USING MULTIPLE VECTOR QUANTIZATION AND SEMICONTINUOUS HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION

Antonio M. Peinado *José C. Segura* *Antonio J. Rubio* *María C. Benítez*
Dpto. de Electrónica y Tecno. de Computadores
Universidad de Granada, 18071-GRANADA (SPAIN)

ABSTRACT

Although Continuous HMM (CHMM) technique seems to be the most flexible and complete tool for speech modeling, it is not always used for the implementation of speech recognition systems due to several problems related to training and computational complexity. Besides, it is not clear the superiority of continuous models over other well-known types of HMMs, such as Discrete (DHMM) or Semicontinuous (SCHMM) models, or Multiple Vector Quantization (MVQ) models, a new type of HMM modeling recently introduced by our group. In this paper, we propose a new variant of HMM models, the SCMVQ HMM models (Semicontinuous Multiple Vector Quantization HMM), that uses one VQ codebook per recognition unit and several quantization candidates. Formally, SCMVQ modeling is the closest one to CHMM, although requiring less computation than SCHMMs. Besides, we show that SCMVQs can obtain better recognition results than DHMMs, SCHMMs or MVQs.

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 first ones use nonparametric discrete output probability distributions, due to a previous VQ process. CHMMs use parametric densities to model the output probabilities, on the assumption that the observed signals have been generated by a mixture gaussian process or an autoregressive process [1]. 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 compute. 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 DHMM [2].

In order to avoid such problems of continuous modeling, Huang et al [2] propose the use of *semicontinuous* HMM (SCHMM) models, a hybrid modeling that uses several VQ candidates instead of only the best one, as in DHMMs. Huang has shown that SCHMMs can achieve better results than CHMMs. Besides, our group has recently proposed a new approach based on the use of *Multiple Vector Quantization* for HMMs (MVQ HMM or, simply, MVQ modeling) [3]. With the same amount of computation, the MVQ modeling can clearly outperform DHMMs and achieve similar or better results than SCHMMs (with less computation).

In this paper, we propose a new variant of HMM modeling based on the generalization of MVQ using several candidates in the VQ process, extending the MVQ models from a discrete to a semicontinuous approach, that we will call *SCMVQ* HMM modeling (semicontinuous HMMs with Multiple Vector Quantization).

In the next section, a generalized framework for HMM modeling, from which MVQ modeling can be derived, will be established. In section 3, we will look for a suitable form of the *pdfs* of the MVQ models (to be used in SCMVQs). Also, this models will be compared with DHMM and SCHMM. SCMVQ models will be introduced and compared in section 4. Finally, we will summarize the conclusions of this work.

The different experiences, developed in this work, were made on an isolated word task, with a vocabulary of 16 words (the 10 spanish digits and 6 keywords). The database contains 1920 signals, uttered by 20 female and 20 male speakers (3 repetitions of each word of the vocabulary by each speaker). The data were analyzed using 32 ms frames, overlapped 16 ms. The feature vectors are made up by 14 cepstral and 14 delta cepstral coefficients, plus delta energy, and compared using an euclidean weighted distance measure [4]. The error rate values have been obtained in speaker independent mode, using a leaving-one-out-like technique, with 5 partitions of the database (32 speakers for training and 8 for testing in each partition).

2. GENERALIZED FRAMEWORK

The difference among the different HMM techniques is in the computation of the output probabilities $b_j(\mathbf{x})$ of a vector $\mathbf{x}$ in state s_i, given a model λ. The most general form corresponds to CHMM modeling, where $b_j(\mathbf{x})$ is modeled by a mixture of *pdfs* of the form,

$$b_j(\mathbf{x}) = p(\mathbf{x}|s_i,\lambda) = \sum_{v_k \in V(s_i,\lambda)} p(\mathbf{x}|v_k,s_i,\lambda)p(v_k|s_i,\lambda) \quad (1)$$

*Work supported by CICYT project TIC92-0662.

where $V(s_i, \lambda)$ is the set of *pdfs* of the mixture of state s_i and model λ, and v_k is an index representing *pdf* number k. The probabilities $p(\mathbf{x}|v_k, s_i, \lambda)$ are usually calculated utilizing gaussian or autoregressive processes.

A first simplification can be made forcing all states to share the same set of *pdfs*, $V(s_i, \lambda) = V \; \forall s_i, \lambda$, which leads to the semicontinuous HMM approach. The output probabilities are now computed as,

$$p(\mathbf{x}|s_i, \lambda) = \sum_{v_k \in V} P(\mathbf{x}|v_k) b_i(v_k) \qquad (2)$$

The set V can be obtained from the construction of a VQ codebook, and the sum of equation (2) is usually reduced only to the best set of candidates.

The DHMM approach is easily obtained from SCHMM modeling keeping only the best VQ candidate (the nearest VQ center). In this case, only $P(O|\lambda)$ (probability of generation) is computed for an input sequence $X = \mathbf{x}_1 \cdots \mathbf{x}_T$ (with $O = o_1 \cdots o_T$ as quantized version), since $P(X|O)$ (probability of quantization) does not depend on the model (the global probability $P(X|\lambda)$ is the product of both).

The MVQ HMM modeling is based on the use of one codebook per model, $V(s_i, \lambda) = V(\lambda)$ for all s_i in model λ, with,

$$P(\mathbf{x}|s_i, \lambda) = P(\mathbf{x}|o, \lambda) b_i(o) \qquad (3a)$$

$$o = \max_{v_k \in V(\lambda)}{}^{-1} \left[P(\mathbf{x}|v_k, \lambda) \right] \qquad (3b)$$

In this case, it is also possible a decomposition of the probability of a sequence X,

$$P(X|\lambda) = P(X|O, \lambda) P(O|\lambda) \qquad (4)$$

The probability of generation can be estimated in the same way as for DHMM models (using each model its own VQ codebook). The main difference between DHMM and MVQ models is that the probability of quantization cannot be removed now, since it is different for each model. It can be considered that a MVQ model is made up by a VQ codebook and a discrete HMM. It has been proven that the ML estimation of MVQ models can be performed independently for the VQ parameters (LBG algorithm) and for the discrete HMM parameters (Baum-Welch algorithm) [5].

3. IMPLEMENTATION OF A MVQ-BASED SYSTEM

Each *pdf* $v_k \in V(\lambda)$ used for MVQ modeling is assumed to be a multivariate gaussian density, with a mean vector μ_k (VQ center) and a diagonal covariance matrix $\Sigma_k = \left\{ \sigma^2_{k_i}, i = 1, \ldots, p \right\}$.

We have tested 3 different forms for the covariance matrices in 3 different experiments:

EXP1 Using a different covariance matrix for each *pdf*.

EXP2 Using only one covariance matrix Σ_λ for all the *pdfs* in the codebook of model λ, where each element of the diagonal is the average distortion of the corresponding feature in that codebook.

EXP3 Using only one covariance matrix $\Sigma_\lambda = \sigma^2_\lambda I$ for all the *pdfs* in the codebook of λ, where σ^2_λ is the average distortion per feature in the codebook, and I is the identity matrix.

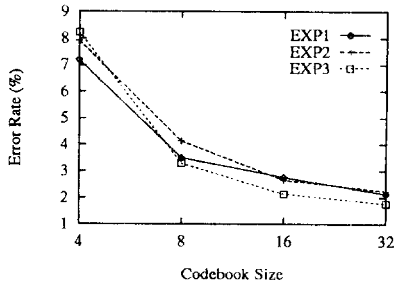

Figure 1. *Error rate vs. the codebook size for EXP1, EXP2 and EXP3.*

Figure 1 shows the results of these 3 experiments, using 4, 8, 16 and 32 centers per codebook. The best results are obtained with EXP3. Two reasons may explain this behavior. First, EXP3 is the only experience for which the probability measure is coherent with the VQ distance used in this work, that is, the nearest center of an input vector represents also the most probable *pdf*. Second, EXP3 uses only one parameter σ^2_λ to represent all the covariance matrices of all the *pdfs* in the codebook, which implies a great reduction in the number of parameters to train, lightening the problem of insufficient training. Each *pdf* can be written (using EXP3) as,

$$P(\mathbf{x}|o_k, \lambda) = (2\pi\sigma^2_\lambda)^{-p/2} \exp\left\{ -\frac{1}{2\sigma^2_\lambda} \|\mathbf{x} - \mu_k\|^2 \right\} \qquad (5)$$

There is a linear relation between the logarithm of (5) and the distance $\|\mathbf{x} - \mu_k\|^2$. Applying logarithms to (4), we can see that the purpose of the MVQ modeling described here is to add to the log-score provided by the discrete HMM model (probability of generation) a new score (probability of quantization) that is linearly related to the average distortion of the input sequence X in the codebook of model λ. The idea of recognizing without time alignment using several VQ codebooks has been already proposed and successfully applied by Burton et al [6].

It is clear that the approach introduced by EXP3 is suboptimal. This means that the composition of probabilities in (4) may be upgraded using a weighting factor α,

$$\log P(X|\lambda) = \log P(X|O, \lambda) + \alpha \log P(O|\lambda) \qquad (6)$$

where $\alpha = \mu/(1-\mu)$, and μ takes values from 0 (only probability of generation) to 1 (only probability of quantization). Figure 2 shows how the error rate varies as a function of μ, in the range 0.25-0.75, for 8, 16 and 32 centers per codebook. Although $\mu = 0.5$ ($\alpha = 1$) is not a bad selection, there is a minimum error rate around $\mu = 0.3, 0.35$. We will use $\mu = 0.35$ ($\alpha = 0.538$) from now on. Figure 2 also shows that the probability of quantization is much more important in recognition than the probability of generation, since for $\mu > 0.3$ the slopes of the plots are smaller than for $\mu < 0.3$.

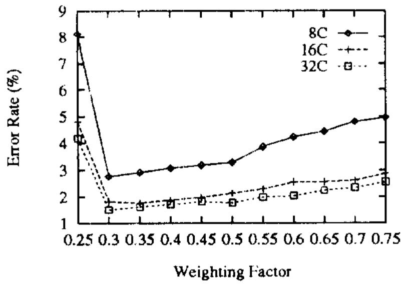

Figure 2. *Error rate vs. weighting factor μ.*

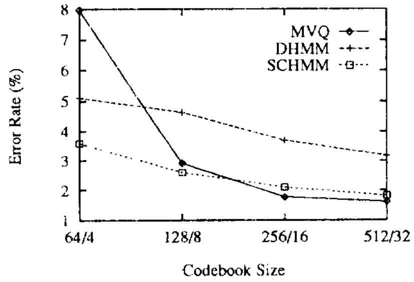

Figure 3. *Error rate versus codebook size MVQ, DHMM and SCHMM.*

A comparison of the designed MVQ system with standard DHMM and SCHMM systems is shown in figure 3. Since we use a 16-word vocabulary, the set of 16 N-center codebooks for MVQ models is equivalent to a (16*N)-center codebook for DHMM and SCHMM. Thus, the results are compared when the same total number of centers (4/64, 8/128, 16/256 and 32/512) is used. The SCHMM were constructed using gaussian *pdfs* and 4 VQ candidates, and trained by means of the joint reestimation proposed by Huang [7] (covariance matrices are not reestimated). It is clear than MVQ modeling outperforms DHMM models (using more than 8 centers) with the same amount of computation in recognition. Besides, there is an important computational saving in training, since the computation involved by 16 N-center codebooks is drastically smaller than that of a single (16*N)-center codebook (due to the exponential complexity of the VQ training algorithm). Also, MVQ models can even achieve similar or better results than SCHMMs.

4. SCMVQ HMM MODELING

The results obtained with MVQ modeling suggest the implementation of a new type of models that generalize MVQ for several quantization candidates, in the same way as SCHMM generalizes DHMM. This yields the SCMVQ HMM modeling, for which the output probabilities must be computed as,

$$b_i(\mathbf{x}) \simeq \sum_{k=1}^{C} P(\mathbf{x}|v_k, \lambda) b_i(v_k) \qquad (7)$$

where C is the number of VQ candidates (in $V(\lambda)$). The SCMVQ modeling is formally the closest one to CHMMs. The only difference is that all the states share the same set $V(\lambda)$ of *pdfs*. In spite of this similarity to CHMM models, it is easy to understand that the computational complexity is smaller than that of SCHMM due to the reduction of the number of parameters in the covariance matrices. This computational saving is drastic in the training stage if a joint reestimation is performed.

The *pdf* of equation (5) must be modified, due to the introduction of the weighting factor α (see eqn. (6)), as,

$$P(\mathbf{x}|o_j, \lambda) = (2\pi\sigma_\lambda^2)^{-p\alpha/2} \exp\left\{-\frac{\alpha}{2\sigma_\lambda^2}\|\mathbf{x} - \mu_j\|^2\right\} \qquad (8)$$

obtaining a non-normalized *pdf*. In this way, SCMVQ models are completely equivalent to MVQ models for the case of only 1 candidate.

The mechanisms of these new models, for training and recognition, are mostly similar to those of SCHMM models described in [2], in the same way as MVQs are similar to DHMMs. In a ML estimation of SCMVQs, the VQ parameters can be jointly estimated along with the discrete HMM parameters (A, B and Π matrices). The reestimation formulas for these parameters in a SCMVQ model are,

$$\hat{\mu}_k(i) = \frac{\sum_{l=1}^{S}\sum_{t=1}^{T^l} S_t^l(k) x_t^l(i)}{\sum_{l=1}^{S}\sum_{t=1}^{T^l} S_t^l(k)} \qquad (9a)$$

$$\hat{\sigma}_\lambda^2 = \frac{1}{p}\frac{\sum_{l=1}^{S}\sum_{t=1}^{T^l} S_t^l(k)\|x_t^l - \hat{\mu}_k\|^2}{\sum_{l=1}^{S}\sum_{t=1}^{T^l} S_t^l(k)} \qquad (9b)$$

where (see [2]),

$$S_t(k) = P(o_t = v_k|X, \lambda) \qquad (10)$$

and S is the number of training sequences.

Two different experiments have been carried out with the SCMVQ models described above:

EXP4 Only the A and B matrices are reestimated, using the same codebooks as for MVQ models.

EXP5 All the model parameters are reestimated, using formulas (9) (overall ML estimation).

These experiments were carried out for 8, 16 and 32 centers per codebook, using from 2 to 6 quantization candidates (1 candidate corresponds to MVQ modeling). The results are shown in table 1. Two main conclusions can be extracted:

#Cands.	1	2	3	4	5	6
8C EXP4	2.91	3.02	3.33	3.07	3.17	3.22
8C EXP5	-	2.70	2.96	2.81	2.81	2.81
16C EXP4	1.77	1.56	1.45	1.51	1.56	1.61
16C EXP5	-	1.45	1.51	1.56	1.61	1.66
32C EXP4	1.61	1.25	1.40	1.35	1.40	1.40
32C EXP5	-	1.25	1.14	1.09	1.09	1.09

Table 1. *Error rate values for SCMVQ modeling with 1-8 candidates, for 8, 16 and 32 centers per codebook. Experiences EXP4 and EXP5.*

#Cents.	D	SC	M	SCM	SC1
64/4	5.10	3.59	7.96	-	4.01
128/8	4.63	2.60	2.91	2.70	2.34
256/16	3.69	2.10	1.77	1.45	1.87
512/32	3.17	1.82	1.61	1.09	1.40

Table 2. *Error rate for DHMM, SCHMM, MVQ, SCMVQ and SCHMM1.*

a) The density of centers must be high enough, in order to obtain better results of recognition. This means that only when there exist more than one close center to an input vector, this is well represented by those centers. Thus, it will be very important to select the appropriate number of quantization candidates depending on the codebook size.

b) It is important to "teach" the system that other centers, different from the nearest one, can represent a given input vector. This is performed by the joint reestimation of EXP5. Thus, it possible to avoid the degradation of the 8-center system in EXP4, and to obtain meaningful improvements in the case of a high density of centers, as for 32 centers.

Finally, table 2 shows a comparison of the error rates achieved by DHMM (D), SCHMM (SC), MVQ (M) and SCMVQ (SCM) (as in EXP5). For SCMVQ, 2 quantization candidates are used for 8 and 16 centers, and 4 candidates for 32 centers. In relation to SCHMM, the computational complexity of SCMVQs is always smaller, due to the use of simplified covariance matrices. This computational reduction is more considerable in the cases of 8 and 16 centers, for which only 2 candidates are used (4 for SCHMMs). It can be observed that MVQs and SCMVQs are always superior for 16 and 32 centers. Also, an experience (labelled SC1) with SCHMM models, using a joint reestimation and the same *pdfs* as in (8), has been performed. It is interesting to observe that SC1 can obtain the same or better results than standard SCHMM, although only superior to SCMVQs for 8/128 centers. This new variation consumes the same computation as SCMVQ in recognition, but, again, more in training. This result ratifies the suitability of the *pdf* given by (8) for speech recognition.

5. SUMMARY

We have introduced in this paper a new type of HMM, called SCMVQ HMM. It is a generalization of MVQ modeling that has been recently introduced to enhance discrete HMMs with the spectral information lost in the VQ process. We looked first for an optimal form for the *pdfs* in a MVQ system. The chosen *pdf* has 3 main features: a) it reduces the needed number of parameters, b) it is coherent with the used distance measure, and c) it is weighted for an optimal composition with discrete HMM probabilities. The comparison of the obtained MVQ system with standard DHMMs and SCHMMs has shown the potential of this approach. The SCMVQ modeling generalizes MVQ using several quantization candidates. In the same way as for SCHMM models, the ML estimation of SCMVQs allows the joint reestimation of the VQ and the discrete HMM parameters. The results with SCMVQ models show that it is very important to have a codebook size high enough to correctly model an input vector with several quantization candidates, and that an appropriate selection of the number of candidates must be made. It is also important to train the system to use several candidates by means of a joint reestimation. The SCMVQs outperforms all the types of HMMs previously tested (DHMMs, SCHMMs and MVQs) using less computation than SCHMMs in recognition (due to the parameter reduction). This decrease in computation is much more important in the training stage.

REFERENCES

[1] 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, Febrero 1989.

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

[3] J. Segura, A. Rubio, A. Peinado, P. García, and R. Román, "Multiple VQ Hidden Markov Modeling for Speech Recognition," *Speech Communication (in press)*, 1994.

[4] A. Peinado, J. López, V. Sánchez, J. Segura, and A. Rubio, "Improvements in HMM-based isolated word recognition system," *IEE Proceedings-I*, vol. 138, pp. 201–206, Junio 1991.

[5] A. Peinado, *Selección y Estimación de Parámetros en Sistemas de Reconocimiento Automático de Voz basados Modelos Ocultos de Markov.* PhD thesis, Universidad de Granada, 1994 (In preparation).

[6] D. Burton and J. S. J. Buck, "Isolated-Word Speech Recognition Using Multisection Vector Quantization Codebooks," *IEEE Trans. on ASSP*, vol. 33, pp. 837–849, Agosto 1985.

[7] X. Huang, K. Lee, and H. Hon, "On Semi-Continuous Hidden Markov Modeling," in *Proc. of ICASSP-90*, pp. 689–692, 1990.