1989

Multiple codebook semi-continuous hidden Markov models for speaker-independent continuous speech recognition

X. D. Huang
Carnegie Mellon University

Hsiao-Wuen Hon

Kai-Fu Lee

Multiple Codebook Semi-Continuous Hidden Markov Models for Speaker-Independent Continuous Speech Recognition

X. D. Huang[1], H. W. Hon and K. F. Lee

April 19, 1989
CMU-CS-89-136

School of Computer Science
Carnegie-Mellon University
Pittsburgh, PA 15213

Abstract

A semi-continuous hidden Markov model based on the multiple vector quantization codebooks is used here for large-vocabulary speaker-independent continuous speech recognition. In the techniques employed here, the semi-continuous output probability density function for each codebook is represented by a combination of the corresponding discrete output probabilities of the hidden Markov model and the continuous Gaussian density functions of each individual codebook. Parameters of the vector quantization codebook and the hidden Markov model are mutually optimized to achieve an optimal model/codebook combination under a unified probabilistic framework. Another advantage of this approach is the enhanced robustness of the semi-continuous output probability density function by the combination of multiple codewords and multiple codebooks. For a 1000-word speaker-independent continuous speech recognition using a word-pair grammar, the recognition error rate of the semi-continuous hidden Markov model was reduced by more than 29% and 40% in comparison to the discrete and continuous mixture hidden Markov model respectively.

This research was sponsored in part by the Defense Advanced Research Projects Agency under Contract N00039-85-C-0163. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency, or the US Government. X.D. Huang is a holder of an Edinburgh University Studentship and ORS Awards.

---

[1]Visiting scientist from CSTR, University of Edinburgh, 80, South Bridge, Edinburgh EH1 1HN, Scotland

TABLE OF CONTENTS

ABSTRACT

A semi-continuous hidden Markov model based on the multiple vector quantization codebooks is used here for large-vocabulary speaker-independent continuous speech recognition. In the techniques employed here, the semi-continuous output probability density function for each codebook is represented by a combination of the corresponding discrete output probabilities of the hidden Markov model and the continuous Gaussian density functions of each individual codebook. Parameters of vector quantization codebook and hidden Markov model are mutually optimized to achieve an optimal model/codebook combination under a unified probabilistic framework. Another advantages of this approach is the enhanced robustness of the semi-continuous output probability by the combination of multiple codewords and multiple codebooks. For a 1000-word speaker-independent continuous speech recognition using a word-pair grammar, the recognition error rate of the semi-continuous hidden Markov model was reduced by more than 29% and 41% in comparison to the discrete and continuous mixture hidden Markov model respectively.

1. INTRODUCTION

The hidden Markov Model (HMM), which can be based either on the discrete observation probability distributions or continuous mixture probability density functions, has been demonstrated as one of the most successful techniques for automatic speech recognition [Jelinek85a, Chow87a, Rabiner88a, Lee89a].

In the discrete HMM, vector quantization (VQ) [Makhoul85a] produces the closet codebword from the codebook for each acoustic observation. This mapping from continuous acoustic space to quantized discrete space may cause serious quantization errors for subsequent hidden Markov modeling. To reduce VQ errors, various smoothing techniques have been proposed for VQ and subsequent hidden Markov modeling [Jelinek76a, Nishimura87a, Tseng87a, Lee88a, Huang88a, Schwartz89a]. A distinctive technique is multiple VQ codebook hidden Markov modeling, which has been shown to offer improved speech recognition accuracy [Gupta87a, Lee88a]. In the multiple VQ codebook approach, VQ distortion can be significantly minimized by partitioning the parameters into separate codebooks. Another disadvantage of the discrete HMM is that the VQ codebook and the discrete HMM are separately modeled, which may not be an optimal combination for pattern classification [Huang89a]. The discrete HMM uses the discrete output probability distributions to model various acoustic events, which are inherently superior to the continuous mixture HMM with mixture of a small number of probability density functions since the discrete distributions could model events with any shapes provided enough training data exist.

On the other hand, the continuous mixture HMM models the acoustic observation directly using estimated continuous probability density functions without VQ, and has been shown to improve the recognition accuracy in comparison to the discrete HMM [Rabiner85a, Poritz86a, Brown87a]. For speaker-independent speech recognition, mixture of a large number of probability density functions [Rabiner88a, Paul89a] or a large number of states in single-mixture case [Doddington89a] are generally required to model characteristics of different speakers. However, mixture of a large number of probability density functions will considerably increase not only the computational complexity, but also the number of free parameters that can be reliablely estimated. In addition, the continuous mixture HMM has to be used with care as continuous probability density functions make more assumption than the discrete HMM, especially when the diagonal covariance Gaussian probability density is used for simplicity [Rabiner85a, Brown87a]. To obtain a better recognition accuracy, acoustic parameters must be well chosen according to the assumption of the continuous probability density functions used.

The semi-continuous hidden Markov model (SCHMM) has been proposed to extend the discrete HMM by replacing discrete output probability distributions with a combination of the original discrete output probability distributions and continuous probability density functions of a Gaussian codebook [Huang88a]. In the SCHMM, each VQ codeword is regarded as a Gaussian probability density. Intuitively, from the discrete HMM point of view, the SCHMM tries to smooth the discrete output probabilities with multiple codeword candidates in VQ procedure. From the continuous mixture HMM point of view, the SCHMM ties all the continuous output probability densities across each individual HMM to form a shared Gaussian codebook, i.e. a mixture of Gaussian probability densities. With the SCHMM, the codebook and HMM can be jointly re-estimated to achieve an optimal codebook/model combination in sense of maximum likelihood criterion. Such a tying can also substantially reduce the number of free parameters and computational complexity in

comparison to the continuous mixture HMM, while maintain reasonablely modeling power of a mixture of a large number of probability density functions. The SCHMM has shown to offer improved recognition accuracy in several speech recognition experiments [Huang88a, Huang89a, Huang89b, Paul89a, Bellegarda89a]*.

In this study, the SCHMM is applied to Sphinx, a speaker-independent continuous speech recognition system. Sphinx uses multiple VQ codebooks for each acoustic observation [Lee88a]. To apply the SCHMM to Sphinx, the SCHMM algorithm must be modified to accommodate multiple codebooks and multiple codewords combination. For the SCHMM re-estimation algorithm, the modified unified re-estimation algorithm for multiple VQ codebooks and hidden Markov models are proposed in this paper. The applicability of the SCHMM to speaker-independent continuous speech is explored based on 200 generalized triphone models [Lee88a]. In the 1000-word speaker-independent continuous speech recognition task using word-pair grammar, the error rate was reduced by more than 29% and 41% in comparison to the corresponding discrete HMM and continuous mixture HMM respectively.

This paper is organized as follows. In Section 2, the mathematical formulation of the HMM is reviewed and re-estimation algorithms for the semi-continuous HMM with multiple codebooks are derived. In Section 3, the implementation of the SCHMM is discussed and experimental results for 1000-word speaker-independent continuous speech recognition are presented. Finally, Section 4 contains a summary and discussions.

* We noticed that J. Bellegarda *et al*. in IBM Watson Research Center developed their techniques independently.

2. SEMI-CONTINUOUS HIDDEN MARKOV MODELS

2.1. Discrete HMMs and Continuous HMMs

An N-state Markov chain with state transition matrix A=[a_{ij}], i,j=1, 2, ..., N, where a_{ij} denotes the transition probability from state i to state j; and a discrete output probability distribution, $b_j(O_k)$, or continuous output probability density function $b_j(\mathbf{x})$ associated with each state j of the unobservable Markov chain is considered here. Here O_k represents discrete observation symbols (usually VQ indices), and $\mathbf{x}$ represents continuous observations (usually speech frame vectors) of K-dimensional random vectors.

With the discrete HMM, there are L discrete output symbols from a L-level VQ, and the output probability is modeled with discrete probability distributions of these discrete symbols. Let $\mathbf{O}$ be the observed sequence, $\mathbf{O}= O_{k_1}O_{k_2}\cdots O_{k_T}$ observed over T samples. Here O_{k_i} denotes the VQ codeword k_i observed at time i. The observation probability of such an observed sequence, $Pr(\mathbf{O}\,|\,\lambda)$, can be expressed as:

$$Pr(\mathbf{O}\,|\,\lambda) = \sum_{S} \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(O_{k_t}) \tag{1}$$

where S is a particular state sequence, $S \in (s_0, s_1, \cdots, s_t)$, $s_t \in \{1, 2, ..., N\}$, and the summation is taken over all of the possible state sequences, S, of the given model λ, which is represented by (π, A, B), where π is the initial state probability vector, A is the state transition matrix, and B is the output probability distribution matrix. In the discrete HMM, classification of O_{k_t} from $\mathbf{x}_t$ in the VQ may not be accurate. The effects of VQ errors may cause the performance of the discrete HMM to be inferior to that of the continuous mixture HMM [Rabiner85a].

If the observation to be decoded is not vector quantized, then the probability density function, $f(\mathbf{X}\,|\,\lambda)$, of producing an observation of continuous vector sequences given the model λ, would be computed, instead of the probability of generating a discrete observation symbol, $Pr(\mathbf{O}\,|\,\lambda)$. Here $\mathbf{X}$ is a sequence of continuous (acoustic) vectors $\mathbf{x}$, $\mathbf{X}=\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_T$. The principal advantage of using the continuous HMM is the ability to directly model speech parameters without involving VQ. However, the continuous HMM requires considerably longer training and recognition times, especially when a mixture of several Gaussian probability density components is used. In the continuous HMM, Eq. (1) can be re-written as:

$$f(\mathbf{X}\,|\,\lambda) = \sum_{S} \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(\mathbf{x}_t) \tag{2}$$

where the output probability density function can be represented by mixture of Gaussian probability density functions. More generally, in the continuous Gaussian (M-component) mixture HMM [Juang85a], the output probability density of state j, $b_j(\mathbf{x})$, can be represented as

$$b_j(\mathbf{x}) = \sum_{k=1}^{M} c_{jk} N(\mathbf{x}, \mu_{jk}, \sum_{jk}) \tag{3}$$

where $N(\mathbf{x}, \mu, \sum)$ denotes a multi-dimensional Gaussian density function of mean vector μ and covariance matrix $\sum$, and c_{jk} is a weighting coefficient for the kth Gaussian component. With such a mixture, any arbitrary distribution cab be approximately modeled, provided the mixture is

large enough.

For either the discrete HMM or continuous mixture HHM, the Baum-Welch re-estimates [Baum72a, Rabiner86a, Poritz86a], $\bar{\lambda}$, guarantee increase of the likelihood, i.e. $Pr(\mathbf{O}\,|\,\bar{\lambda}) \geq Pr(\mathbf{O}\,|\,\lambda)$ or $f(\mathbf{X}\,|\,\bar{\lambda}) \geq f(\mathbf{X}\,|\,\lambda)$, unless a local maximum has been reached. This algorithm can be used to iteratively improve the model parameters until some criterion of convergence is satisfied.

2.2. Semi-Continuous Hidden Markov Models

In the discrete HMM, the discrete probability distributions are sufficiently powerful to model any random events with a reasonable number of parameters. The major problem with the discrete output probability is that the VQ operation partitions the acoustic space into separate regions according to some distortion measure. This introduces errors as the partition operations may destroy the original signal structure. An improvement is to model the VQ codebook as a family of Gaussian density functions such that the distributions are overlaped, rather than disjointed. Each codeword of the codebook can then be represented by one of the Gaussian probability density functions and may be used together with others to model the acoustic event. The use of a parametric family of finite mixture densities (a mixture density VQ) can then be closely combined with the HMM methodology. From the continuous mixture HMM point of view, the output probability in the continuous mixture HMM is shared among the Gaussian probability density functions of the VQ. This can reduce the number of free parameters to be estimated as well as the computational complexity. From the discrete HMM point of view, the partition of the VQ is unnecessary, and is replaced by the mixture density modeling with overlap, which can effectively minimize the VQ errors.

The problems of estimating the parameters which determine a mixture density has been the subject of a large, diverse body of literature spanning some ninety years [Redner84a]. The procedure, known as the EM algorithm [Dempster77a], is a specialization, to the mixture density context, of a general algorithm for obtaining maximum likelihood estimates for incomplete problems, i.e. the mixture density estimation problem is an estimation problem involving incomplete data by regarding an unlabeled observation on the mixture as *missing* a label indicating its component population of origin. This has been defined earlier by Baum [Baum70a] in a similar way and has been widely used in HMM-based speech recognition methods. Thus, the VQ problems and HMM modeling problems can be unified under the same probabilistic framework to obtain an optimized VQ/HMM combination, which forms the foundation of the SCHMM.

Provided that each codeword of the VQ codebook is represented by a Gaussian density function, for a given state s_t of HMM, the probability density function that s_t produces a vector $\mathbf{x}$ can then be written as:

$$\begin{aligned} b_{s_t}(\mathbf{x}) &= f(\mathbf{x}\,|\,s_t) \\ &= \sum_{j=1}^{L} f(\mathbf{x}\,|\,O_{j_t}, s_t) Pr(O_{j_t}\,|\,s_t) \end{aligned} \tag{4.a}$$

where L denotes the VQ codebook level. For the sake of simplicity, the output probability density function conditioned on the codewords can be assumed to be independent of the Markov states s_t, (4.a) can then be written as:

$$f(\mathbf{x}|s_t) = \sum_{j=1}^{L} f(\mathbf{x}|O_{j_t}) Pr(O_{j_t}|s_t)$$
$$= \sum_{j=1}^{L} f(\mathbf{x}|O_{j_t}) b_{s_t}(O_{j_t}) \quad (4.b)$$

This equation is the key to the semi-continuous hidden Markov modeling. Given the VQ codebook index O_{j_t}, the probability density function f($\mathbf{x}|O_{j_t}$) can be estimated with the EM algorithm [Redner84a], or maximum likelihood clustering [Huang89c]. It can also be obtained from the HMM parameter estimation directly as explained later. Using (4) to represent the semi-continuous output probability density, it is possible to combine the codebook distortion characteristics with the parameters of the discrete HMM under a unified probabilistic framework. Here, each discrete output probability is weighted by the continuous conditional Gaussian probability density function derived from VQ. If these continuous VQ density functions are considered as the continuous output probability density function in the continuous mixture HMM, this also resembles the L-mixture continuous HMM with all the continuous output probability density functions shared with each other in the VQ codebook. Here the discrete output probability in state i, $b_i(O_{j_t})$, becomes the weighting coefficients for the mixture components.

In implementation of the SCHMM [Huang89b], Eq. (4) can be replaced by finding M most significant values of $f(\mathbf{x}|O_j)$ (with M be one to five, the algorithm converges well in practice) over all possible codebook indices O_j, which can be easily obtained in the VQ procedure. This can significantly reduce the amount of computational load for subsequent output probability computation since M is of lower order than L. Experimental results show this to perform well in speech recognition [Huang89b], and result in an L-mixture continuous HMM with a computational complexity significantly lower than the continuous mixture HMM.

2.3. Re-estimation formulas for the SCHMM

If the $b_i(O_{j_t})$ are considered as the weighting coefficients of different mixture output probability density functions in the continuous mixture HMM, the re-estimation algorithm for the weighting coefficients can be extended to re-estimate $b_i(O_{j_t})$ of the SCHMM [Juang85a]. The re-estimation formulations can be more readily computed by defining a forward partial probability, $\alpha_t(i)$, and a backward partial probability, $\beta_t(i)$ for any time t and state i as:

$$\alpha_t(i) = Pr(\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_t, s_t=i \,|\, \lambda)$$
$$\beta_t(i) = Pr(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \cdots \mathbf{x}_T \,|\, s_t=i, \lambda) \quad (5.a)$$

The forward and backward probability can be computed recursively as:

$$\alpha_t(i) = \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} b_i(\mathbf{x}_t), \quad 2 \le t \le T;$$
$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{x}_{t+1}) \beta_{t+1}(j). \quad 1 \le t \le T-1. \quad (5.b)$$

where $\alpha_1(i)=\pi_i$ and $\beta_T(i)=1$ if i is a final state, otherwise $\beta_T(i)=0$.

The intermediate probabilities, $\chi_t(i,j,k)$, $\gamma_t(i,j)$, $\gamma_t(i)$, $\zeta_t(i,j)$, and $\zeta_t(j)$ can be defined as

follows for efficient re-estimation of the model parameters:

$$\chi_t(i,j,k) = Pr(s_t=i,\ s_{t+1}=j,\ O_{k_{t+1}} | \mathbf{X}, \lambda)$$
$$= \frac{\alpha_t(i) a_{ij} b_j(O_{k_{t+1}}) f(\mathbf{x}_{t+1} | O_{k_{t+1}}) \beta_{t+1}(j)}{Pr(\mathbf{X}|\lambda)}$$
$$\gamma_t(i,j) = Pr(s_t=i,\ s_{t+1}=j | \mathbf{X}, \lambda) \tag{6.a}$$
$$\gamma_t(i) = Pr(s_t=i | \mathbf{X}, \lambda)$$
$$\zeta_t(i,k) = Pr(s_t=i, O_{k_t} | \mathbf{X}, \lambda)$$
$$\zeta_t(k) = Pr(O_{k_t} | \mathbf{X}, \lambda)$$

All these intermediate probabilities can be represented by $\chi_t()$ as

$$\gamma_t(i,j) = \sum_{k=1}^{L} \chi_t(i,j,k) \tag{6.b}$$

$$\gamma_t(i) = \sum_{j=1}^{N} \gamma_t(i,j) \tag{6.c}$$

$$\zeta_t(i,k) = \sum_{j=1}^{N} \chi_{t-1}(j,i,k) \tag{6.d}$$

$$\zeta_t(k) = \sum_{i=1}^{N}\sum_{j=1}^{N} \chi_{t-1}(i,j,k) \tag{6.e}$$

Using Eq. (5) and (6), the re-estimation equations for π_i, a_{ij}, and $b_i(O_j)$ can be written as:

$$\pi_i = \gamma_1(i), \quad 1 \le i \le N; \tag{7}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)}, \quad 1 \le i, j \le N; \tag{8}$$

$$\bar{b}_i(O_j) = \frac{\sum_{t=1}^{T} \zeta_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)}, \quad 1 \le i \le N; \ 1 \le j \le L. \tag{9}$$

The means and covariances of the Gaussian probability density functions can also be re-estimated to update the VQ codebook separately with Eq. (5) and (6). The feedback from the HMM estimation results to the VQ codebook implies that the VQ codebook is optimized based on the HMM likelihood maximization rather than minimizing the total distortion errors from the set of training data. Although re-estimation of means and covariances of different models will involve interdependencies, the different density functions which are re-estimated are strongly correlated. To re-estimate the parameters of the VQ codebook, i.e. the means, μ_j, and covariance matrices, Σ_j, of

the codebook index j, it is not difficult to extend the continuous mixture HMM re-estimation algorithm with modified Q function [Huang89c]. In general, it can be written as:

$$\bar{\mu}_j = \frac{\sum_{\nu}[\sum_{t=1}^{T}\zeta_t(j)\mathbf{x}_t]}{\sum_{\nu}[\sum_{t=1}^{T}\zeta_t(j)]}, \quad 1 \le j \le L; \tag{10}$$

and

$$\bar{\Sigma}_j = \frac{\sum_{\nu}[\sum_{t=1}^{T}\zeta_t(j)(\mathbf{x}_t-\bar{\mu}_j)(\mathbf{x}_t-\bar{\mu}_j)^t]}{\sum_{\nu}[\sum_{t=1}^{T}\zeta_t(j)]}, \quad 1 \le j \le L. \tag{11}$$

where ν denotes the HMM used; and expressions in [ ] are variables of model ν. In Eq. (10) and (11), the re-estimation for the means and covariance matrices in the output probability density function of the SCHMM are tied up with all the HMM models, which is similar to the approach with tied transition probability inside the model [Jelinek80a]. From Eq. (10) and (11), it can be observed that they are merely a special form of EM algorithm for the parameter estimation of mixture density functions [Redner84a], which are closely welded into the HMM re-estimation equations.

2.4. Multiple Codebook Case

When multiple codebooks are used, each codebook represents a set of different speech parameters. One way to combine these multiple output observations is to assume that they are independent, and the output probability is computed as the product of the output probability of each codebook. It has been shown that performance using multiple codebook can be substationally improved [Lee89a]. In the semi-continuous HMM, the semi-continuous output probability of multiple codebooks can also be computed as the product of the semi-continuous output probability for each codebook as Eq. (4), which consists of L-mixture continuous density functions. In other word, the semi-continuous output probability could be modified as:

$$b_{s_t}(\mathbf{x}) = \prod_{c} \sum_{j=1}^{L} f^{c}(\mathbf{x} \mid O_{j_t}^{c}) b_{s_t}^{c}(O_{j_t}^{c}) \tag{12}$$

where c denotes the codebook used. The re-estimation algorithm for the multiple codebook based HMM could be extended if Eq. (6.a) is computed for each codeword of each codebook c with the combination of the rest codebook probability. Since multiplication of the semi-continuous output probability density of each codebook lead to several independent items in Q function [Huang89c], for codebook c_l, $\chi_t(i,j,k^{c_l})$ could be extended as:

$$\chi_t(i,j,k^{c_l}) = \frac{\alpha_t(i) a_{ij} b_j^{c_l}(O_{k_{t+1}}^{c_l}) f^{c_l}(\mathbf{x}_{t+1} \mid O_{k_{t+1}}^{c_l}) \prod_{c \neq c_l} [\sum_{m=1}^{L} f^{c}(\mathbf{x} \mid O_{m_t}^{c}) b_{s_t}(O_{m_t}^{c})] \beta_{t+1}(j)}{Pr(\mathbf{X} \mid \lambda)} \tag{13}$$

Other intermediate probability can also be computed similar to Eq. (13). It can be easily proved that this is consistent with maximum likelihood criterion.

Experimental observation shows that the increase of the likelihood of training data due to the re-estimation of codebook is about 3-6 times as large as the re-estimation of HMM output and transition probabilities alone, which indicates in part the significance of re-estimation of the VQ codebook.

3. EXPERIMENTAL EVALUATION

3.1. Analysis Conditions

For both training and evaluation, the standard Sphinx analysis conditions consist of the following:

sampling rate	16 kHz
analysis method	bilinear transformed LPC cepstrum
LPC analysis order	14
cepstrum order	12
bilinear transformation constant	0.6
window type	Hamming window
window length and shift	20 ms and 10 ms
pre-emphasis	$1\text{-}0.97z^{-1}$

The complete database consists of 4358 training sentences from 105 speakers (june-train) and 300 test sentences from 12 speakers. For tuning experiments conducted here, the training data consist of 2880 sentences from 72 speakers, the tuning test data consist of 45 sentences from 12 speakers, which are extracted randomly from the 300 test sentences.

The vocabulary of the Resource Management database is 991 words. There is also an "official" word-pair recognition grammar, which is just a list of allowable word pairs without probabilities for the purpose of reducing the recognition perplexity to about 60.

3.2. Experimental Results Using Bilinear Transformed Cepstrum

Discrete HMMs and continuous mixture HMMs based on 200 generalized triphones are first experimented as benchmarks. The discrete HMM is the same as Sphinx except only 200 generalized triphones are used [Lee88a].

In the continuous mixture HMM implemented here, the cepstrum, difference cepstrum, normalized energy, and difference energy are packed into one vector. This is similar to the one codebook implementation of the discrete HMM [Lee88a]. Each continuous output probability consists of 4 diagonal Gaussian probability density function as Eq. (3). To obtain reliable initial models for the continuous mixture HMM, the Viterbi alignment with the discrete HMM is used to phonetically segment and label training speech. These labeled segments are then clustered by using the k-means clustering algorithm to obtain initial means and diagonal covariances. The forward-backward algorithm is used iteratively for the monophone models, which are then used as initial models for the generalized triphone models. Though continuous mixture HMM was reported to significantly better the performance of the discrete HMM [Rabiner85a], for the experiments conducted here, it is *significantly* worse than the discrete HMM. Why is this paradox? One explanation is that multiple codebooks are used in the discrete HMM, therefore the VQ errors for the discrete HMM are

not so serious here. Another reason may be that the diagonal covariance assumption is not appropriate for the bilinear transformed LPC cepstrum since many coefficients are strongly correlated after the transformation. Indeed, observation of average covariance matrix for the bilinear transformed LPC cepstrum shows that values of off-diagonal components are generally quite large.

For the semi-continuous model, multiple codebooks are used instead of packing different feature parameters into one vector. The initial model for the SCHMM comes directly from the discrete HMM with the VQ variance obtained from k-means clustering for each codeword. Though diagonal Gaussian assumption may be inappropriate here, the SCHMM outperformed either the discrete HMM or continuous mixture HMM. In computing the semi-continuous output probability density function, only M most significant codewords are used for subsequent processing. Experiments with top one and top four codewords were conducted.

Under the same analysis condition, the percent correct (correct word percentage) and word accuracy (percent correct - percent insertion) results of the discrete HMM, the continuous mixture HMM, and the SCHMM are shown in Table 1. Parameters tuned include floor to the diagonal covariance, language weight, and number of iterations. The optimum value for the iteration number is 1-3; for language weight is 2.5-3.8; for covariance floor is 10^{-3} to 0 depending on different acoustic parameters used.

Table 1
Average recognition accuracy based on 200 generalized triphones 4358 training sentences; 300 test sentences

types	*percent correct (word accuracy)*
Discrete HMM	89.5% (88.0%)
Continuous Mixture HMM	84.2% (81.3%)
Semi-continuous HMM + top1	87.2% (84.0%)
Semi-continuous HMM + top4	90.6% (89.1%)

From Table 1, it can be observed that the SCHMM with top 4 codewords works better than both the discrete and continuous mixture HMM. The SCHMM with top 1 codeword works actually worse than the discrete HMM. Though bilinear transformed cepstral coefficients could not be well modeled by the diagonal Gaussian assumption (which was proven by the poor performance of the continuous mixture HMM and the SCHMM with top 1 codeword), the SCHMM with top 4 codeword works modestly better than the discrete HMM. The improvement may primarily come from smoothing effect of the SCHMM, i.e. the robustness of multiple codewords and multiple codebooks in the semi-continuous output probability representation, albeit 200 generalized triphone models are relatively well trained in comparison to standard Sphinx version [Lee88a], where 1000 generalized triphone models are used. Detailed observations suggest that the SCHMM can significantly improve the performance of some speakers, but not others. Overall, it is only slightly better than the discrete HMM.

3.3. Experimental Results Using Less Correlated Data

If the diagonal Gaussian covariance is used, each dimension in speech vector should be uncorrelated. In practice, this can be partially satisfied by using less correlated feature as acoustic observation representation.

One way to reduce correlation is principal component projection. In the implementation here, the projection matrix is computed by first pooling together the bilinear transformed cepstrum of the whole training sentences, and then computing the eigenvector of that pooled covariance matrix. For the tuning database, result comparison between projected data and un-projected data is shown in Table 2. Unfortunately, only insignificant improvements are obtained based on such a projection. This is because the covariance for each codeword is quite different, and such a projection only makes *average* covariance diagonal, which is inadequate.

Table 2
Average accuracy of projected and un-projected data
2880 training sentences; 45 tuning test sentences

types	*percent correct (word accuracy)*
Semi-continuous HMM + top1	87.8% (85.3%)
Semi-continuous HMM + top1 + projection	88.3% (85.8%)

As bilinear transformed cepstral coefficients could not be well modeled by diagonal Gaussian probability density function, experiments without bilinear transformation are conducted. Here, 18th order cepstral coefficients derived from 18th order LPC analysis are compared with 12th order cepstral coefficients derived from 14th order LPC analysis. Results for the discrete HMM are shown in Table 3. The 12th order cepstrum with bilinear transformation is also listed in Table 3 for reference. Though previous experimental results suggest that the recognition accuracy of the 18th order bilinear transformed cepstrum is about the same as that of the 12th order bilinear transformed cepstrum, the recognition accuracy of the 18th order cepstrum is better than that of the 12th order cepstrum, but worse than that of the bilinear transformed cepstrum.

Table 3
Average accuracy of discrete HMMs based on 200 generalized triphones 2880 training sentences; 45 tuning test sentences

types	*percent correct (word accuracy)*
12th LPC	84.4% (81.6%)
12th LPC + Bilinear Transformation	88.2% (86.2%)
18th LPC	86.1% (82.9%)

The 18th order cepstrum is used here for the SCHMM because of less correlated characteristics of the cepstrum. With 4358 training sentences (june-train), test results of 300 sentences (june-test) are listed in Table 4.

Table 4
Average accuracy of 18th order cepstrum based on 200 generalized triphones 4358 training sentences; 300 test sentences

types	*percent correct (word accuracy)*
Discrete HMM	86.3% (83.8%)
Semi-continuous HMM + top1	86.6% (85.5%)
Semi-continuous HMM + top2	88.8% (87.6%)
Semi-continuous HMM + top4	89.3% (88.5%)
Semi-continuous HMM + top6	89.6% (88.6%)
Semi-continuous HMM + top8	89.3% (88.2%)

Here, the recognition accuracy of the SCHMM is significantly improved in comparison with the discrete HMM, and error reduction is over 29%. Even the SCHMM with top one codeword is used, it is still better than the discrete HMM (85.5% vs. 83.8%). Use of multiple codewords (top4 and top6) in the semi-continuous output probability density function greatly improves the word accuracy (from 85.5% to 88.6%). Further increase of codewords used in the semi-continuous output probability density functions shows no improvement on word accuracy, but substantial growth of computational complexity. From Table 4, it can be seen that the SCHMM with top four codewords is adequate (88.5%). In contrast, when bilinear transformed data was used (Table 2), the error reduction is less than 10% in comparison to the discrete HMM, and the SCHMM with top one codeword is actually slightly worse than the discrete HMM. This strongly indicates that appropriate feature is very important if continuous probability density function, especially diagonal covariance assumption, is used. If assumption is inappropriate, maximum likelihood estimation

will only maximize the *wrong* assumption.

Although more than 29% error reduction has been achieved for 18th order LPC analysis using diagonal covariance assumption, the last results with the discrete HMM (bilinear transformed cepstrum, 88.3%) and the SCHMM (18th order cepstrum, 88.6%) are about the same. This suggest that bilinear transformation is helpful for recognition, but have correlated coefficients, which is inappropriate to the diagonal Gaussian assumption. It can be expected that with the full covariance SCHMM and bilinear transformed cepstral data, even better recognition accuracy can be obtained.

4. CONCLUSIONS

Semi-continuous hidden Markov models based on multiple vector quantization codebooks take the advantages of both the discrete HMM and continuous HMM. With the SCHMM, it is possible to model a mixture of a large number of probability density functions with a limited amount of training data and computational complexity. Robustness is enhanced by using multiple codewords and multiple codebooks for the semi-continuous output probability representation. In addition, the VQ codebook itself can be adjusted together with the HMM parameters in order to obtain the optimum maximum likelihood of the HMM. A unified modeling approach can therefore achieve an optimum HMM parameters and VQ codebook parameters combination. From the continuous HMM point of view, the SCHMM can be considered as a special form of continuous mixture HMM with tied mixture continuous density functions. Because of the binding of the continuous density functions, in the SCHMM, the number of free parameters and computational complexity are reduced in comparison to the continuous mixture HMM while retaining the modeling powers of continuous HMM with a mixture of a large number of probability density functions.

The applicability of the continuous mixture HMM or the SCHMM relies on appropriately chosen acoustic parameters and assumption of the continuous probability density function. Acoustic features must be well represented if diagonal covariance is applied to the Gaussian probability density function. This is strongly indicated by the experimental results based on the bilinear transformed cepstrum and cepstrum. With bilinear transformation, high frequency components are compressed in comparison to low frequency components [Shikano86a, Lee88a]. Such a transformation converts the linear frequency axis into a mel-scale-like one. The discrete HMM can be substantially improved by bilinear transformation. However, bilinear transformation introduces strong correlations, which is inappropriate for the diagonal Gaussian assumption modeling. Using the cepstrum without bilinear transformation, the diagonal SCHMM can be substantially improved in comparison to the discrete HMM. However, if the bilinear transformed cepstrum is used, the recognition accuracy of the diagonal SCHMM is only slightly higher than the discrete HMM.

All experiments conducted here were based on only 200 generalized triphones; as smoothing can play a more important role in those less-well-trained models, more improvement can be expected for 1000 generalized triphones (where the word accuracy for the discrete HMM is 91% with bilinear transformed data). In addition, removal of diagonal covariance assumption by use of full covariance can be expected to further improve recognition accuracy [Doddington89a]. Regarding use of full covariance, the SCHMM has a distinctive advantage. Since Gaussian probability density functions are tied to the VQ codebook, by chosing M most significant codewords, computational complexity can be several order lower than the conventional continuous mixture HMM while maintaining the modeling power of large mixture components.

Experimental results have clearly demonstrated that the SCHMM offers improved recognition accuracy in comparison to both the discrete HMM (a error reduction of 29%) and the continuous mixture HMM (a error reduction of 41%) in speaker-independent continuous speech recognition. We conclude that the SCHMM is indeed a powerful technique for modeling non-stationary stochastic processes with multi-modal probabilistic functions of Markov chains.

ACKNOWLEDGEMENTS

We would like to thank Professor M.A. Jack, University of Edinburgh, Professor R. Reddy, CMU, and Professor D.T. Fang, Tsinghua University for their help and insight shared in this research.

References

Baum70a.

Baum, L.E., Petrie, T., Soules, G., and Weiss, N., "A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.* **41** pp. 164-171 (1970).

Baum72a.

Baum, L.E., "An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes," *Inequalities* **3** pp. 1-8 (1972).

Bellegarda89a.

Bellegarda, J. and Nahamoo, D., "Tied mixture continuous parameter models for large vocabulary isolated speech recognition," *ICASSP 89, Glasgow, Scotland* , (1989).

Brown87a.

Brown, P.F., "Acoustic-phonetic modeling problem in automatic speech recognition," *Ph.D. Thesis, Dept. of Computer Science, Carnegie-Mellon University*, (1987).

Chow87a.

Chow, Y.L., Dunham, M.D., Kimball, O.A., Kranser, M.A., Kubala, G.F., Makhoul, J., Price, P.J, Roucos, S., and , R.M. Schwartz, "BYBLOS: The BBN continuous speech recognition system," *IEEE ICASSP 87, Dallas, USA*, pp. 89-92 (1987).

Dempster77a.

Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum-likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc. Ser. B (methodological)* **39** pp. 1-38 (1977).

Doddington89a.

Doddington, G.R., "Phonetically sensitive discriminants for improved speech recognition," *ICASSP 89, Glasgow, Scotland*, (1989).

Gupta87a.

Gupta, V.N., Lennig, M., and Mermelstein, P., "Integration of Acoustic Information in a Large Vocabulary Word Recognizer," *IEEE ICASSP*, pp. 697-700 (1987).

Huang88a.

Huang, X.D. and Jack, M.A., "Hidden Markov modelling of speech based on a semi-continuous model," *IEE Electronics Letters* **24**(1) pp. 6-7 (1988).

Huang89a.

Huang, X.D. and Jack, M.A., "Unified modeling of vector quantization and hidden Markov model using semi-continuous hidden Markov models," *IEEE ICASSP 89, Glasgow, Scotland*, (1989).

Huang89b.

Huang, X.D. and Jack, M.A., "Semi-continuous hidden Markov models for speech recognition," *Computer Speech and Language, to be published*, (1989).

Huang89c.

Huang, X.D., "Unified theory of vector quantization and hidden Markov modelling," *Ph.D. thesis proposal, Dept. of Electrical Engineering, University of Edinburgh*, (1989).

Jelinek76a.

Jelinek, F., "Continuous speech recognition by statistical methods," *Proceedings of IEEE* **64** pp. 532-556 (1976).

Jelinek80a.

Jelinek, F. and Mercer, R.L., "Interpolated estimation of Markov source parameters from sparse data," *Proceedings of the workshop on pattern recognition in practice, Amsterdam, The Netherlands: North-Holland*, (1980).

Jelinek85a.

Jelinek, F., "The development of an experimental discrete dictation recognizer," *Proceedings of IEEE* **73** pp. 1616-1624 (1985).

Juang85a.

Juang, B.H., "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chain," *AT&T Technical Journal* **64** pp. 1235-1249 (1985).

Lee88a.

Lee, K.F., "Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system," *Ph.D. thesis, Dept. of C. Science, Carnegie-Mellon University*, (1988).

Lee89a.

Lee, K.F., Hon, H.W., and Reddy, R., "The SPHINX speech recognition system," *IEEE ICASSP 89, Glasgow, Scotland*, (1989).

Makhoul85a.

Makhoul, J., Roucos, S., and Gish, H., "Vector quantisation in speech coding," *Proceedings of IEEE* 73 pp. 1551-1588 (1985).

Nishimura87a.

Nishimura, M. and Toshioka, K., "HMM-based speech recognition using multi-dimensional multi-labeling," *IEEE ICASSP 87, Dallas, USA*, pp. 1163-1166 (1987).

Paul89a.

Paul, D., "The Lincoln continuous speech recognition system: recent developments and results," *DARPA 1989 Feb meeting*, (1989).

Poritz86a.

Poritz, A.B. and Richter, A.G., "On hidden Markov models in isolated word recognition," *IEEE ICASSP 86, Tokyo, Japan*, pp. 705-708 (1986).

Rabiner85a.

Rabiner, L.R., Juang, B.H., Levinson, S.E., and , M.M. Sondhi, "Recognition of isolated digits using hidden Markov models with continuous mixture densities," *AT&T Technical Journal* **64** pp. 1211-1234 (1985).

Rabiner86a.

Rabiner, L.R. and Juang, B.H., "An introduction to hidden Markov models," *IEEE ASSP Magazine*, pp. 4-16 (Jan. 1986).

Rabiner88a.

Rabiner, L.R., Wilpon, J.G., and Soong, F.K., "High Performance Connected Digit Recognition Using Hidden Markov Models," *IEEE ICASSP*, (1988).

Redner84a.

Redner, R.A. and Walker, H.F., "Mixture densities, maximum likelihood and the EM algorithm," *SIAM review* **26** pp. 195-239 (1984).

Schwartz89a.

Schwartz, R., Kimball, O., Kubala, F., Feng, M., Chow, Y., Barry, C., and Makhoul, J., "Robust smoothing methods for discrete hidden Markov models," *IEEE ICASSP 89, Glasgow, Scotland*, (1989).

Shikano86a.

Shikano, K., "Evaluation of LPC spectral matching measures for phonetic unit recognition," *CMU Technical Report CMU-CS-86-108, Computer Science Dept*, (1986).

Tseng87a.

Tseng, H.P., Sabin, M., and Lee, E., "Fuzzy vector quantization applied to hidden Markov modeling," *IEEE ICASSP 87, Dallas, USA*, pp. 641-644 (1987).