Discriminative Codebook Design Using Multiple Vector Quantization in HMM-Based Speech Recognizers

Antonio M. Peinado, *Member, IEEE*, José C. Segura, *Member, IEEE*, Antonio J. Rubio, Pedro García, and José L. Pérez, *Member, IEEE*

Abstract— **Recent research on multiple vector quantization (MVQ) has shown the suitability of such technique for speech recognition. Basically, MVQ proposes the use of one separated VQ codebook for each recognition unit. Thus, a MVQHMM model is composed of a VQ codebook and a discrete HMM model. This technique allows the incorporation in the recognition dynamics of the input sequence information wasted by discrete HMM models in the VQ process. The use of distinct codebooks also allows to train them in a discriminative manner. In this paper, we propose a new VQ codebook design method for MVQ-based systems, obtained from a modified maximum mutual information estimation. This method provides meaningful error reductions and is performed independently from the estimation of the discrete HMM part of the MVQ model. The results show that the proposed discriminative design turns the MVQHMM technique into a powerful acoustic modeling tool in comparison with other classical methods as discrete or semicontinuous HMM's.**

I. Introduction

During the last years, hidden Markov models (HMM) have been successfully applied to acoustic modeling for speech recognition. Two main variations of HMM's have been widely used: discrete HMM's (DHMM) and continuous HMM's (CHMM). The first ones use nonparametric discrete output probability distributions, due to a previous VQ process. CHMM's use parametric densities to model the output probabilities [1]. The main problem of DHMM's is the loss of information about the input signal during the VQ process. CHMM's avoid this problem using probability density functions (*pdf*s). Thus, CHMM modeling seems to be a more flexible and complete tool for speech modeling. In spite of this, they are not always used for the implementation of speech recognition systems. There are several reasons for it. The main problem is the large number of parameters to obtain. In order to obtain a good estimation of them, a big amount of computation and a large database is required. These requirements can not be always satisfied with the available resources.

In order to avoid such problems of continuous modeling, Huang *et al* [2] propose the use of *Semicontinuous* HMM (SCHMM) models, similar to CHMM's but forcing all the models to share the same set of *pdf*s. Huang has shown that SCHMM's can achieve better results than CHMM's. Besides, our group has recently proposed new approaches based on the use of *multiple vector quantization* (MVQ) for HMM's [3,4]. The basic MVQ-based model is the MVQHMM, or simply MVQ, and is composed of a VQ codebook and a discrete HMM. These new models have been introduced as a direct way to incorporate to the system dynamics the information lost in the VQ process when using the discrete approach. In order to do this, each MVQ model uses its own VQ codebook to evaluate the average distortion of the input utterance. With the same amount of computation in recognition, the MVQ modeling can outperform DHMM's and achieve similar or better results than SCHMM's (with less computation) [5].

In the case of MVQ models, the use of one specific codebook per recognition unit allows us to train it in a discriminative manner. In this paper we propose a discriminative method for codebook design in a MVQ-based system, derived from the maximum mutual information (MMI) estimation [6], modified to control the error rate reduction. The resultant centroid updating formulas, obtained from gradient techniques, resemble the LVQ procedures for classifier design [7], although avoiding the fixed duration restriction of these procedures (see [8]). One important feature of the proposed design is its independence from the estimation of the discrete HMM part of the model, simplifying the computational complexity of the discriminative training. The experimental results show that this design improves the reference error rates for all the tested codebook sizes. This improvement is more important for small codebooks, for which the standard ML-estimated MVQ model performs worse in relation to DHMM's and SCHMM's.

The rest of the paper is organized as follows. Section II describes the MVQ modeling and its maximum likelihood (ML) estimation, comparing its performance with that of other HMM models. Section III discusses the MMI estimation over DHMM and MVQ models. In Section IV, a MMI-based codebook design is proposed for MVQ models. Section V deals with the application of the proposed VQ design to speech recognition, introducing several modifications to the original MMI estimation in order to obtain a meaningful test error rate reduction. The paper ends with the conclusions and a discussion of the future applications of the work.

Manuscript received April 24, 1994; revised October 25, 1995. The associate editor coordinating the review of this paper and approving it for publication was Dr. Amro El-Jaroudi.

The authors are with the Dipartimento de Electrónica y Tecnologia de Computadores, Facultad de Ciencias, Universidad de Granada, 18071-Granada, Spain.

Publisher Item Identifier S 1063-6676(96)02440-6.

II. Multiple Vector Quantization HMM Modeling

A continuous HMM model uses a mixture of *pdf*s to model the output probabilities in the following form:

$$b_i(\boldsymbol{x}) = P(\boldsymbol{x}|s_i,\lambda) = \sum_{v_k \in V(s_i,\lambda)} P(\boldsymbol{x}|v_k,s_i,\lambda)P(v_k|s_i,\lambda) \tag{1}$$

where each $P(\boldsymbol{x}|v_k,s_i,\lambda)$ is a log-concave or elliptically symmetric density [1] (Gaussian along this work) with mean vector $\boldsymbol{y}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, and $\boldsymbol{x}$ is the input vector. Each *pdf* is labeled by one symbol v_k that varies in the set $V(s_i,\lambda)$ defined for state s_i in model λ.

The simplification of (1) leads to different HMM approaches. For example, doing $V(s_i,\lambda) = V \forall s_i,\lambda$, we obtain a SCHMM. Furthermore, if we assume nonoverlapped *pdf*s a SCHMM becomes a DHMM.

Another simplification (MVQ models) can be derived by assuming a different set of *pdf*s $V(\lambda)$ for each model λ, and considering nonoverlapped densities. Thus

$$b_i(\boldsymbol{x}) = P(\boldsymbol{x}|o,\lambda)P(o|s_i,\lambda) \tag{2a}$$

$$o = \max_{v_j \in V(\lambda)}{}^{-1}[P(\boldsymbol{x}|v_j,\lambda)]. \tag{2b}$$

It can be proved that, for an input sequence $X = \boldsymbol{x}_1 \cdots \boldsymbol{x}_T$, the density $P(X|\lambda)$ can be expressed as

$$P(X|\lambda) = P(X|O,\lambda)P(O|\lambda) \tag{3}$$

where $O = o_1 \cdots o_T$ is the sequence of symbols obtained by (2b) corresponding to X for the model λ. We shall refer to $P(X|O,\lambda)$ and $P(O|\lambda)$ as *quantization* and *generation* probabilities, respectively.

If we consider that the model parameter set can be decomposed as $\lambda = (\theta,\phi)$, where θ represents the parameter set of densities $P(\boldsymbol{x}|v_j,\lambda)$ and ϕ is the parameter set of the discrete HMM model, it can be proved that the ML estimation of λ is obtained from the independent ML estimation of θ and ϕ [5] (this statement is not exactly true in CSR due to the use of subword units). The first parameter set (mean vectors and covariance matrices) can be obtained from a VQ codebook $\{\boldsymbol{y}_j, j = 1,\cdots,M\}$ (trained using the LBG algorithm, for example). The second one is estimated by applying the Baum–Welch algorithm, as for DHMM models.

A. Recognition with MVQ Models

A convenient form for the densities $P(\boldsymbol{x}|o,\lambda)$ in expression (2a) is Gaussian with covariance matrix $\Sigma_\lambda = \sigma_\lambda^2 I$, where I is the identity matrix and σ_λ^2 is the average distortion per center and per feature of the codebook θ associated to model λ [4]. Thus, the quantization log-probability for an input sequence X is written as

$$\log P(X|O,\lambda) = \sum_{t=1}^{T} \log P(\boldsymbol{x}_t|o_t,\lambda) = -\frac{pT}{2}\left[\log(2\pi\bar{D}_\lambda/p) + \frac{D_\lambda(X)}{\bar{D}_\lambda}\right] \tag{4}$$

where

$$\bar{D}_\lambda = p\sigma_\lambda^2$$

$$D_\lambda(X) = \frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{x}_t - \boldsymbol{y}_t\|^2 \tag{5}$$

and p is the dimension of the considered feature space. $\bar{D}_\lambda$ is the average distortion of the codebook associated to model λ, and $D_\lambda(X)$ is the average distortion of the input sequence X in the same codebook. The log-probability of (4) can be weighted in order to obtain an optimal composition with the log-generation probability.

For simplicity, the different techniques introduced in this paper are tested and tuned on an isolated word recognition system (due to the large number of performed experiments and the computation required by some of them). The vocabulary is made up of 16 words, the 10 Spanish digits and six keywords, uttered three times by 20 male and 20 female speakers. The average SNR measured over this database is 24 dB. The speakers are separated in five disjoint groups containing utterances from eight different speakers (four male, four female), to be utilized for test (the rest for training). Thus, each experiment is composed of five different speaker-independent experiments, whose error results are averaged. This procedure is similar to the well-known *leaving-one-out* technique [9] for error probability estimation. Feature vectors incorporate liftered cepstrum, delta cepstrum and delta energy (appropriately weighted), and are compared with an euclidean distance measure.

For comparison of MVQ with DHMM models in the recognition stage, it must be taken into account that with a 16-word vocabulary, the use of 16 N-center codebooks in a MVQ system is computationally equivalent to the use of a single $(16 \times \mathrm{N})$-center codebook in a DHMM system. However, for the training stage the MVQ procedure is always less time-consuming due to the exponential complexity (with the codebook size) of the LBG algorithm. Besides, the MVQ models are always simpler than SCHMM's in both recognition and training [3]. Fig. 1 shows that MVQ modeling clearly outperforms DHMM modeling with the same computational cost in recognition (for eight or more centers per codebook). Besides, MVQ models can achieve similar or even better results than SCHMM's with a meaningful computational saving. SCHMM's have been designed using Gaussian multivariate *pdf*s with diagonal covariance matrices [2].

III. MMI Estimation of DHMM and MVQ Models

It can be proved that the ML estimation method is optimal under certain assumptions such as the true model is known or the training data is large enough [10]. Besides, there is a well-known algorithm, called Baum–Welch procedure, that provides a straightforward way for the ML estimation of Markov models. These reasons explain the success of this type of estimation for HMM models. However, the required optimality assumptions are basically false and the Baum–Welch method is not specifically designed for error minimization. On the other hand, the MMI estimation method does not need any of the ML optimality assumptions and is related to error rate

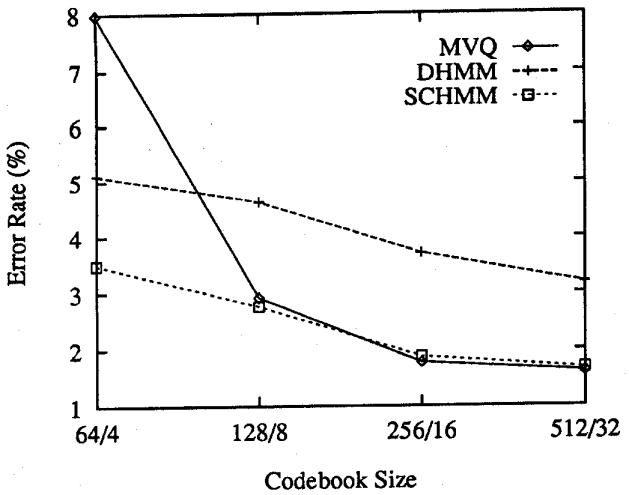


Fig. 1. Error rate versus number of VQ centers for MVQ, DHMM, and SCHMM.

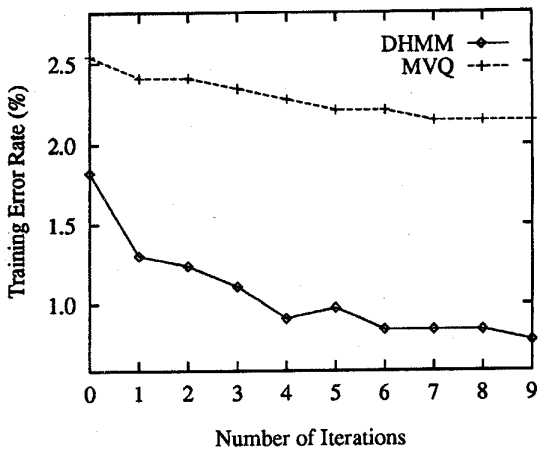


Fig. 2. Training Error Rate evolution for a MMI estimation of matrices Π, A, and B using DHMM models (64 centers) and MVQ (eight centers) with LBG codebooks.

minimization [11]. The criterion function to be maximized in a MMI estimation for a training symbol utterance O and DHMM modeling is (obviating language modeling considerations) given by

$$I_m(O,\Lambda) = \log P_m(O,\Lambda) \tag{6}$$

where $\Lambda = \{\lambda_1, \cdots, \lambda_L\}$ ($\lambda_s = \{A_s, B_s, \Pi_s\}$) is the set of acoustic models, m is the index corresponding to the correct class model of the input sequence X, and

$$P_s(O,\Lambda) = \frac{P(O|\lambda_s)}{\sum_{l=1}^{L} P(O|\lambda_l)}, \qquad s = 1, \cdots, L. \tag{7}$$

Obviously, $s = m$ in (6). The maximization over the parameter set Λ is carried out in this work by means of a standard gradient descent (see reference [11] for more details). Table I shows the recognition error rates using ML and MMI estimations in the DHMM-based system for several codebook sizes. There are no clear improvements due to MMI estimation on the tested task.

When using MVQ models, it must be taken into account that the conditional probabilities P_s include the quantization scores, that is, $P_s = P_s(X,\Lambda)$ [with the same expression as in (7)]. The derivatives for the (A, B, Π) matrices are exactly the same as for DHMM's. The main difference, related to discrete modeling, is that the MMI estimation must be extended to the VQ codebooks. Thus, two new derivatives must be included in the gradient

$$\frac{\partial I_m}{\partial \boldsymbol{y}_j^s} = (\delta_{m,s} - P_s) \sum_{t=1}^{T} \frac{\boldsymbol{x}_t - \boldsymbol{y}_j^s}{\sigma_\lambda^2} \delta_{o_t, v_j} \tag{8a}$$

$$\frac{\partial I_m}{\partial \sigma_{\lambda_s}^2} = (\delta_{m,s} - P_s) \frac{Tp}{2\sigma_{\lambda_s}^2} \left[\frac{D_{\lambda_s}(X)}{\bar{D}_{\lambda_s}} - 1 \right] \tag{8b}$$

where s and j indicate the model and the center under consideration, respectively, and δ_{o_t, v_j} denotes the Kronecker delta function between the symbol o_t corresponding to the nearest neighbor center to $\boldsymbol{x}_t$ and the symbol v_j corresponding to the considered center $\boldsymbol{y}_j^s$.

IV. Discriminative Codebook Design in MVQ Models By MMI

The MMI estimation of MVQ models does not generate two independent training processes like the ML estimation [for the VQ codebook and the (A, B, Π) matrices, respectively], since the P_s probability includes both quantization and generation probabilities.

In spite of the fact that the MMI estimation of MVQ models is not fully decoupled, there are several reasons that support the idea of obtaining a discriminative VQ training independent from the discrete HMM training. These reasons are:

1) The results of MMI estimation in DHMM's do not show clear improvements (in the reference system) related to a ML estimation, and, in fact, the MVQ models are, removing the VQ part, discrete HMM models.
2) The MMI estimation of the discrete HMM part in a MVQ system would increase the training complexity.
3) The codebook sizes used in MVQ modeling are meaningfully smaller than those of DHMM or SCHMM modeling. Thus, the recognition with only generation probabilities is not accurate, and error retrieval is not easy. Fig. 2 shows how training error rate evolves when (A, B, Π) matrices are obtained by a MMI estimation for DHMM and MVQ models (LBG-trained codebooks with 64 and eight centers, respectively, are used). It can be observed that the error reduction is much more pronounced for DHMM's, and negligible for MVQHMM's.

The above items point out that only a MMI codebook design could be profitable for error rate reduction, using equation (8a). Parameters σ_λ^2 are still estimated as average distortions and are not updated by (8b) (in the next sections we will go back to this point). In order to obtain a VQ design method totally independent from the discrete HMM training, the P_s probabilities must include only quantization probabilities

$$P_s(X,\Theta) = \frac{P(X|O_s,\lambda_s)}{\sum_{l=1}^{L} P(X|O_l,\lambda_l)} = \frac{P(X|\theta_s)}{\sum_{l=1}^{L} P(X|\theta_l)} \tag{9}$$

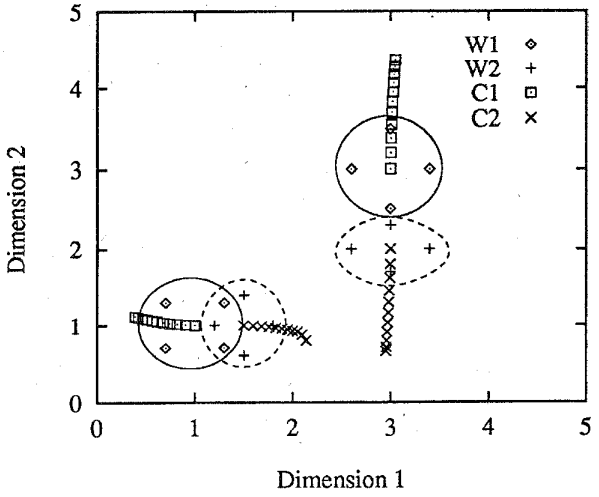


Fig. 3. MVQ codebook evolution by means of MMI-MVQ design.

where $\Theta = \{\theta_1, \cdots, \theta_L\} \subset \Lambda$ only contains the VQ parameters (center vectors and average distortions). We shall refer to this codebook implementation as MMI-MVQ design.

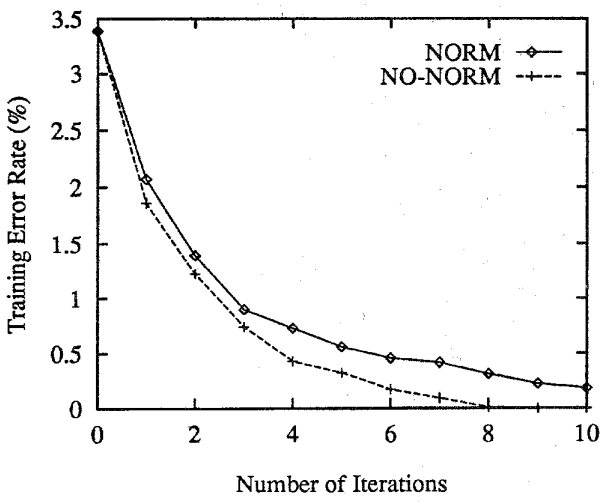
The MVQ system training can be summarized in the three following steps:

1) Construction of one codebook per recognition unit using the LBG algorithm.
2) Reestimation of codebook centers using derivative (8a) [with P_s corresponding to (9)] for iterative updating. Parameters σ_λ^2 maintain the average distortion sense.
3) Estimation of the discrete HMM part (matrices Π, A, and B) using the Baum–Welch algorithm.

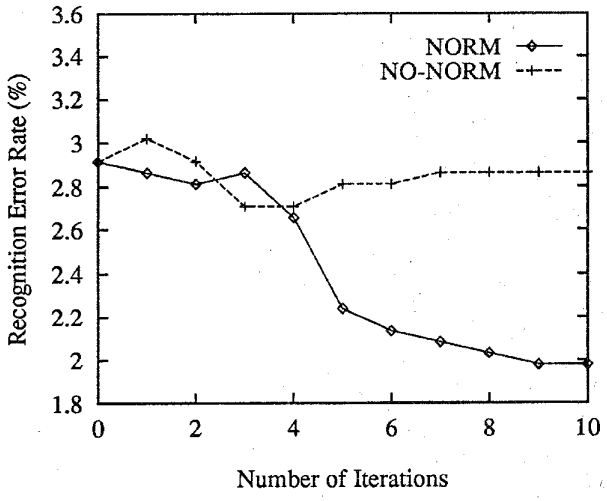
To illustrate the codebook evolution (step 2), let us consider the following example depicted in Fig. 3: we have a set of training sequences belonging to two different classes W1 (framed by solid lines) and W2 (framed by dotted lines). A LBG two-center codebook is constructed for each class (C1 and C2). This recognition scheme presents one error using probabilities $P(X|\theta)$ for decision. We apply 10 iterations using a gradient descent and equation (8a) for center updating. Fig. 3 shows the center trajectories. After five iterations, the classification error disappears. The main effect of the MMI-MVQ design is to move away the VQ centers from incorrect vectors, maintaining a correct classification. It is interesting to observe the great difference between the original LBG codebooks and the error-free ones.

V. Applying the MMI-MVQ Design to Speech Recognition

In the expression (8a), the corrections are weighted by the factor $(\delta_{m,s} - P_s)$, where P_s is given by (9). According to (4), the quantization probabilities $P(X|\theta_l)$, and therefore P_s, strongly depend on the duration T. Since $P(X|\theta_l)$ is obtained as the product of T Gaussian densities, the Tth root provides a time normalization that removes such dependence from $P(X|\theta_l)$, and, therefore, from P_s and $(\delta_{m,s} - P_s)$. Probabilities



(a)



(b)

Fig. 4. Error rate evolution with and without temporal normalization. Effect of σ_λ^2 estimation. (a) Training error rate evolution. (b) Recognition error rate evolution.

P_s are expressed now as

$$P_s(X, \Theta) = \frac{P(X|\theta_s)^{\beta/T}}{\sum_{l=1}^{L} P(X|\theta_l)^{\beta/T}} \tag{10}$$

where $\beta = 1$ in the following discussion.

In addition to the time equalization effect of the normalization, there is another issue that is addressed next: let us assume that no Tth root is applied and $T \to \infty$. Then, we have $P_s \to 1$ if $P(X|\theta_s)$ is the maximum probability, and $P_s \to 0$ otherwise. For such limit, it can be easily verified that $|\delta_{m,s} - P_s|$ coincides with the empirical error, that is, $|\delta_{m,s} - P_s|$ tends to zero in the case of a correct classification, and to one otherwise. We can conclude that the main effect of the time normalization is the smoothing of $|\delta_{m,s} - P_s|$, increasing its value for less probable models. In fact, the

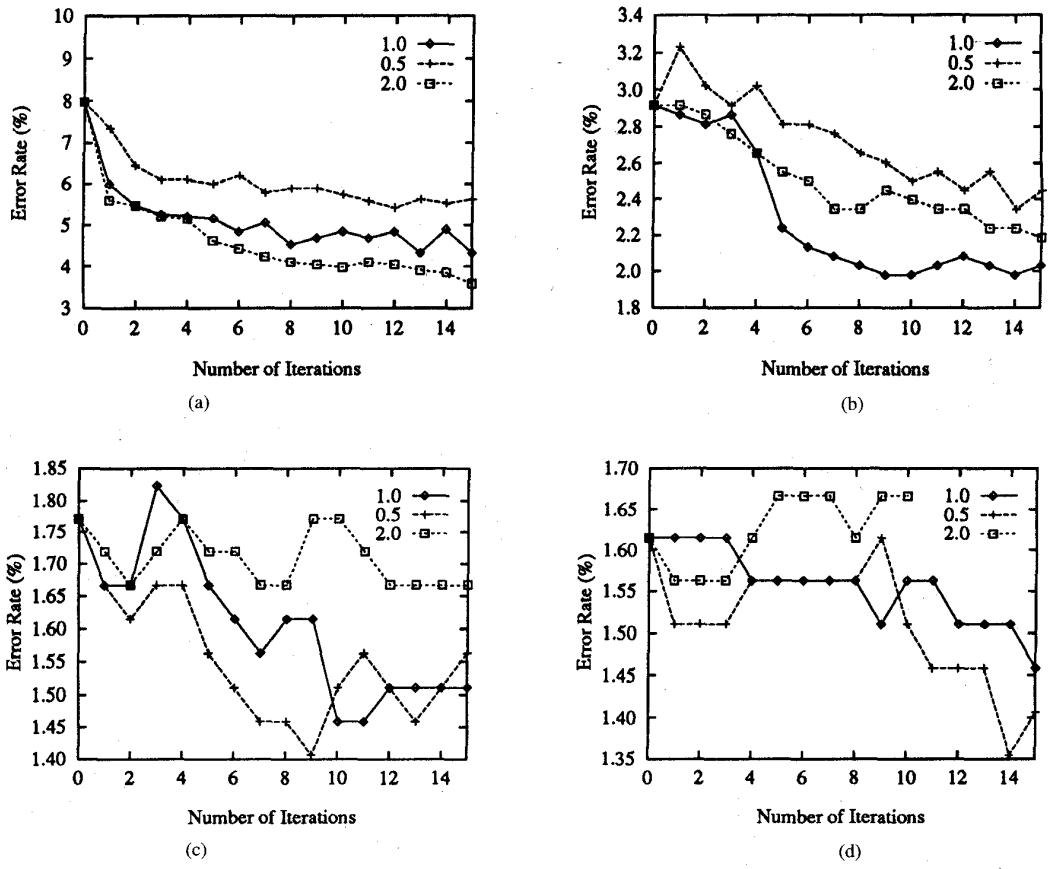


Fig. 5. Error rate evolution for 4, 8, 16, and 32 centers and $\beta = 0.5, 1.0$, and 2.0 versus the number of iterations (MMI-MVQ method). (a) 4 centers. (b) 8 centers. (c) 16 centers. (d) 32 centers.

smoothing of the empirical error function is necessary to obtain analytically tractable discriminative estimation methods, since the empirical error is a nondifferentiable function.

The effect of the time normalization over probabilities P_s is shown in Fig. 4, where an eight-center codebook size with and without time normalization (10 iterations) is used. The training error rate is obtained using only quantization probabilities, and the recognition one from the whole system (including generation probabilities). No noticeable differences are observed in the training stage, although the plot corresponding to no normalization (NO-NORM) converges slightly faster than the plot corresponding to the use of temporal normalization (NORM). This is a logical behavior since NO-NORM implies a more direct attack to the empirical training error rate. However, the NORM plot shows that the temporal normalization method is much more powerful in testing. This result indicates that not only the errors are important to perform corrections, but also that near-misses can be even more effective to correct possible recognition errors. Both experiments, NORM and NO-NORM, use the updating (8a). It must be observed that a full time normalization should include a division of $(\delta_{m,s} - P_s)$ by duration T in (8a). Thus, although all the training utterances have the same importance for the calculation of P_s in experiment NORM, each feature vector contributes to the gradient in the same degree, without considering whether it belongs to a longer or a shorter sequence. A full time normalization is also depicted in Fig. 4 (experiment TOT-NORM). The error rates provided by TOT-NORM in test are slightly worse that those form NORM. The next experiments are performed with the normalization applied in NORM.

A. Estimation of σ_λ^2

In the proposed MMI-MVQ codebook design, σ_λ^2 keeps the meaning of average distortion, instead of being reestimated by means of (8b). In order to justify this point, another experience (labeled as VAR) has been plotted in Fig. 4, similar to NORM but including the reestimation of σ_λ^2 using (8b). This reestimation makes training convergence slightly faster than NORM. This could be expected, since there is now one more parameter to be discriminatively trained. In the case of recognition experiences, the behavior is the same as in training for the first iterations. However, from the fifth iteration on, the previous trend is inverted, and VAR works worse. The explanation for this behavior is that σ_λ^2 is a normalization factor in score (4). If this parameter is discriminatively trained, then the composition of quantization and generation probabilities becomes suboptimal (as we pointed out in Section II-A).

B. MMI-MVQ Codebook Design Performance

In the expression of P_s proposed in (10) we introduced a factor β. Its role is to control the approximation of $|\delta_{m,s} -$

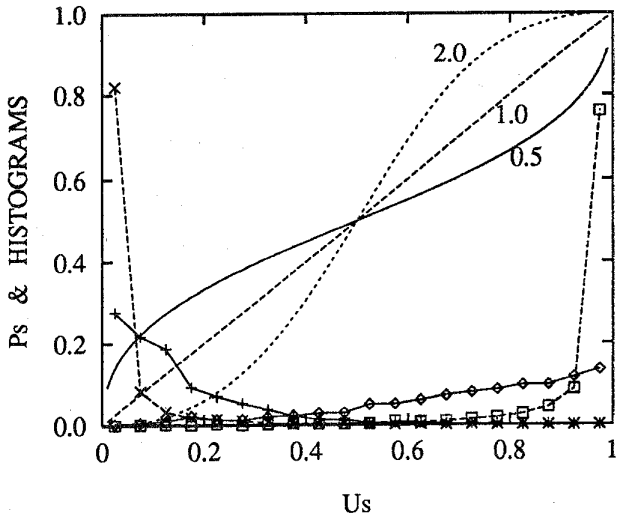


Fig. 6. Function $P_s(u_s)$ for $\beta = 0.5, 1$, and 2 and normalized histograms of variable u_s for four (solid lines) and 16 (dotted lines) centers per codebook and for correct ($\diamond$ four centers, $\square$ 16 centers) and incorrect ($+$ four centers, $\times$ 16 centers) utterances.

$P_s|$ to the empirical error. Note that with all the applied modifications the resultant MMI-MVQ codebook design is not an MMI estimation. However, we will keep the MMI-MVQ abbreviation for the described procedure.

Fig. 5 shows the recognition error rate evolution plots as a function of the number of iterations of the MMI-MVQ procedure for four, eight, 16, and 32 center codebooks, and for $\beta = 0.5, 1.0,$ and 2.0 in (10). Iteration zero denotes the original MVQ system error rate. It is observed that the value of factor β must be decreased when the codebook size is increased. Thus, $\beta = 2$ is appropriate for four centers, $\beta = 1$ for eight, and $\beta = 0.5$ for 16 and 32 centers. That is, $|\delta_{m,s} - P_s|$ must be smoothed when the codebook size is increased.

An appropriate value of β can be graphically determined from the histograms of the probabilities $p_s = P(X|\theta_s)^{1/T}$. Fig. 6 shows the normalized histograms of the average values of $u_s = p_s / \sum_l p_l$ for the correct and the best incorrect class for four and 16 centers per codebook (correct utterances are concentrated near $u_s = 1$ and incorrect ones are near $u_s = 0$). These plots are superposed to those of the function

$$P_s(u_s) = \frac{u_s^\beta}{u_s^\beta - (1 - u_s)^\beta} \tag{11}$$

for $\beta = 0.5, 1.0,$ and 2.0, where $(\delta_{m,s} - P_s(u_s))$ would be the correction factor in (8a) for an input sequence in a two classes problem (we are only considering the correct and best incorrect classes). For $\beta = 2$, it is observed that the most contributing utterances to the gradient are those placed in the center of the plot ($u_s = 0.3$ to 0.7). In fact, most of the error minimization methods tend to take training vectors from the boundaries between classes, that is, where an error can take place. That is not a problem for the case of four centers, since there are enough training utterances in that area, but when 16 centers are used this area is almost empty, and insufficient training problems can arise. Thus, the value of β must be decreased to include utterances from the side areas.

TABLE I
RECOGNITION ERROR RATE FOR ML AND MMI ESTIMATIONS IN DHMM MODELING

Codebook Size	Error (%) DHMM-ML	Error (%) DHMM-MMI
64	5.10	5.41
128	4.63	4.58
256	3.69	3.48
512	3.17	3.22

TABLE II
ERROR RATE FOR DHMM, SCHMM AND MVQ MODELS WITH ML ESTIMATION AND MVQ WITH MMI-MVQ CODEBOOK DESIGN

# Centers	DHMM	SCHMM	MVQ	MMI-MVQ
64/4	5.10	3.48	7.96	3.58
128/8	4.63	2.76	2.91	1.97
256/16	3.69	1.87	1.77	1.40
512/32	3.17	1.66	1.61	1.35

VI. CONCLUSIONS

We have proposed in this paper a new variant of the MMI estimation method, thought for a proper test error rate reduction. The results show that the discriminative codebook design obtained from this MMI-based method can be very efficient for the pursued goal when using MVQ models. Table II summarizes the obtained error rates and compares them with those from ML estimations of DHMM, SCHMM and MVQ models. It can be observed that the use of discriminative codebooks can approximate the performance of a MVQ system to that of a SCHMM system for four centers per codebook, and provides the best results for eight, 16, and 32 centers. It must be taken into account that a MVQ system involves the same computation as a DHMM-based system in recognition. Thus, the computational saving of MVQ's in relation to SCHMM's is the same as that obtained with DHMM's. The proposed MMI-MVQ codebook design is more effective for small codebook sizes. This result agrees with the statement that establishes the suitability of the MMI estimation in the case that the true model is not known [11] (the model is more incorrect for small codebook sizes). The similarity of the results for 16 and 32 centers can be exploited to reduce the computational complexity of a high performance system. Furthermore, even with only eight centers, an error rate value below 2% can be reached.

Although we have utilized here a MMI-based approach, the proposed method is opened to other discriminative schemes yielding center updating formulas of the form

$$\hat{\boldsymbol{y}}_k = \boldsymbol{y}_k + \eta f(X, \Lambda) \sum_{t=1}^{T} (\boldsymbol{x}_t - \boldsymbol{y}_k) \delta_{o_t, v_k} \tag{12}$$

where function $f(X, \Lambda)$ is defined by the utilized estimation method, and η is a small real positive number. For example, a minimum classification error (MCE) estimation [12] can be applied (for details of the implementation see reference [13]), obtaining similar results, although it involves more

experimental parameters to be fitted. It must be pointed out the similarity between the proposed codebook designs, summarized in (12), and the LVQ techniques proposed by Kohonen [7]. The main difference is that (12) needs no time alignment as other LVQ-based applications to speech recognition [8]. In our system, the time alignment information is processed by the discrete HMM part of the MVQ model.

Finally, we also think that the proposed techniques could be very useful in a continuous speech recognition task, because of the high degree of confusion among subword units.

REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, Feb. 1989.

[2] X. Huang and M. Jack, "Unified techniques for vector quantisation and hidden Markov modeling using semicontinuous models," in *Proc. ICASSP-89*, Glasgow, Scotland, May 1989, pp. 639–642.

[3] J. Segura, A. Rubio, A. Peinado, P. García, and R. Román, "Multiple VQ hidden Markov modeling for speech recognition," *Speech Commun.*, vol. 14, pp. 163–170, Apr. 1994.

[4] A. Peinado, J. Segura, A. Rubio, and M. Benítez, "Using multiple vector quantization and semicontinuous hidden Markov models for speech recognition," in *Proc. ICASSP-94*, 1994.

[5] A. Peinado, J. Segura, A. Rubio, V. Sanchez, and P. Garcia, "On the use of multiple vector quantization for semicontinuous-HMM speech recognition," *IEE Proc. Vision Image and Signal Processing*, Dec. 1994, vol. 141, pp. 391–396.

[6] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov models parameters for speech recognition," in *Proc. ICASSP-86*, 1986, Tokyo, pp. 49–52.

[7] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, pp. 1464–1480, Sept. 1990.

[8] S. Katagiri and C.-H. Lee, "A new hybrid algorithm for speech recognition based on HMM segmentation and learning vector quantization," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 421–430, Oct. 1993.

[9] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[10] A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," *IEEE Trans. Acoust. Signal Processing*, vol. ASSP-31, pp. 814–817, Aug. 1983.

[11] P. Brown, "The Acoustic-Modeling Problem in Automatic Speech Recognition," Ph.D. dissertation, Carnegie Mellon Univ., 1987.

[12] B. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification," *IEEE Trans. Signal Processing*, vol. 40, pp. 3043–3054, Dec. 1992.

[13] A. M. Peinado, A. J. Rubio, J. C. Segura, V. E. Sánchez, and J. E. Díaz, "MCE estimation of VQ parameters for MVQHMM speech recognition," in *Proc. EUROSPEECH-95*, Sept. 1995, vol. 1, pp. 533–536.

Antonio M. Peinado (M'95) was born in Guadix, Granada, Spain, in 1963. He received the Licenciate, Grado, and Doctor degrees in physics from the University of Granada, Spain, in 1987, 1989 and 1994, respectively. He developed his Ph.D. thesis on HMM parameter estimation.

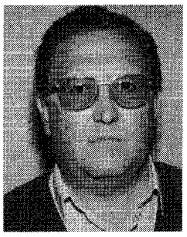
Since 1988, he has been working with the Research Group on Signal Processing and Communications of the Department of Electronics and Computer Technology of the University of Granada on several topics related to speech recognition and coding. In 1989, he was a Consultant in the Speech Research Department, AT&T Bell Labs, Murray Hill 07974 NJ, USA. Currently, he is Associate Professor at the same department in Granada and his research interests are in discriminative training, robust speech recognition and speech coding.

Dr. Peinado is a member of ESCA and AERFAI.

José C. Segura (M'95) was born in Alicante, Spain, in 1961. He received the Licenciado degree in 1984 and the Ph.D. degree in 1991 from the University of Granada, Spain. He developed his Ph.D. thesis on MVQHMM modelling.

From 1987 to 1991 he was working as an Assistant Professor, and since 1991 as an Associate Professor at the Department of Electronics and Computer Technology of Granada University. Since 1984, he has been working with the Research Group on Signal Processing and Communications of the Department of Electronics and Computer Technology of the University of Granada, on several topics related to speech recognition and coding. His curient lesealch interests are in speech recognition and coding.

Dr. Segura is member of AERFAI.



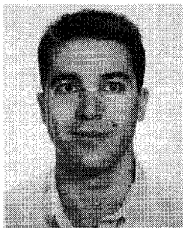
Antonio J. Rubio studied physics at the University of Seville, with a specialty in electronics in 1972. He received the Ph.D. degree in 1978.

In 1972 he joined the University of Granada as an Assistant Professor. Since then he has been dedicated to the speech recognition and coding research. He spent a one-year period working as a consultant in the Speech Research Department, AT&T Bell Labs, Murray Hill, NJ. Currently, he is a Profesor Titular in the Department of Electronics and Computer Technology of the University of Granada, Spain. Dr. Rubio organized the 1993 NATO-ASI on "New Advances and Trends in Speech Recognition and Coding", held at Bubion, Granada, Spain.

Dr. Rubio is a member of AEIA, EURASIP, and ESCA Association.

Pedro García received the Licenciate degree in physics in 1988.

In 1989 and 1990 he was granted by Fujitsii Espana and IBM, respectively. Since 1990 he has been working as an Assistant Professor at the Department of Electronics and Computer Technology of the University of Granada, where is currently developing his researching on speech recognition and coding.



José L. Pérez (M'95) was born in Archidona, Malaga, Spain. He received the Licenciado en Ciencias Fisicas degree with a speciality in electronics, in 1987, from the University of Granada, Spain.

In 1989 he was granted by the Government of Spain and joined the Department of Electronics and Computer Technology where he is now working as an Assistant Professor. From September 1993 to February 1994 he was visiting with the System Research Center, University of Maryland, USA. His research interests include speech coding and recognition.

Mr. Perez is member of the the European Speech Communication Association.