MCE ESTIMATION OF VQ PARAMETERS FOR MVQHMM SPEECH RECOGNITION

Antonio M. Peinado, Antonio J. Rubio, José C. Segura, Victoria Sánchez, Jesús E. Díaz
Research Group on Signal Processing and Communications
Dpto. de Electrónica y Tecnología de Computadores
Universidad de Granada - 18071 Granada (SPAIN)

Abstract

Recent research on Multiple Vector Quantization (MVQ) has shown the suitability of such technique to Speech Recognition. Basically, MVQ proposes the use of one separated VQ codebook for each recognition unit. Thus, a MVQ HMM 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 that provides meaningful error reductions and is performed independently from the estimation of the discrete HMM part of the MVQ model. This codebook design uses a Minimum Classification Error scheme and have certain similarities with the LVQ techniques proposed by Kohonen, but overcoming any time alignment requisite.

1 Introduction

During the last years, Hidden Markov Models (HMM) have been successfully applied to acoustic modeling for speech recognition. Two main variations of HMMs have been widely used: discrete HMMs (DHMM) and continuous HMMs (CHMM). The main problem of DHMMs is the loss of information about the input signal during the VQ process. CHMMs avoid this problem using probability density functions (*pdfs*). 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. These are strong restrictions that may make advisable the use of the DHMM approach.

In order to avoid such problems of continuous modeling, Huang *et al* [1] propose the use of *Semicontinuous* HMM models (SCHMM). Our research group has also proposed a new approach based on the use of *Multiple Vector Quantization* (MVQ) for HMMs, that has been called MVQHMM or, simply, MVQ modeling [2]. A MVQHMM model 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. To do this, each MVQ model uses its own VQ codebook to evaluate the average distortion of the input utterance (the sequential information is evaluated by the discrete HMM part). This type of modeling can be generalized, in the same way as SCHMMs generalize DHMMs, using several quantization candidates. This way, the SCMVQ-HMM modeling is obtained [3]. The training computational complexity of MVQs and SCMVQs is lower than that of DHMMs and SCHMMs, respectively. Besides, when the number of centroids per codebook is high enough, MVQ and SCMVQ models have been shown to be more accurate than DHMMs or SCHMMs [3]. However, for a very small number of centroids the MVQ approach is not able to correctly model the acoustic variety of the events being modeled, and the performance degrades. In a previous work [4], we have shown that the performance of a MVQ system with few centers per codebook can be greatly improved applying a MMI estimation to the estimation of the codebook parameters. This VQ estimation can be performed independently from the discrete HMM parameters, which are estimated via Baum-Welch, since most of the discriminative information is in the VQ part.

Since the final goal of speech recognition is the minimization of the error rate, and taking into account the above discussion, a great benefit could be obtained from the minimization of the error rate obtained recognizing only with VQ distortions using MVQ models. In this work we propose a new method of estimating the VQ parameters (of MVQHMM models) based on the minimization of the classification error using the MCE (Minimum Classification Error) criterion proposed by Juang et al [5]. The resulting estimation method have certain similarities with the LVQ techniques proposed by Kohonen, but overcoming the time alignment requisite of LVQ.

The rest of the paper is organized as follows. In the next section, we briefly review the fundamentals of MVQ HMM modeling and set the experimental conditions of the work. In section 3 we propose the MCE-based VQ design procedure. Section 4 is devoted to our experimental results. The paper finishes with the conclusions of the work.

2 MVQ Modeling

A continuous HMM model uses a mixture of *pdfs* to model the output probabilities in the following form,

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

where each $P(\mathbf{x}|v_k, s_i, \lambda)$ is a log-concave or elliptically symmetric density [6] (gaussian along this work) with mean vector μ_k and covariance matrix Σ_k, and $\mathbf{x}$ is the input vector. Each *pdf* is labelled by one index v_k that varies in the set $V(s_i, \lambda)$ defined for state s_i in model λ. Factors $P(v_k|s_i, \lambda)$ are the mixture coefficients (their sum extended over $V(s_i, \lambda)$ must be one).

The simplification of expression (1) leads to different HMM approaches. For example, doing $V(s_i, \lambda) = V \quad (\forall s_i, \lambda)$, we obtain a SCHMM. Furthermore, if we assume non-overlapped *pdfs*, a SCHMM becomes a DHMM. A third simplification can

ESCA. EUROSPEECH'95. 4th European Conference on Speech Communication and Technology. Madrid, September 1995. ISSN 1018-4074

be derived by assuming a different set of *pdfs* $V(\lambda)$ (shared by all the states of the model) for each model λ, and considering non-overlapped densities. Thus,

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

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

We have just defined a MVQ HMM model. It can be proved that, for an input sequence $X = \mathbf{x}_1 \cdots \mathbf{x}_T$, the density $P(X|\lambda)$ can be expressed as,

$$P(X|\lambda) = P(X|O,\lambda)P(O|\lambda) \quad (3)$$

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 MVQ model parameter set can be decomposed as $\lambda = (\theta, \phi)$, where θ represents the parameter set of densities $P(\mathbf{x}|v_j,\lambda)$ and ϕ is the parameter set of the discrete HMM model, it can be proved that the ML estimation of λ is obtained from a separated ML estimation of θ and ϕ [3]. The first parameter set (mean vectors and covariance matrices) can be obtained from a VQ codebook $\{\mathbf{y}_j, j = 1,\ldots,M\}$ (trained using the LBG algorithm, for example). The second one is estimated applying the Baum-Welch algorithm, as for DHMM models.

A convenient form for the densities $P(\mathbf{x}|o,\lambda)$ 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 [3]. Thus, the quantization probability for an input sequence X is written as,

$$P(X|O,\lambda) = \prod_{t=1}^{T} P(\mathbf{x}_t|o_t,\lambda) = \quad (4)$$

$$= \prod_{t=1}^{T} (2\pi\sigma_\lambda^2)^{-p/2} \exp\left\{-\frac{1}{2\sigma_\lambda^2}\|\mathbf{x}_t - \mathbf{y}_t\|^2\right\}$$

where p is the number of features and $\mathbf{y}_t$ is the nearest center to the input vector $\mathbf{x}_t$.

For sequence evaluation, the following score is utilized,

$$\log P(X|\lambda) = \alpha \log P(X|O,\lambda) + \log P(O|\lambda) \quad (5)$$

where α is a tuning factor to optimize the composition of quantization and generation probabilities ($\alpha = 0.5$ in this work) [3].

It can be observed in (4) that the log-probability of quantization has a linear relation with the average distortion of the input sequence X in the codebook θ. In fact, this is the original motivation of the MVQHMM modeling: it incorporates to the decision criterion the distortion information generated in the VQ process and wasted when using DHMM models.

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 realized experiments and the computation required by some of them). The vocabulary is made up of 16 words, the 10 Spanish digits and 6 keywords, uttered 3 times by 20 male and 20 female speakers, what means 1920 different signals in the database. The average SNR measured over this database is 24 dB, that corresponds to the environment of a work room with computer noise. The speakers are separated in 5 disjoint groups containing utterances from 8 different speakers (4 male, 4 female), to be utilized for test (the rest for training). The result is the realization of 5 different speaker-independent

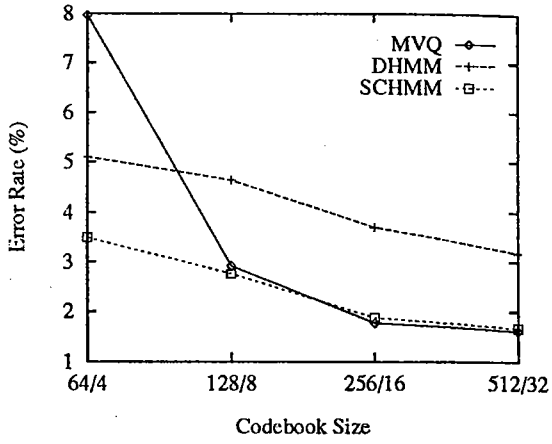

Figure 1: *Error Rate vs. number of VQ centers for MVQ, DHMM and SCHMM.*

experiments, whose error results are finally averaged. Feature vectors incorporate liftered cepstrum, delta cepstrum and delta energy, and are compared with an euclidean weighted 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 equivalent to the use of a single (16*N)-center codebook in a DHMM system, in order to obtain the same computational complexity. However, for the training stage, the MVQ procedure is always less time-consuming since the exponential complexity of the LBG algorithm grows exponentially with the codebook size. Besides, the MVQ models are always simpler than SCHMMs in both recognition and training.

Figure 1 shows that MVQ modeling clearly outperforms DHMM modeling (when more than 4 centers per codebook are used) with the same computational cost in recognition. Besides, MVQ models can achieve similar (8 centers) or even better results than SCHMMs with a meaningful computational saving. SCHMMs have been designed using gaussian multivariate *pdfs* with diagonal covariance matrices [1]. The objective of this work is to improve the performance of the MVQ for 4 to 8 centers per codebook, that is, maintaining a low computational complexity.

3 MCE estimation of MVQ parameters

The MCE estimation is based on the minimization of a cost function. Let us suppose a set of classes $W = \{W_1,\ldots,W_L\}$. For an input sequence X belonging to the mth class, the cost function has a sigmoid shape,

$$l_m(d_m) = \frac{1}{1 + e^{-\alpha d_m}} \quad (6)$$

where,

$$d_k(X,\Lambda) = -g_k(X,\lambda_k) + \quad (7)$$

$$\log\left[\frac{1}{L-1}\sum_{j\neq k} e^{\beta g_j(X,\lambda_j)}\right]^{1/\beta}$$

$$g_k(X,\lambda_k) = \log P(X|O,\lambda_k) \quad (8)$$

The classifier parameter set is $\Lambda = \{\lambda_1,\ldots,\lambda_L\}$. When the error measure $d_m \gg 0$, then the input sequence is incorrectly

classified, and $l_m = 1$. On the other hand, when $d_m \ll 0$, $l_m = 0$. For a multiple training sequence, we can sum the particular cost functions (of each training sequence) to obtain an Empirical Average Cost (EAC), that is an approximation to the total number of training errors. It can be also derived an expression for the Expected Cost (EC). The MCE estimation is based on the minimization of one of these total costs. The EAC can be minimized by means of a common gradient descent (GD), meanwhile the EC needs a Generalized Probability Descent (GPD) [5], quite similar to a GD but using the training data "on line".

In any case (using GD or GPD), the codebook centers are be iteratively updated using the following gradient,

$$\frac{\partial l_m(X,\Lambda)}{\partial \mathbf{y}_j^s} = F(X,\Lambda)\frac{\sum_{t=1}^{T}(\mathbf{x}_t - \mathbf{y}_j^s)\,\delta_{o_t,v_j}}{\sigma_{\lambda_s}^2} \qquad (9)$$

where $\mathbf{y}_j^s$ is the jth center of the codebook of model λ_s, δ_{o_t,v_j} is a Kronecker delta equal to 1 when the nearest center to $\mathbf{x}_t$ is $\mathbf{y}_j^s$ (0 otherwise), and,

$$F(X,\Lambda) = \begin{cases} -\nu_m(X,\Lambda) & s = m \\ \nu_m(X,\Lambda)\phi_{sm}(X,\Lambda) & s \neq m \end{cases} \qquad (10)$$

$$\nu_m(X,\Lambda) = \alpha l_m(X,\Lambda)[1 - l_m(X,\Lambda)] \qquad (11)$$

$$\phi_{sm}(X,\Lambda) = \frac{e^{\beta g_s(X,\lambda_s)}}{\sum_{l\neq m} e^{\beta g_l(X,\lambda_l)}} \qquad (12)$$

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 (9) (for both, GD or GPD). 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.

4 MCE codebook design performance

In this section, we present some results obtained applying MCE to the codebook parameter estimation for MVQ models. As we discussed in the Introduction section, we are specially interested on small codebook sizes. Thus, we will develop experiments for 4 and 8 centers per codebook. We have first considered a GPD implementation. Since the resultant GPD procedure is an "on line" process, the computation of σ_λ^2 as an average distortion (as in the MMI-MVQ method) would be very time-consuming. Thus, as a first approximation, no reestimation will be performed for this parameter (the original LBG value is kept constant). The correction factors ν_m and ϕ_{sm} are obtained from the time-normalized discriminant functions (to remove the influence of sequence length),

$$g_k(X,\lambda_k) = \frac{1}{T}\log P(X|O,\lambda_k) \qquad (13)$$

The results of using $\beta = 2.0, 4.0, 8.0$ (in eqn. (7)) are depicted in figure 2 ($\alpha = 1$ in sigmoid (6)). Each iteration means one presentation of the whole training data. The best results are obtained with $\beta = 4.0, 8.0$. $\beta = 4.0$ will be used from now on.

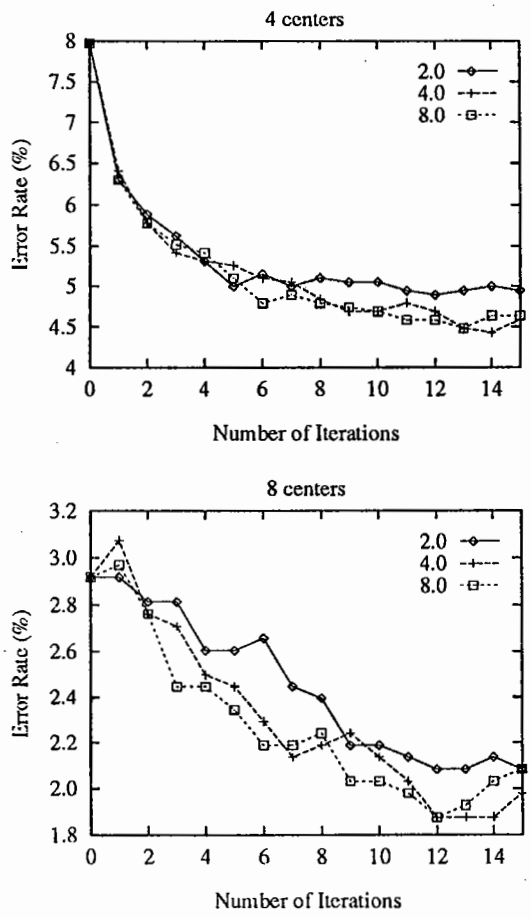

Figure 2: *Error rate evolution for 4 and 8 centers versus the number of iterations for* $\beta = 2.0, 4.0, 8.0$.

Figure 3 shows the previous experiment (for $\beta = 4.0$) (labelled as GPD) along with two new ones. The experiment labelled with GD corresponds to a gradient descent minimizing the EAC (σ_λ^2 is updated in each iteration). Experiment GPDR is the same as GPD, but updating σ_λ^2 at the end of each iteration. As observed, no noticeable differences are detected with respect to the initial GPD design. Two conclusions can be extracted:

1) There are no meaningful differences between the minimization of the empirical average cost (GD) and the expected cost (GPD and GPDR) for codebook design purposes.

2) The MCE method does not produce noticeable deviations in the optimal composition of quantization and generation probabilities (see eqn. (5)), as it is extracted from the fact that GPD, GD and GPDR obtain similar results and estimate σ_λ^2 in different ways.

Also, the influence of the sigmoid transition parameter α must be studied, since it controls the effective number of utterances and their weights for training, as we can see in figure 4, where $\nu_m = \nu_m(d_m)$ (see equation (11)) is depicted. The experiments for that are ongoing at the moment of writing this paper, and the results will be shown at the conference.

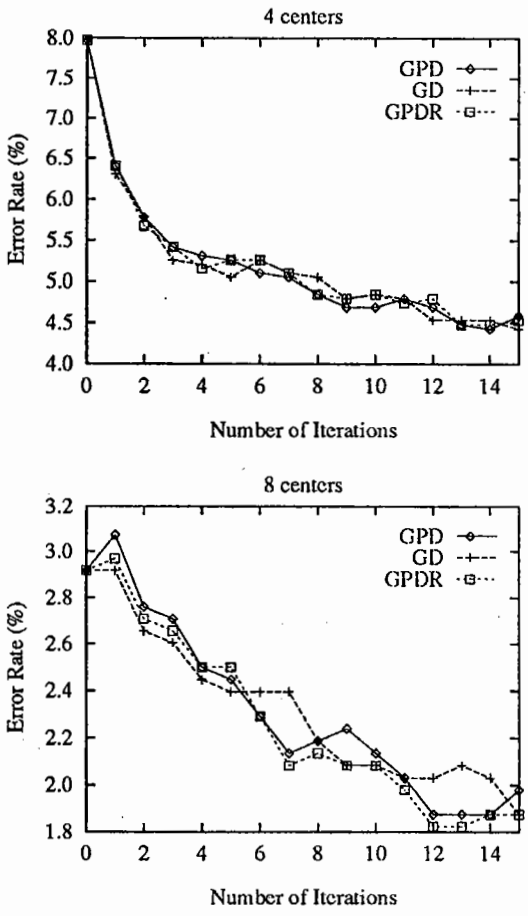

Figure 3: *Error rate evolution for 4 and 8 centers versus the number of iterations for experiments GPD, GD and GPDR.*

5 Conclusions

We have proved that the MCE parameter estimation method suits quite well to the codebook design of MVQ models and have proposed a procedure to train MVQ HMM models using this codebook design. Table 1 shows the best results obtained with the MCE-based VQ design for 4, 8, 16 and 32 centers per codebook compared with those obtained for DHMMs, SCHMMs and basic MVQs. 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 4 centers per codebook, and provides the best results for 8, 16 and 32 centers. 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 8 centers, an error rate value below 2% can be reached. The computational cost of the proposed procedure estimation is not too high, since it is only applied to VQ codebooks, and not to the discrete HMM parts.

References

[1] X. Huang and M. Jack, "Unified Techniques for Vector Quantisation and Hidden Markov Modeling Using Semi-Continuous Models," in *Proc. of ICASSP-89*, (Glasgow (Scotland)), pp. 639–642, May 1989.

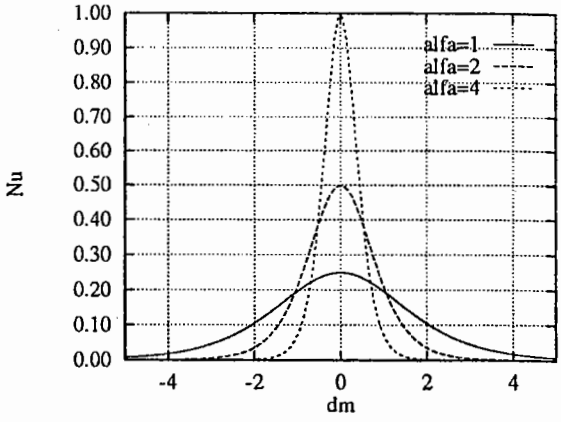

Figure 4: *Factor ν_m as a function of d_m.*

# Centers	DHMM	SCHMM	MVQ	MCE
64/4	5.10	3.48	7.96	4.42
128/8	4.63	2.76	2.91	1.87
256/16	3.69	1.87	1.77	1.45
512/32	3.17	1.66	1.61	1.45

Table 1: *Error rate for DHMM, SCHMM and MVQ models with ML estimation and MVQ with MCE-based codebook design*

[2] J. Segura, A. Rubio, A. Peinado, P. García, and R. Román, "Multiple VQ Hidden Markov Modelling for Speech Recognition," *Speech Communication*, vol. 14, pp. 163–170, April 1994.

[3] A. Peinado, J. Segura, A. Rubio, and M. Benítez, "Using Multiple Vector Quantization and Semicontinuous Hidden Markov Models for Speech Recognition," in *Proc. of ICASSP-94*, 1994.

[4] A. Peinado, J. Segura, A. Rubio, and V. Sánchez, *New Advances and Trends in Speech Recognition and Coding*, ch. A MMI Codebook Design for MVQHMM Speech Recognition. NATO ASI 93 (in press), 1993.

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

[6] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *Proceedings of the IEEE*, vol. 77, pp. 257–285, Feb. 1989.