiple VQ Hidden Markov Modelle in bezug auf die sprecherspezifischen Abweichungen zuverlässiger zu sein.

Résumé

Cet article présente une nouvelle variante des Modèles de Markov cachés (HMM) "à Quantification Vectorielle Multiple" (MVQHMM). Sa caractéristique principale est d'utiliser un dictionnaire différent pour chaque modèle. On décrit les algorithmes d'apprentissage et de reconnaissance pour ces modèles. La procédure de reconnaissance combine le calcul de l'erreur de quantification de la séquence de vecteurs avec celui de la probabilité de sa génération par le HMM. On donne des résultats sur un système de reconnaissance de mots isolés, en comparant les

* Corresponding author. Tel. 34-58-243283; Fax. 34-58-243230; e-mail: segura@hal.ugr.es

SSDI 0167-6393(93)E0083-8

MVQHMM et les HMM discrets et semi-continus: utiliser des dictionnaires différents et inclure l'erreur de quantification dans le calcul améliore les performances du système. De plus, les MVQHMM apparaissent plus robustes à la variabilité interlocuteur que les modèles discrets et semi-continus.

Key words: Hidden Markov models; Speech recognition

1. Introduction

Discrete hidden Markov models have been successfully applied to speech recognition, as acoustics models for different types of decision units (phones, words, etc). Their main advantages are moderate computational requirements and high versatility. Nevertheless, one of the main disadvantages is the implicit discretization of the observations which produces information loss that in turn may deteriorate the models performance (Rabiner et al., 1985).

At least two alternatives to the discrete modelling have been proposed in the literature. The first one corresponds to continuous models which obviate the quantization problem by directly modelling the observations as continuous probability density functions (pdf's). The main problems of this modelling are their high computational cost and the large number of parameters to be estimated (mainly covariance matrices). When the training data are insufficient the performance of the system is also deteriorated (Rabiner et al., 1985).

The second alternative corresponds to the semi-continuous Markov models, proposed in (Huang and Jack, 1989). Under this approach, the VQ codewords are modelled as multivariate pdf's and the VQ process is modified in such a way that multiple candidates are generated, one per codeword. A probability value is associated to each candidate, according to the corresponding codeword pdf. These probabilities are used to obtain the observation probabilities (see Eq. (5)). From the point of view of continuous mixtures HMM, the semi-continuous HMM approach uses a common set of pdf's (from the VQ codebook) to build all models state mixtures, which reduces the number of parameters to be estimated. From the discrete modelling point of view, the use of multiple candidates reduces the information loss in the VQ process.

In this work, a new variant of HMM is introduced. This new approach uses a VQ process in which every model has its own codebook. For every input vector sequence, the VQ process generates a symbol sequence corresponding to each one of the codebooks. A more precise VQ is obtained by using a specific codebook for each model, which can characterize more precisely its acoustic productions. Furthermore, as we will show later, this modified VQ process gives useful information that can be used in the classification procedure of unknown vector sequences. This information is essentially contained in quantization distortions of the unknown vector sequence with the different model codebooks.

The remaining of the paper is organized as follows. In Section 2, we describe the formalism of multiple VQ hidden Markov models (MVQHMM). Evaluation formulas are derived from a general formulation of hidden Markov modelling and compared with discrete, continuous and semi-continuous hidden Markov models.

In Section 3, we describe the particular implementation of MVQHMM modelling used in the isolated word recognition system presented later in this paper, as well as the algorithm used for models training.

In Section 4, we present the experimental environment used for testing the proposed recognition system. In this section we also present comparative results between three recognition systems based on discrete, semi-continuous and MVQHMM models.

Finally in Section 5, we summarize the conclusions of the present work.

2. Multiple VQ HMM

An MVQHMM is composed of a VQ codebook, modelling the different acoustic produc-

tions of the modelled unit, and a standard discrete HMM modelling the temporal behavior of the process observations (VQ codewords). Each decision unit (phone, word, etc) has its own set of acoustic prototypes and therefore it can be expected that the quantization process with the appropriate codebook will be more accurate than the traditional VQ with a shared codebook. Unlike that in the traditional VQ-based HMM approach, there is no a priori selection of the optimal quantization and it is carried out along with the final decision about the correct model.

As shown in (Shore and Burton, 1983; Furui, 1988; Bergh et al., 1985), the quantization distortions of an input sequence with the different model codebooks can be used to classify unknown vector sequences. When evaluating MVQHMM generation probabilities, quantization distortions are combined with the generation probabilities of the discrete HMM to obtain a final probability, which is used as the classifying criterion.

In Fig. 1 the block diagram of an MVQHMM-based isolated word recognition system is depicted. Given a model λ, $P(X_1^T \mid \lambda)$ is the generation probability of the vector sequence $X_1^T = x_1 x_2 \ldots x_t \ldots x_T$, $P(X_1^T \mid O_1^T, \lambda)$ is the quantization probability of the vector sequence X_1^T in the symbol sequence $O_1^T = o_1 o_2 \ldots o_t \ldots o_T$, and $P(O_1^T \mid \lambda)$ is the generation probability of the symbol sequence O_1^T.

Given an observation sequence $X_1^T = x_1 x_2 \ldots x_t \ldots x_T$, where x_t is a vector of acoustic

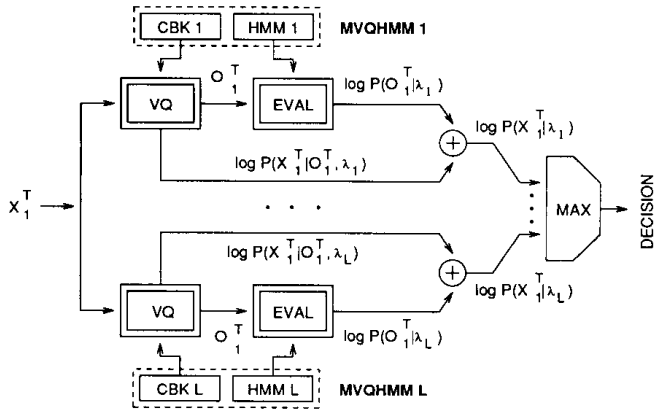

Fig. 1. MVQHMM-based isolated word recognition system.

characteristics, the generation probability $P(X_1^T \mid \lambda)$ for an HMM λ can be expressed in the following way:

$$P(X_1^T \mid \lambda) = \sum_{S_1^T} P(X_1^T \mid S_1^T, \lambda) P(S_1^T \mid \lambda), \tag{1}$$

$$P(X_1^T \mid S_1^T, \lambda) = \prod_{t=1}^{T} P(x_t \mid s_t, \lambda), \tag{2}$$

$$P(S_1^T \mid \lambda) = P(s_1 \mid \lambda) \prod_{t=2}^{T} P(s_t \mid s_{t-1}, \lambda), \tag{3}$$

where $S_1^T = s_1 s_2 \ldots s_t \ldots s_T$ is a state sequence, and s_t is the model state at time t. The summation in S_1^T represents the sum over all possible states sequences of the model. Assuming that $P(x_t \mid s_t, \lambda)$ can be expressed as a probability density function mixture, we can write

$$P(x_t \mid s_t, \lambda) = \sum_{o_t \in V(s_t, \lambda)} P(x_t \mid o_t, s_t, \lambda) P(o_t \mid s_t, \lambda), \tag{4}$$

where $V(s_t, \lambda)$ is a set of acoustic prototypes belonging to the state s_t of the model λ. $P(x_t \mid o_t, s_t, \lambda)$ are the mixtures pdf's and $P(o_t \mid s_t, \lambda)$ are the mixture coefficients. This formulation is essentially the same used in continuous mixture HMM (Rabiner et al., 1985).

Assuming that the set of prototypes V is independent of both the state and model considered, it can be written

$$P(x_t \mid s_t, \lambda) = \sum_{o_t \in V} P(x_t \mid o_t) P(o_t \mid s_t, \lambda). \tag{5}$$

This expression is equivalent to the one used in (Huang and Jack, 1989) in the formulation of semi-continuous HMM.

Furthermore, assuming that the classes represented by the prototypes are disjoint or little overlapped, the former expression can be approximated as

$$P(x_t \mid s_t, \lambda) = P(x_t \mid o_t^*) P(o_t^* \mid s_t, \lambda), \tag{6}$$

$$o_t^* = \arg\max_{o_t \in V} \{P(x_t \mid o_t)\}. \tag{7}$$

Eq. (7) represents the quantization condition of input vector x_t, where o_t^* is the corresponding

codeword and V is the codebook. Using Eqs. (1), (2) and (6) the following relation can be obtained:

$$\mathrm{P}(X_1^{\mathrm{T}} \mid \lambda) = \mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*})\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda), \tag{8}$$

where

$$\mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*}) = \prod_{t=1}^{T} \mathrm{P}(x_t \mid o_t^*), \tag{9}$$

$$\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda) = \sum_{S_1^{\mathrm{T}}} \mathrm{P}(O_1^{\mathrm{T}*} \mid S_1^{\mathrm{T}})\mathrm{P}(S_1^{\mathrm{T}} \mid \lambda). \tag{10}$$

In Eq. (9) $O_1^{\mathrm{T}*} = o_1^* \ldots o_T^*$ is the best symbol sequence (in the sense of Eq. (7)). Note that due to the fact that the quantization probability $\mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*})$ is independent of the current model, we only have to compute the generation probability of the symbol sequence, $\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda)$. This is essentially the discrete HMM formulation.

Finally, the evaluation formulas for the MVQHMM can be obtained under the assumption that the classes set V is disjoint and it only depends on the considered model. Thus, we can write

$$\mathrm{P}(x_t \mid s_t, \lambda) = \mathrm{P}(x_t \mid o_t^*, \lambda)\mathrm{P}(o_t^* \mid s_t, \lambda), \tag{11}$$

$$o_t^* = \arg \max_{o_t \in V(\lambda)} \{\mathrm{P}(x_t \mid o_t, \lambda)\}. \tag{12}$$

The only difference between (6)–(7) and (11)–(12) is that the codebook V now depends on model λ and therefore the VQ process must be model specific. From Eqs. (1), (2) and (11) it is easy to obtain the following relation:

$$\mathrm{P}(X_1^{\mathrm{T}} \mid \lambda) = \mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*}, \lambda)\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda), \tag{13}$$

where

$$\mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*}, \lambda) = \prod_{t=1}^{T} \mathrm{P}(x_t \mid o_t^*, \lambda), \tag{14}$$

$$\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda) = \sum_{S_1^{\mathrm{T}}} \mathrm{P}(O_1^{\mathrm{T}*} \mid S_1^{\mathrm{T}}, \lambda)\mathrm{P}(S_1^{\mathrm{T}} \mid \lambda). \tag{15}$$

In this expression $\mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*}, \lambda)$ represents the quantization probability of the observation sequence X_1^{T} into the symbol sequence $O_1^{\mathrm{T}*}$ of λ model and $\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda)$ is the generation probability of the symbols sequence $O_1^{\mathrm{T}*}$. Now the quantization probability depends on the model, and it must be incorporated into the classifying criterion which must be based on MVQHMM generation probabilities $\mathrm{P}(X_1^{\mathrm{T}} \mid \lambda)$ instead of on symbol generation probability $\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda)$.

3. Implementation of the MVQHMM modelling

In this section, we are going to describe the implementation used in this work for the MVQHMM modelling presented in the previous section.

The first subsection describes the parametric model used for the codewords of the different models codebook and, therefore, a method for estimating the quantization probabilities presented in the previous section.

The last subsection describes an algorithm for training MVQHMM models. This algorithm is based on a maximum likelihood approach to the estimation of both the codebook and the HMM of an MVQHMM model.

3.1. Quantization probabilities

In order to evaluate quantization probabilities, it is necessary to choose a parametric model for the codewords. In this work each one of them is modelled as a Gaussian with an identity covariance matrix, and therefore the following equations can be written:

$$\Sigma_\lambda = \sigma_\lambda^2 I, \tag{16}$$

$$p(x_t \mid o_t, \lambda) = (2\pi)^{-p/2}(\sigma_\lambda^2)^{-p/2} \times \exp\left\{-\frac{\| x_t - \mu_{o_t,\lambda} \|^2}{2\sigma_\lambda^2}\right\}, \tag{17}$$

$$o_t^* = \arg \max_{o_t \in V(\lambda)} \{\mathrm{P}(x_t \mid o_t)\} = \arg \min_{o_t \in V(\lambda)} \left\{\| x_t - \mu_{o_t,\lambda} \|^2\right\}, \tag{18}$$

$$\frac{1}{T}\log \mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*}, \lambda) = -\frac{p}{2}\log 2\pi - \frac{p}{2}\log \sigma_\lambda^2 - \frac{D_\lambda(X_1^{\mathrm{T}})}{2\sigma_\lambda^2}, \tag{19}$$

$$D_\lambda(X_1^{\mathrm{T}}) = \frac{1}{T}\sum_{t=1}^{T} \| x_t - \mu_{o_t^*,\lambda} \|^2, \tag{20}$$

codeword and V is the codebook. Using Eqs. (1), (2) and (6) the following relation can be obtained:

$$\mathrm{P}(X_1^{\mathrm{T}} \mid \lambda) = \mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*})\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda), \tag{8}$$

where

$$\mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*}) = \prod_{t=1}^{T} \mathrm{P}(x_t \mid o_t^*), \tag{9}$$

$$\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda) = \sum_{S_1^{\mathrm{T}}} \mathrm{P}(O_1^{\mathrm{T}*} \mid S_1^{\mathrm{T}})\mathrm{P}(S_1^{\mathrm{T}} \mid \lambda). \tag{10}$$

In Eq. (9) $O_1^{\mathrm{T}*} = o_1^* \ldots o_T^*$ is the best symbol sequence (in the sense of Eq. (7)). Note that due to the fact that the quantization probability $\mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*})$ is independent of the current model, we only have to compute the generation probability of the symbol sequence, $\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda)$. This is essentially the discrete HMM formulation.

Finally, the evaluation formulas for the MVQHMM can be obtained under the assumption that the classes set V is disjoint and it only depends on the considered model. Thus, we can write

$$\mathrm{P}(x_t \mid s_t, \lambda) = \mathrm{P}(x_t \mid o_t^*, \lambda)\mathrm{P}(o_t^* \mid s_t, \lambda), \tag{11}$$

$$o_t^* = \arg \max_{o_t \in V(\lambda)} \{\mathrm{P}(x_t \mid o_t, \lambda)\}. \tag{12}$$

The only difference between (6)–(7) and (11)–(12) is that the codebook V now depends on model λ and therefore the VQ process must be model specific. From Eqs. (1), (2) and (11) it is easy to obtain the following relation:

$$\mathrm{P}(X_1^{\mathrm{T}} \mid \lambda) = \mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*}, \lambda)\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda), \tag{13}$$

where

$$\mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*}, \lambda) = \prod_{t=1}^{T} \mathrm{P}(x_t \mid o_t^*, \lambda), \tag{14}$$

$$\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda) = \sum_{S_1^{\mathrm{T}}} \mathrm{P}(O_1^{\mathrm{T}*} \mid S_1^{\mathrm{T}}, \lambda)\mathrm{P}(S_1^{\mathrm{T}} \mid \lambda). \tag{15}$$

In this expression $\mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*}, \lambda)$ represents the quantization probability of the observation sequence X_1^{T} into the symbol sequence $O_1^{\mathrm{T}*}$ of λ model and $\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda)$ is the generation probability of the symbols sequence $O_1^{\mathrm{T}*}$. Now the quantization probability depends on the model, and it must be incorporated into the classifying criterion which must be based on MVQHMM generation probabilities $\mathrm{P}(X_1^{\mathrm{T}} \mid \lambda)$ instead of on symbol generation probability $\mathrm{P}(O_1^{\mathrm{T}*} \mid \lambda)$.

3. Implementation of the MVQHMM modelling

In this section, we are going to describe the implementation used in this work for the MVQHMM modelling presented in the previous section.

The first subsection describes the parametric model used for the codewords of the different models codebook and, therefore, a method for estimating the quantization probabilities presented in the previous section.

The last subsection describes an algorithm for training MVQHMM models. This algorithm is based on a maximum likelihood approach to the estimation of both the codebook and the HMM of an MVQHMM model.

3.1. Quantization probabilities

In order to evaluate quantization probabilities, it is necessary to choose a parametric model for the codewords. In this work each one of them is modelled as a Gaussian with an identity covariance matrix, and therefore the following equations can be written:

$$\Sigma_\lambda = \sigma_\lambda^2 I, \tag{16}$$

$$p(x_t \mid o_t, \lambda) = (2\pi)^{-p/2}(\sigma_\lambda^2)^{-p/2} \times \exp\left\{-\frac{\| x_t - \mu_{o_t,\lambda} \|^2}{2\sigma_\lambda^2}\right\}, \tag{17}$$

$$o_t^* = \arg \max_{o_t \in V(\lambda)} \{\mathrm{P}(x_t \mid o_t)\} = \arg \min_{o_t \in V(\lambda)} \left\{\| x_t - \mu_{o_t,\lambda} \|^2\right\}, \tag{18}$$

$$\frac{1}{T}\log \mathrm{P}(X_1^{\mathrm{T}} \mid O_1^{\mathrm{T}*}, \lambda) = -\frac{p}{2}\log 2\pi - \frac{p}{2}\log \sigma_\lambda^2 - \frac{D_\lambda(X_1^{\mathrm{T}})}{2\sigma_\lambda^2}, \tag{19}$$

$$D_\lambda(X_1^{\mathrm{T}}) = \frac{1}{T}\sum_{t=1}^{T} \| x_t - \mu_{o_t^*,\lambda} \|^2, \tag{20}$$

where I is an identity matrix, $D_\lambda(X_1^T)$ represents the average quantization distortion of the vector sequence X_1^T with the λ model codebook and $\mu_{o_t^*,\lambda}$ is the mean vector corresponding to the codeword with least Euclidean distance to the x_t vector. The value p is the number of vector components, and therefore $p\sigma_\lambda^2$ is the expected value of the mean VQ distortion of a vector sequence belonging to the model λ. The value σ_λ^2 can be estimated from the training set in the following manner:

$$\sigma_\lambda^2 = \frac{\overline{D}_\lambda}{p}, \tag{21}$$

where $\overline{D}_\lambda$ is the mean VQ distortion of λ model training set. Eq. (18) is the quantization criterion, therefore the distortion measure is simply a Euclidean distance.

3.2. MVQHMM training

Construction of an MVQHMM requires the specification of the model codebook and the HMM transition and production probabilities. In a maximum likelihood estimation approach, this is done by maximizing the a posteriori observation probabilities (13). This can be done in a two step procedure.

In the first step, the model VQ codebooks are built with a K-means clustering algorithm with binary splitting initialization and Euclidean distance like the one described in (Peinado et al., 1991). Due to the definitions (16)–(18), this process maximizes the quantization probabilities (14) by means of a minimization of the Euclidean distance between training vectors and codebook centroids.

In the second step, the previously built VQ codebooks are used to quantize the training vector sequences, each one with the corresponding codebook. With the obtained symbol sequences the discrete HMM are trained using a standard Baum–Welch re-estimation procedure, which maximizes the symbol generation probabilities (15). Initial HMM models are built from a linear segmentation of the training sequences.

4. Experimental results

Comparative experiments have been carried out between MVQHMM and discrete and semi-continuous HMM on an isolated word recognition system with a 16-words vocabulary. The vocabulary is formed by the 10 Spanish digits and six key words (CUERPO, HOMBRO, CODO, MUÑECA, MANO, DEDOS). There are 3 repetitions of every word uttered by 40 speakers (20 male and 20 female).

Words have been sampled at 8 kHz with 12 bits. An order 10 LPC analysis has been carried out on 32 ms frames every 8 ms. Each frame has been characterized by a 25 component vector formed by 12 liftered cepstrum coefficients (Juang et al., 1987), 12 liftered delta cepstrum coefficients and delta energy (Furui, 1986). Delta cepstrum and delta energy have been weighed with 0.925 and 0.728, respectively, in order to optimize the performance of the Euclidean distance used in the codebook construction. Note that this is equivalent, but formally simpler, to the use of a weighed Euclidean distance measure. In addition, vectors sequences are decimated to a period of 16 ms, after the parameter extraction procedure.

In all cases 10 states left-to-right HMM have been used. A postprocessor that takes into account the duration of the model states is always added to the system. Details on preprocessing, parameter extraction and post-processing along with a complete description of the recognition system based on discrete Markov models can be found in (Peinado et al., 1991).

Due to the limited number of speakers in the speech database, the testing of the system is carried out by means of a procedure similar to the so-called *leaving-one-out* (Duda and Hart, 1973).

In the semi-continuous HMM implementation used in this work, only $L = 8$ most likely values of the pdf's are selected because the other ones have significantly lower probability values and can be therefore neglected. This approach significantly reduces the computational cost, and sometimes improves the recognition performance of the system (Huang and Jack, 1989).

With the previous modifications, the conver-

sion process from a discrete HMM based recognition system to an SCHMM based one is as follows:

1. Estimation of codewords covariance matrices.
2. Quantization of the training sequences generating the $L = 8$ most likely codewords and their corresponding probability values.
3. Using the discrete models obtained for the DHMM based recognition system as initial models, train the semi-continuous models.

Multi-speaker error rates

When testing the system in a multi-speaker environment, two repetitions of each word uttered by all the 40 speakers have been used to train the system and the remaining repetition to test it. This procedure is repeated for three disjoint partitions of the database and the obtained results are averaged.

Table 1 shows the error rates obtained for the multi-speaker environment test. In this environment, the results are similar to those on the speaker-independent environment. For 4 codewords per model, the error rate for MVQHMM models is significantly greater than the corresponding to DHMM and SCHMM. This is due to the fact that 4 codewords are not enough to properly model the acoustic productions of most of the words in the vocabulary. Nevertheless, for 8 codewords per model the error rate is similar to the obtained for DHMM, and for 16 and 32 codewords, the error rate is similar or lower than the corresponding to both DHMM and SCHMM.

Speaker-independent error rates

When testing the system in a speaker-independent environment, thirty two speakers (16 male and 16 female) have been used to train the system and the remaining eight (4 male and 4 female) to test it. This procedure is repeated for five disjoint partitions of the database and the obtained results are averaged. This is equivalent to use a database with a training set of 32 speakers and a testing set of 40 different ones.

Table 1
Multi-speaker error rates

CDWDS	DHMM	SCHMM	MVQHMM
4– 64	3.07%	1.77%	4.64%
8–128	1.93%	1.15%	2.19%
16–256	1.35%	0.73%	0.73%
32–512	0.99%	0.52%	0.36%

Table 2
Speaker-independent error rates

CDWDS	DHMM	SCHMM	MVQHMM
4– 64	4.38%	3.02%	4.27%
8–128	3.59%	2.14%	2.45%
16–256	2.81%	1.56%	0.94%
32–512	2.03%	1.41%	0.89%

Table 2 shows the error rates for the system with discrete HMM (DHMM), semi-continuous HMM (SCHMM) and MVQHMM. The first column (CDWDS) indicates both the number of codewords of every model codebook for MVQHMM and the number of codewords in the codebook for DHMM and SCHMM. In the case of 4 codewords per MVQHMM codebook, the error rate is similar to the DHMM's but higher than SCHMM's. As in the multi-speaker test, this is due to the fact that four codewords are insufficient to properly model the acoustic productions of the model. Nevertheless, for 8 codewords the MVQHMM error is lower than for DHMM and similar to SCHMM, and for 16 or 32 codewords is even lower than for SCHMM.

Inter-speaker robustness

Figs. 2, 3 and 4 show the multi-speaker and speaker-independent error rates for the DHMM, SCHMM and MVQHMM models, respectively.

For both the DHMM and MVQHMM models, there is a significant increment in the error rate between multi-speaker and speaker-independent tests, about a 1.5% for the DHMM models and about a 1% for the SCHMM models. This increment in the error rate is due to the additional inter-speaker variability of the speaker-independent test.

Nevertheless, for the MVQHMM models only

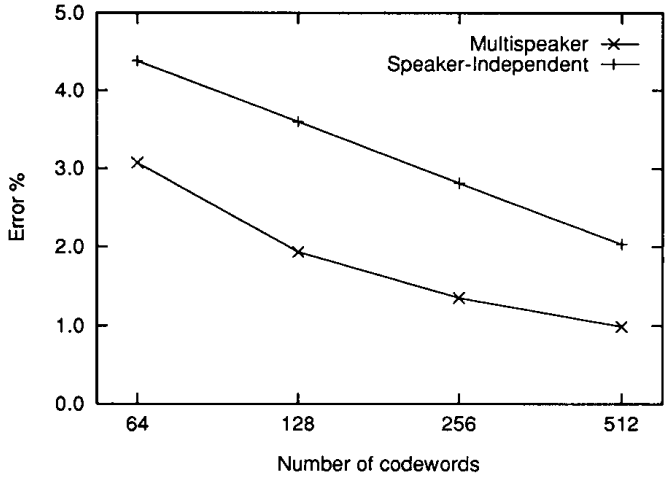

Fig. 2. Error rates for DHMM.

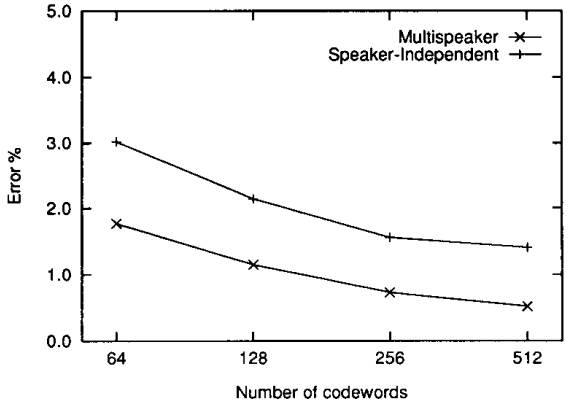

Fig. 3. Error rates for SCHMM.

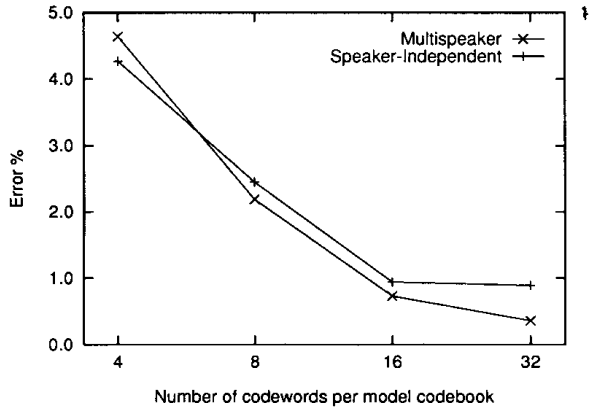

Fig. 4. Error rates for MVQHMM.

a little increment of the error rate (about a 0.25%) between multi-speaker and speaker-independent test is obtained. This result seems to indicate that MVQHMM models are more robust than the DHMM and SCHMM models with respect to inter-speaker variability.

5. Conclusion

In this work we have presented a new variant of discrete hidden Markov modelling, whose main characteristic is the use of specific codebooks for each model.

We have obtained the evaluation formulas for these new models from a general hidden Markov model approach, and we have also presented an algorithm for model training based on a maximum likelihood approach. The evaluation and re-estimation procedures do not require the use of algorithms different from the ones used in discrete hidden Markov modelling.

The experimental results show that, in the speaker-independent situation, the MVQHMM based system performance is always better than the discrete HMM based one for an equivalent number of codewords. The computational cost is essentially the same except for a little overhead due to the probability composition.

In comparison with the results obtained with the semi-continuous HMM based recognition system, these ones yield a lower error rate for a small size codebook (i.e., 64 and 128). However, for a sufficient number of codewords (i.e., 256 and 512), the MVQHMM models have a meaningfully lower error rate.

MVQHMM models seem to be more robust than DHMM and SCHMM ones with respect to inter-locutor variability.

6. References

A. Bergh, F. Soong and L. Rabiner (1985), "Incorporation of temporal structure into a vector-quantization-based preprocessor for speaker-independent, isolated word recognition", *ATT Tech. J.*, Vol. 64, No. 5, pp. 1047–1063.

R. Duda and P. Hart, "Estimating the error rate", *Pattern Classification and Scene Analysis*, Vol. I, pp. 211–256.

S. Furui (1986), "Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-34, No. 1, pp. 52–59.

S. Furui (1988), "A VQ-based preprocessor using cepstral dynamic features for speaker-independent large vocabulary word recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-36, No. 7, pp. 980–987.

X. Huang and M. Jack (1989), "Unified techniques for vector quantisation and hidden Markov modeling using semi-continuous models", *Proc. Internat. Conf. Acoust. Speech Signal Process. '89*, Vol. 2, pp. 639–642.

B. Juang, L. Rabiner and J. Wilpon (1987), "On the use of bandpass liftering in speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-35, No. 7, pp. 947–954.

A. Peinado, J. López, V. Sánchez, J. Segura and A. Rubio (1991), "Improvements in HMM-based isolated word recognition system", *IEE Proc. I*, Vol. 138, No. 3, pp. 201–206.

L. Rabiner, B. Juang, S. Levinson and M. Sondhi (1985), "Recognition of isolated digits using Hidden Markov Models with continuous mixture densities", *ATT Tech. J.*, Vol. 64, No. 6, pp. 1211–1234.

J. Shore and D. Burton (1983), "Discrete utterance speech recognition without time alignment", *IEEE Trans. Inform. Theory*, Vol. IT-29, pp. 473–491.