# DISCRETE MIXTURE HMM

*Satoshi Takahashi  Kiyoaki Aikawa  and  Shigeki Sagayama*

NTT Human Interface Laboratories

1-1 Hikarino-oka, Yokosuka-shi, Kanagawa 239 Japan

{taka, aik, saga}@nttspch.hil.ntt.co.jp

## ABSTRACT

This paper proposes a new type of acoustic model called the discrete mixture HMM (DMHMM). As large scale speech databases have been constructed for speaker-independent HMMs, continuous mixture HMMs (CMHMMs) are needed to increase the number of mixture components in order to represent complex distributions. This leads to a high computational cost for calculating output probabilities. The DMHMM represents the feature parameter space by using the mixtures of multivariate distributions in the same way as the diagonal covariance CMHMM. Instead of using Gaussian mixtures to represent feature distributions in each dimension, the DMHMM uses the mixtures of the discrete distributions based on the scalar quantization (SQ). Since the discrete distribution has a higher degree-of-freedom in terms of representation, the DMHMM is advantageous in representing the feature distributions efficiently with fewer mixture components. In isolated word recognition experiments for telephone speech, we have found that the DMHMM outperformed the CMHMMs when those models had the same number of mixture components.

## 1. INTRODUCTION

Overviewing the past progress in acoustic modeling, the discrete Hidden Markov Model (DHMM) based on vector quantization (VQ) was often used in the early years. The discrete distribution is non-parametric and can represent an arbitrary distribution shape. However, the VQHMM suffers from VQ distortions. In order to decrease the distortion, multiple codebooks were introduced to the VQHMM (See Figure 1).

There is an another stream of acoustic modeling techniques based on the continuous distribution. The continuous HMM (CHMM) represents feature distributions parametrically using Gaussian pdfs. When the amount of training data is small, using the parametric distribution is effective in estimating suitable distributions because the constraint on the distribution shape interpolates the unseen data. Currently, the continuous mixture HMM (CMHMM) is widely used for speaker-independent context-dependent phoneme models. The Gaussian mixture with the diagonal covariance matrices is preferred than the full covariance version from the reason of the parameter estimation.

The conditions surrounding the acoustic modeling have been changing, recently. Since large speech databases collecting from various speakers and recording conditions were built, the number of mixture components in an HMM state have been increased to represent complex feature distributions more accurately. Some speech recognition systems use models having more than 30 mixture components, and succeed in achieving higher level of recognition performance [1][2]. In general, model complexity trades off against model robustness for recognition. In recent situations, models need higher complexity to represent the distributions in detail since it is possible to use a large amount of data for model training.

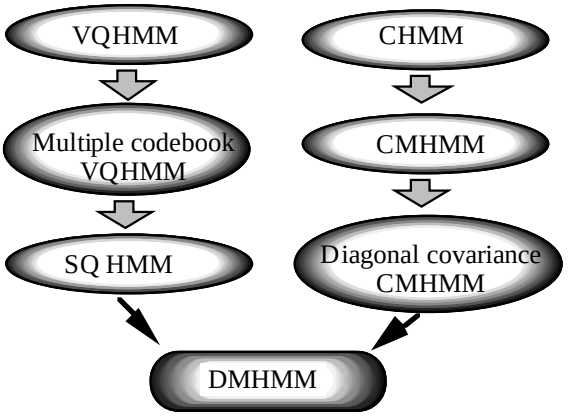


Figure 1. Historical view of acoustic modeling.

Applying Gaussian pdfs is a good idea for representing relatively simple distributions. However, is a Gaussian mixture the best way of representing complex distributions? Many distributions are required to represent a complex probability distribution in a multidimensional feature space since the Gaussian pdf has a lower degree-of-freedom in terms of representation. For the sake of improving the recognition performance, high computational cost is required, which basically increase monotonously in proportion to the number of mixture components. Simply increasing the number of mixture components of the CMHMM is not a good solution for efficient acoustic modeling. To solve this problem, this paper proposes the DMHMM which incorporates the advantages of both the DHMM and the CMHMM.

## 2. EFFICIENT MODELING USING DISCRETE MIXTURE HMM

### 2.1. Discrete mixture HMM

We propose the discrete mixture HMM (DMHMM) which uses discrete distributions as the mixture components in each feature dimension. Although the structure of this model is similar to the conventional multivariate CMHMM with diagonal covariance matrices, its Gaussian distributions in each dimension are replaced by discrete distributions. Each distribution is represented by the probabilities for the scalar quantization (SQ) codes. Since the discrete distribution can represent an arbitrary distribution shape (nonparametric), the DMHMM is considered to be more advantageous in accurately representing a complex distribution than the CMHMM.

Figure 2 illustrates an example of the two-dimensional feature space represented by the mixture Gaussian pdfs with diagonal covariance matrices. In this example, six one-dimensional Gaussian distributions in each dimension (twelve Gaussian distributions in total) represent the feature parameter space. By using the DMHMM, the feature parameter space can be covered by three one-dimensional discrete distributions

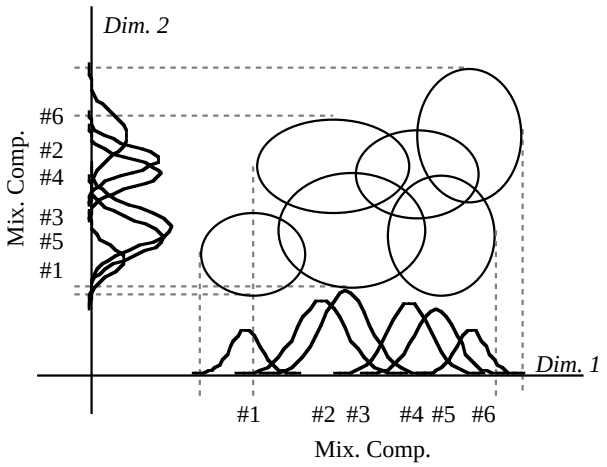


Figure 2. Feature parameter space represented by continuous mixture HMM (CMHMM).

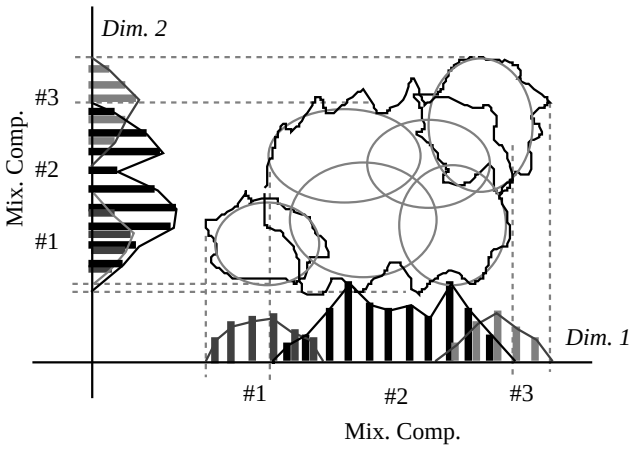


Figure 3. Feature parameter space represented by discrete mixture HMM (DMHMM).

(six discrete distributions in total) as shown in Figure 3. This example indicates that the same level of recognition performance will be obtained by the DMHMM with fewer mixture components than the CMHMM. In other words, when the DMHMM and the CMHMM having the same number of mixture components are trained using a large amount of data, the DMHMM will outperform the CMHMM.

In a continuous distribution framework, the diagonal covariance CMHMM is preferred over the full covariance CMHMM. There is a large number of parameters to be estimated in the full covariance version, and this makes reestimation difficult. The same story can be said for the discrete distribution case. The mixture of multivariate distribution composed by the SQ based one-dimensional discrete distribution is expected to be more efficient than the VQ based discrete distribution.

The output probability, $b_s(x_t)$, of state $s$ for $p$-dimensional input observation vector at time $t$, $x_t = (x_t^1, x_t^2, \ldots x_t^p)^T$, is calculated as

$$b_s(x_t) = \sum_{k \in K_s} \phi_{s,k}(x_t) \qquad (1)$$

$$\phi_{s,k}(x_t) = w_{s,k} \prod_{i=1}^{p} \psi_{s,k,i}(\tilde{x}_t^i) \qquad (2)$$

where $K_s$ denotes the subset of distributions comprising the mixture components for state $s$, and $w_{s,k}$ denotes the mixture weight coefficient for the $k$-th mixture component in state $s$. $\tilde{x}_t^i$ is the SQ code for $x_t^i$. $\psi_{s,k,i}(\tilde{x}_t^i)$ is the output probability distribution for the $k$-th mixture component in $i$-th dimension of state $s$.

### 2.2. Computational aspects

In the DMHMM, the output probabilities of the discrete distribution are stored in a table beforehand and the probability is referred to according to the SQ code. The probability is often calculated in the logarithmic domain. In the log likelihood calculation for each distribution, the Gaussian distribution requires the arithmetic operations $(x_t^i - \mu_{ki})^2 / 2\sigma_{ki}^2$, where $\mu_{ki}$ and $\sigma_{ki}^2$ are the mean and variance values. On the other hand, the discrete distribution requires SQ, which can be rapidly executed by float-to-integer conversion, and table look-up according to the SQ codes. This is the significant advantage of SQ, and is different from the VQ which requires a number of distance calculations. When the SQ code books for each feature dimension are shared by all discrete distributions in all models, the SQ is needed once at each frame. Thus theoretically, the DMHMM is faster than the CMHMM. In practical terms, the computation time depends on the employed computer architecture because the time needed for the table look-up varies.

In the previous work [3], it was found that substituting the SQ discrete distribution for the Gaussian pdf of the CMHMM did not degrade the recognition performance until reaching 16 SQ levels (SQHMM). This is also supported by the experiment in which the 10-th order LSP parameters with 4-bit quantization could represent LPC speech spectra at an average spectral distortion of 0.77 dB [4]. Also in [3], the

technique for fast output probability computation in the SQ based frame work was presented. The technique neglects the computation of Eq. (1) including a low SQ probability, which may not contribute to the state output probability. The technique attained an approximate 30% reduction in computational cost for the output probability. All these properties and techniques can be inherited in the DMHMM.

### 2.3. ML reestimation procedure for discrete mixture distributions

Maximum likelihood (ML) reestimation procedure for the output probability distributions of the DMHMM are described in this section. We first define $\xi_t(r,s,k)$, the probability of being in state $r$ at time $t$, and mixture component $k$ in state $s$ at time $t+1$, given model $\lambda$ and discrete observation sequence $X=(x_1,x_2,\cdots,x_t,\cdots,x_T)$.

$$\begin{aligned}\xi_t(r,s,k) &= \frac{P(q_t=r,q_{t+1}=s,k,X\mid\lambda)}{P(X\mid\lambda)} \\ &= \frac{\alpha_t(r)a_{rs}\phi_{s,k}(x_{t+1})\beta_{t+1}(s)}{P(X\mid\lambda)}\end{aligned} \qquad (3)$$

where $\alpha_t(r)$ and $\beta_{t+1}(s)$ are the forward and backward probabilities, respectively. $\phi_{s,k}(x_{t+1})$ is the discrete output probability for the $k$-th mixture component in state $s$, and is calculated as described in Eq. (2). Thus, the probability of being in the $k$-th mixture component in state $s$ at time $t$ is

$$\gamma_t(s,k)=\sum_{r=1}^{N}\xi_t(r,s,k) \qquad (4)$$

where $N$ is the number of states connected to state $s$. Finally, the reestimation formula for the output probability distribution (one-dimensional distribution in dimension $i$ of the $k$-th mixture component in state $s$ is

$$\begin{aligned}\psi_{s,k,i}(m) &= \frac{\sum_{t=1}^{T}\gamma_t(s,k)\delta(\tilde{x}_t^i,v_m^i)}{\sum_{t=1}^{T}\gamma_t(s,k)} \\ \delta(\tilde{x}_t^i,v_m^i) &= 1 \quad (if \quad \tilde{x}_t^i=v_m^i) \\ &= 0 \quad (otherwise)\end{aligned} \qquad (5)$$

where $\tilde{x}_t^i$ is the SQ code of the observation in dimension $i$ at time $t$, and $v_m^i$ is the SQ code for the $m$-th quantization point. The reestimation formulae for the state transition probability and the mixture coefficient are identical to those for the CMHMM.

### 2.4. Initial model estimation

There are a number of techniques for obtaining the initial models of the DMHMM training. In the following experiments, the initial models were generated based on the CMHMM:

(1) Training a CMHMM having the same number of mixture components with the desired DMHMM.

(2) To generate the SQ code books for each dimension, the range of quantization in dimension $i$ are set as shown below.

$$\left[\min_{k=1,2,\cdots K}(\mu_{ki}-5\sigma_{ki}),\ \max_{k=1,2,\cdots K}(\mu_{ki}+5\sigma_{ki})\right]$$

where $\mu_{ki}$ and $\sigma_{ki}$ are the mean and variance values of the $k$-th one-dimensional Gaussian distribution in dimension $i$. $K$ indicates the total number of distributions in all models. The range is equally divided by the quantization levels to determine the SQ centroids.

(3) Exchanging the continuous distribution in each dimension to the discrete distribution. Probability densities for the SQ centroids are calculated within $\pm 5\sigma$. The sum of probability densities is normalized by 1.0 for each distribution.

The state transition probabilities and the mixture coefficients are identical to those of the continuous distribution HMM generated in step (1).

## 3. EXPERIMENTAL EVALUATION

### 3.1. Experimental conditions

The DMHMMs were compared with the CMHMMs in speaker-independent isolated-word telephone speech recognition experiments. The performances were evaluated in both the context-independent modeling (26 models) and the context-dependent modeling (1,504 models). Model-level and state-level tying were carried out for the context-dependent HMMs [5], resulting in 400 shared states. The gender-dependent models were trained using 16,103 phoneme-balanced words uttered by male speakers, and 15,346 words uttered by female speakers. Data were collected through real telephone networks. 100-word sets each uttered by 9 male and 8 female speakers were used for evaluation. The vocabulary size for the recognition test was 1,200 words. The utterances were sampled at 8 kHz, and 12 mel-scaled frequency cepstrum coefficients (MFCCs) were calculated every 8 ms. The feature vector consists of 12 MFCCs, 12 first order time derivatives of the MFCCs and one dimensional delta-power (25 dimensions in total).

### 3.2. Results

The training procedure for the DMHMM was as follows. First, the context-independent CMHMMs were trained using data subsets with hand labels. Then, the models were trained using all the data. After two training iterations, the initial models for the DMHMMs were generated based on the CMHMMs as described in Sec. 2.4. The reestimation procedure iterated three times for both the CMHMM and the DMHMM after generation. Figure 4 shows the training curve along with the training iteration.

Table 1 shows the average word recognition performance using gender-dependent *context-independent* HMMs. The baseline system used the four-mixture CMHMM. The error reduction rate was calculated based on the baseline performance. For both the male and the female models, the CMHMM with sixteen mixture components attained an approximate 20% reduction in error. The DMHMM with four discrete mixture components were evaluated with two different

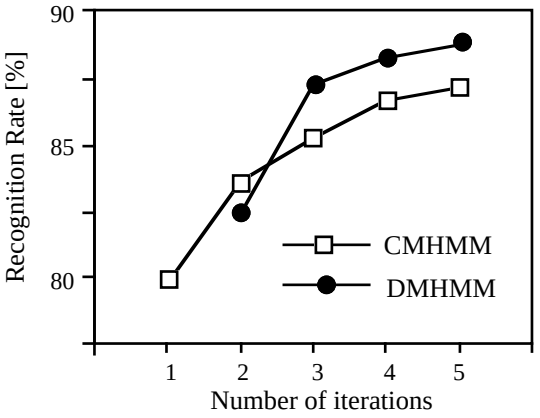


Figure 4. Recognition rates along with training iterations.

SQ levels. All feature dimensions were linearly scalar quantized into 20 SQ points and 40 SQ points within the range of $\pm 5\sigma$. Although the performance varies depending on the SQ levels, the DMHMM outperformed the CMHMM when those have the same number of mixture components. The performance did not reach to the sixteen mixture component CMHMM.

The computation time needed for calculating output probabilities was also measured for all models. Although a large amount of arithmetic operations were replaced by table look up, the CPU time needed for the DMHMM was increased by approximately 30% compared with the CMHMM (using HP735). This is mainly caused by the time needed for table look up which requires memory access, and the memory access time highly depends on the memory cache size.

When the models are compared at the same performance level, the DMHMM still has the advantage. To obtain the same performance as the DMHMM with four mixtures, namely the 10% error reduction, the CMHMM will require about eight mixture components if the performance increases proportionally to the number of mixtures. The double number of mixture components results in a 100% increase in the CPU time. Furthermore, the sixteen-mixture CMHMM required the computational cost as much as four times of the baseline system while its performance was the best.

Table 2 lists the performance using gender-dependent *context-dependent* HMMs. For male models, the DMHMM achieved higher recognition performance than the CMHMM also in the context-dependent model structure. Currently, the DMHMM failed to improve the performance of the female models.

## 4. CONCLUSION

This paper proposed the DMHMM to represent the feature parameter space efficiently and improve the recognition performance. Instead of using the Gaussian mixtures, the new model uses the mixtures of the discrete distributions to represent the feature distributions. Each mixture component composes the multivariate distribution similar to the CMHMM. The model has the advantage for a model with complex distributions using a large amount of training data.

The experimental results show that the DMHMM obtained a better recognition performance than the CMHMM when the number of mixture components were the same. From the view point of the computational cost, the DMHMM succeeded in obtaining the same performance with a low computational cost.

Table 1. Comparison in performance for the gender-dependent context-***independent*** models.

| Gender | | | Male | | Female | |
|---|---|---|---|---|---|---|
| Model | | | Recog. Rate [%] | Error Reduction Rate [%] | Recog. Rate [%] | Error Reduction Rate [%] |
| CM-HMM | 4 mix. (baseline) | | 87.1 | N/A | 77.8 | N/A |
| | 16 mix. | | 89.9 | 21.7 | 82.5 | 21.2 |
| DM-HMM | 4 mix. | 20 SQ levels | 87.4 | 2.3 | 81.8 | 18.0 |
| | | 40 SQ levels | 88.8 | 13.2 | 79.0 | 5.4 |

Table 2. Comparison in performance for the gender-dependent context-***dependent*** models.

| Gender | | | Male | | Female | |
|---|---|---|---|---|---|---|
| Model | | | Recog. Rate [%] | Error Reduction Rate [%] | Recog. Rate [%] | Error Reduction Rate [%] |
| CM-HMM | 4 mix. (baseline) | | 92.9 | N/A | 86.8 | N/A |
| | 16 mix. | | 94.6 | 23.9 | 89.8 | 22.7 |
| DM-HMM | 4 mix. | 20 SQ levels | 93.9 | 14.1 | 87.2 | 3.0 |

## REFERENCES

[1] J. L. Gauvain, L. Lamel and M. A.-Decker, "Developments in Continuous Speech Dictation using the ARPA WSJ Task," Proc. of ICASSP95, pp. 65-68, 1995.

[2] C. H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini and A. E. Rosenberg, "Improved Acoustic Modeling for Large Vocabulary Continuous Speech Recognition," Computer Speech and Language, 6, pp. 103-127, 1992.

[3] S. Sagayama and S. Takahashi, "On the Use of Scalar Quantization for Fast HMM Computation," Proc. of ICASSP95, pp. 213-216, 1995.

[4] F. Itakura and N. Sugamura, "LSP Speech Synthesizer, Its Principle and Implementation," Technical Report of Acoustic Society of Japan, S79-46, pp. 349-356, 1979 (in Japanese).

[5] J. Takami and S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. of ICASSP92, 1996.