FACTOR ANALYSIS OF ACOUSTIC FEATURES FOR STREAMED HIDDEN MARKOV MODELING

Chuan-Wei Ting and Jen-Tzung Chien

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, ROC
{cwting,chien}@chien.csie.ncku.edu.tw

ABSTRACT

This paper presents a new streamed hidden Markov model (HMM) framework for speech recognition. The factor analysis (FA) is performed to discover the common factors of acoustic features. The streaming regularities are governed by the correlation between features, which is inherent in common factors. Those features corresponding to the same factor are generated by identical HMM state. Accordingly, we use multiple Markov chains to represent the variation trends in cepstral features. We develop a FA streamed HMM (FASHMM) and go beyond the conventional HMM assuming that all features at a speech frame conduct the same state emission. This streamed HMM is more delicate than the factorial HMM where the streaming was empirically determined. We also exploit a new decoding algorithm for FASHMM speech recognition. In this manner, we fulfill the flexible Markov chains for an input sequence of multivariate Gaussian mixture observations. In the experiments, the proposed method can reduce word error rate by 36% at most.

Index Terms— factor analysis, Markov chain, streamed HMM, speech recognition

1. INTRODUCTION

Hidden Markov models (HMMs) [8] have been becoming the main stream approach for speech recognition for the past decades. Due to the powerfulness of HMMs in stochastic modeling, many approaches have been presented to deal with different limitations in HMM. In [5][6], the factorial hidden Markov model (FHMM) was proposed to model the loosely coupled random processes. Using FHMM, different processes were represented by individual HMM topologies through streams of Markov chains instead of simply adopting one stream for conventional HMM. This approach has achieved desirable performance in real-world applications [1][3]. In [3], the FHMM combining two HMM processes was employed for simultaneous recognition of utterances from multiple speakers. Similar structure was used in noisy speech recognition. FHMM with two streams of processes was adopted to capture the statistics of speech as well as noise [1][12]. The log-max approximation was performed to determine the stream of an observation [1][3][7].

Even though HMM and FHMM are successful for speech recognition, these methods are limited due to the assumptions that all features in a speech frame come from the same Markov chain, namely the state transition of different cepstral features happens at the same moment. This assumption shall result in the limitation that the dynamics of individual features cannot be properly characterized. For this reason, the streamed FHMM [6] was proposed to represent acoustic features separately by different streams or Markov chains for sub-vectors. For example, a full acoustic feature vector could be separated into three sub-vectors of cepstral features, energy features and delta features. Using the streamed FHMM, the input features between sub-vectors were distinct. The features within the sub-vector were significantly correlated. The streaming was determined artificially. However, in conventional HMM, if we only used a full covariance matrix to model the correlations among features. The dynamics such as state transition were not explicitly expressed.

In this work, we go beyond the streamed FHMM by modeling the correlation among features according to factor analysis (FA) principle. We perform FA projection of acoustic features and group the correlated features to form inputs for streams. In an extreme case, we can model individual feature via its own Markov chain. However, this case is impractical and computationally inefficient. Here, we balance the tradeoff between coarse and precise streams through the control of common factors. The dynamics or the moments of state transition are flexibly determined. We find the features with high correlation and use a separate Markov chain to modeling its statistics. In what follows, we first describe how FA is able to perform correlation analysis of cepstral features. In section 3, we present the topology of FA streamed HMM (FASHMM) and its solution to parameter estimation. In section 4, we report the experiments on recognizing TIDIGIT utterances and investigating different realizations of FASHMM. A decoding algorithm is described. In section 5, we draw the conclusions from this study.

2. CEPSTRAL FACTOR ANALYSIS

In speech recognition procedure, we usually extract the Mel-frequency cepstral coefficients (MFCCs) as features

and use them to characterize state emission for speech frames in HMM framework. When representing speech process by Markov chains, it is important to conduct the correlation analysis of cepstral features and discover the regularities of individual features. One single Markov chain shall not be sufficient to reflect the dynamics of acoustic features observed in a phone segment. Figure 1 shows an example of different cepstral coefficients of digit "4" sampled from a TIDIGIT database. We find that the cepstral coefficients make considerable changes at different time moments. The 1st MFCC and log-energy have similar transition time, but the movement of 4th MFCC is different from that of 1st MFCC and log-energy. The transitions in Markov chains or the boundaries of HMM states should be represented by the streamed HMM where the features in the same stream are similarly behaved and modeled by the shared Markov chain. In this study, we are presenting a factor analysis approach to capture the correlations among features and the regularities for state transitions.

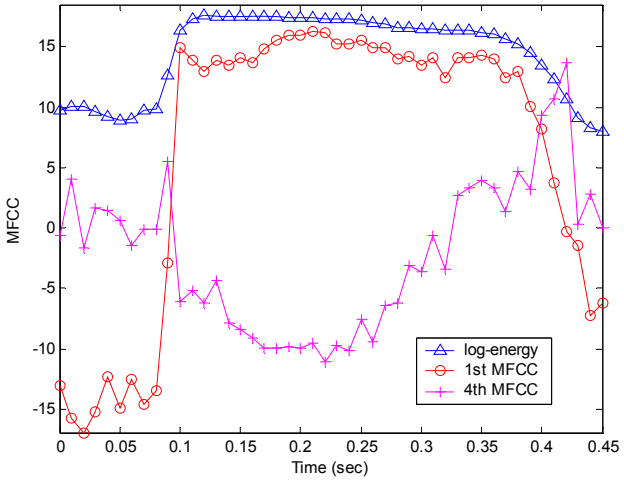

Figure 1: Dynamics of different cepstral features of TIDIGIT "4".

2.1. Factor Analysis

Statistical factor analysis (FA) is a machine learning approach to discovering the correlations inherent in observation data. FA was applied for front-end preprocessing of noisy speech signal [2]. It was also used to construct the HMM covariance matrix for speech recognition [9][10]. Different from the processing in signal domain [2] and model domain [9][10], we are applying FA approach both in feature domain and in model domain. We are detecting the correlations among features via FA method and merging these correlations in HMM modeling.

FA conducts data analysis of the multivariate observations using the common factors and the specific factors. For a D dimensional feature vector $\mathbf{y} = [y_1, \cdots y_D]^T$, the general form of FA model is given by

$$\mathbf{y} = \mathbf{W}_{\mathrm{f}}\mathbf{f} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ denotes $M \times 1$ common factor and $D \times 1$ specific factor and are independently Gaussian distributed by densities $N(0, I)$ and $N(0, \boldsymbol{\psi})$, respectively. The $D \times M$ matrix $\mathbf{W}_{\mathrm{f}}$ is called the factor loading matrix with each entry recording the correlation between feature y_d and common factor f_m. We should estimate FA parameters $\{\mathbf{W}_{\mathrm{f}}, \mathbf{f}, \boldsymbol{\varepsilon}\}$. Using the maximum likelihood estimation [11], we can find a $D \times D$ transformation matrix $\mathbf{W}$ by FA procedure. By dividing $\mathbf{W}$ into two complementary sub-matrices $\mathbf{W} = [\mathbf{W}_{\mathrm{f}}\ \mathbf{W}_{\mathrm{r}}]$, we rewrite (1) by $\mathbf{y} = \mathbf{y}_{\mathrm{f}} + \mathbf{y}_{\mathrm{r}} = \mathbf{W}_{\mathrm{f}}\mathbf{f} + \mathbf{W}_{\mathrm{r}}\mathbf{r}$ and accordingly estimate $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ by $\mathbf{f} = \mathbf{W}_{\mathrm{f}}^T \mathbf{y}_{\mathrm{f}}$ and $\boldsymbol{\varepsilon} = \mathbf{y}_{\mathrm{r}} = \mathbf{W}_{\mathrm{r}}\mathbf{r}$, respectively. Here, we are motivated to interpret different dynamics in acoustic features by common factors and specific factors. The features with high correlation should contribute the specific factor loading weights. We use separate Markov chains to model the dynamics caused by the common factor and the residual factor. We construct a new graphical model to delicately express complex dynamics in a sequence of feature vectors. The conventional state transition probability tied by all features in a speech frame is improved by using multiple transition probabilities activated by the common factors and the residual factors.

2.2. Rotation of Loading Matrix

Common factor is inherent to represent the tying of acoustic features with high correlations. There exists the phenomenon that one feature is correlated to several common factors. The estimated common factors shall be confusing. To determine common factors with good discriminability, we perform a rotation process for finding factor loading matrix. There are two rotation approaches for FA. One is orthogonal rotation and the other is oblique rotation. Orthogonal rotation claims that the resulting factors are uncorrelated after transformation while oblique rotation does not guarantee this orthogonal property. Basically, the orthogonal rotation is desirable to attain discrimination of common factors. Hence, we apply the Varimax rotation [11]. Using this approach, we yield the rotated factor loading matrix by

$$\mathbf{H} = \mathbf{W}\boldsymbol{\Gamma}, \tag{2}$$

where $\boldsymbol{\Gamma}$ is a $D \times D$ orthogonal matrix and $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = I_D$. $\mathbf{H} = \{h_{ij}\}$ represents the rotated loading matrix. Varimax rotation assures that the variance in each row vector of rotated loading matrix is maximized. Let

$$q_i = \sum_{j=1}^{D} h_{ij}^2, \ i = 1, \cdots, D. \tag{3}$$

Then $\mathbf{H}$ can be obtained by maximizing

$$\sum_{j=1}^{D}\left\{\left[\sum_{i=1}^{D}\left(\frac{h_{ij}^2}{q_i}\right)^2 \Big/ D\right] - \left[\sum_{i=1}^{D}\left(\frac{h_{ij}^2}{q_i}\right) \Big/ D\right]^2\right\}. \tag{4}$$

In Table 1, we display the elements of factor loading matrix before and after the orthogonal rotation. Here, we use the same data as that used in Figure 1. Before rotation, the first common factor captures high correlation in the 1st MFCC and log-energy. This is matching the result shown in Figure 1. After rotation, the discriminability between two factors is increased. For example, the correlations between 4th MFCC and 1st factor and 2nd factor are -0.312 and -0.724, respectively. After rotation, we increase the difference of correlations. The correlations become 0.266 and 0.791.

Table 1: Comparison of some elements of $\mathbf{W}$ and $\mathbf{H}$.

(a)	1st Factor	2nd Factor
1st MFCC	0.842	0.011
4th MFCC	-0.312	-0.724
log-energy	0.896	0.120

(b)	1st **Rotated** Factor	2nd **Rotated** Factor
1st MFCC	-0.892	-0.004
4th MFCC	0.266	0.791
log-energy	-0.933	-0.135

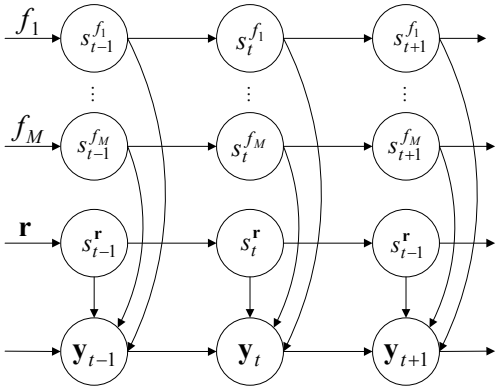

Figure 2: Topology of FASHMM.

3. FA STREAMED HMM

Using FA, the processes of observed features and hidden states are represented by common factors and residual factors. We intend to use separate Markov chain to model the movements of acoustic features corresponding to different factors. Figure 2 shows new topology based on the proposed FA streamed HMM (FASHMM). Markov chains are driven by common factors $\{f_1,\cdots,f_M\}$ and residual factor $\mathbf{r}$. At each frame t, we use several states to generate the feature vector $\mathbf{y}_t$. This is similar to the topology of FHMM [5]. FHMM was seen as a Bayesian belief network which was composed of more than one stream. Markov chain at each stream was used to characterize the dynamics of feature vector. Nevertheless, the streams in FHMM had the inputs of observed features instead of those of common factors and residual factors used in FASHMM.

3.1. Survey of Different HMMs

In standard HMM, the joint probability of observation sequence $Y=\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_T\}$ and state sequence $S=\{s_1,s_2,\cdots,s_T\}$ was represented by

$$p(S,Y)=p(s_1)p(\mathbf{y}_1\mid s_1)\prod_{t=2}^{T}p(s_t\mid s_{t-1})p(\mathbf{y}_t\mid s_t), \quad (5)$$

where $\mathbf{y}_t$ was a $D\times 1$ feature vector. Using FHMM, the state at time t was extended to M states, i.e. $s_t=s_t^{(1)},\cdots,s_t^{(m)},\cdots,s_t^{(M)}$. FHMM was also related to the multi-stream HMM [4]. In multi-stream HMM, the likelihoods at different streams were combined at subword level. This was different from FHMM where the combination was done at frame level. Using FHMM, the state transition of one stream was constrained to be independent to that of other streams. The state transition probability was given by

$$p(s_t\mid s_{t-1})=\prod_{m=1}^{M}p(s_t^{(m)}\mid s_{t-1}^{(m)}). \quad (6)$$

Also, the calculation of likelihood function $p(\mathbf{y}_t\mid s_t)$ involved M states. The likelihood function was defined by a Gaussian density with a common covariance matrix and a mean vector which was a linear combination of means for a HMM state [5]

$$p(\mathbf{y}_t\mid s_t)=(2\pi)^{-D/2}\mid\Sigma\mid^{-1/2}\times\exp\left\{-\frac{1}{2}(\mathbf{y}_t-\sum_{m=1}^{M}\mu_m)^T\Sigma^{-1}(\mathbf{y}_t-\sum_{m=1}^{M}\mu_m)\right\}, \quad (7)$$

where μ_m was the mean of state $s_t^{(m)}$ and Σ was a shared covariance matrix. According to (5)-(7), EM algorithm was carried out for maximum likelihood estimation. In [5], an efficient approximation algorithm was proposed for parameter estimation for FHMM.

Although FHMM considered different dynamics, all features were "tied" together to model the dynamics in each stream. Namely, all features were generated by the same Markov chain. The state transition of these features occurred at the same time. Gaussian process could not be properly modeled. For this reason, the streamed FHMM (SFHMM) [6] was proposed through partitioning the observed feature vector into several streams of sub-vectors. For example, the acoustic feature vector could be divided into three subvectors of cepstral coefficients and their first and second delta coefficients. These subvectors were grouped into different streams. Using SFHMM, the grouping was artificially determined. In this study, we present FASHMM framework where the streaming of acoustic features is driven by the factor analysis principle.

3.2. FASHMM

According to FA method, the common factor f_m is associated with some features, which are highly correlated. This factor f_m is viewed as the input of FASHMM in stream m. In this case, we don't need to put original features as the inputs as performed in FHMM [5]. Correlated features are grouped together to form a univariate common factor f_m for a stream and shared by the same FA parameters. Since the input of each stream is a

univariate variable, the model size of FASHMM is smaller than that of FHMM. In addition to streams of common factors, we also create one stream for residual vector $\mathbf{r}$ so as to fulfill FA spirit. Here, we extend (1) and represent D dimensional feature vector by

$$\begin{aligned}\mathbf{y} &= \mathbf{W}_{\mathrm{f}}\mathbf{f} + \mathbf{W}_{\mathrm{r}}\mathbf{r} \\ &= [\mathbf{w}_{\mathrm{f}_1} \quad \mathbf{w}_{\mathrm{f}_2} \quad \cdots \quad \mathbf{w}_{\mathrm{f}_M} \quad \mathbf{W}_{\mathrm{r}}][f_1 \quad f_2 \quad \cdots \quad f_M \quad \mathbf{r}]^T,\end{aligned} \tag{8}$$

where $\mathbf{w}_{\mathrm{f}_m}$ is the mth column vector of factor loading matrix $\mathbf{W}_{\mathrm{f}}$, $\mathbf{W}_{\mathrm{r}}$ is a $D\times(D-M)$ sub-matrix of transformation matrix and $\mathbf{r}$ is a $(D-M)\times 1$ residual vector. In FASHMM, the state transition probability is expressed by

$$\begin{aligned}p(s_t \mid s_{t-1}) &= p(s_t^{f_1}, s_t^{f_2}, \cdots, s_t^{f_M}, s_t^{\mathrm{r}} \mid s_{t-1}^{f_1}, s_{t-1}^{f_2}, \cdots, s_{t-1}^{f_M}, s_{t-1}^{\mathrm{r}}) \\ &= p(s_t^{\mathrm{r}} \mid s_{t-1}^{\mathrm{r}}) \cdot \prod_{m=1}^{M} p(s_t^{f_m} \mid s_{t-1}^{f_m})\end{aligned} \tag{9}$$

Different streams are assumed to be independent. The probability of observation vector $\mathbf{y}_t$ at state s_t is yielded by

$$\begin{aligned}p(\mathbf{y}_t \mid s_t) &= p(f_1, f_2, ..., f_M, \mathbf{r} \mid s_t^{f_1}, s_t^{f_2}, ..., s_t^{f_M}, s_t^{\mathrm{r}}) \\ &= p(\mathbf{r} \mid s_t^{\mathrm{r}}) \cdot \prod_{m=1}^{M} p(f_m \mid s_t^{f_m}) \\ &= \left\{\sum_{k=1}^{K_r} c_k^{\mathrm{r}} \cdot N(\boldsymbol{\mu}_k^{\mathrm{r}}, \boldsymbol{\Sigma}_k^{\mathrm{r}})\right\} \cdot \prod_{m=1}^{M}\left\{\sum_{k=1}^{K_m} c_k^m \cdot N(\mu_k^m, {\sigma_k^m}^2)\right\},\end{aligned} \tag{10}$$

where $(\mu_k^m, {\sigma_k^m}^2)$ and $(\boldsymbol{\mu}_k^{\mathrm{r}}, \boldsymbol{\Sigma}_k^{\mathrm{r}})$ are the kth Gaussian parameters for common factor f_m and residual vector $\mathbf{r}$, respectively. Given the initial state probability $\{p(s_1 = i) = \pi_i\}$ and the state transition probability $\{p(s_t = j \mid s_{t-1} = i) = a_{ij}\}$, we calculate the joint likelihood

$$\begin{aligned}p(S,Y) &= p(S)p(Y \mid S) = \left\{\pi_i^{\mathrm{r}} \cdot \prod_{m=1}^{M} \pi_i^{f_m}\right\} \\ &\cdot \left\{\prod_{t=1}^{T}\left[a_{ij}^{\mathrm{r}} \cdot \prod_{m=1}^{M} a_{ij}^{f_m}\right]\right\}\left\{\prod_{t=1}^{T}\left[p(\mathbf{r} \mid s_t^{\mathrm{r}}) \cdot \prod_{m=1}^{M} p(f_m \mid s_t^{f_m})\right]\right\}.\end{aligned} \tag{11}$$

Parameters $\lambda = \{\pi_i^{\mathrm{r}}, \pi_i^{f_m}, a_{ij}^{\mathrm{r}}, a_{ij}^{f_m}, c_k^{\mathrm{r}}, c_k^m, \boldsymbol{\mu}_k^{\mathrm{r}}, \mu_k^m, \boldsymbol{\Sigma}_k^{\mathrm{r}}, {\sigma_k^m}^2\}$ are constructed. In model training procedure, we perform the maximum likelihood (ML) estimation of FASHMM parameters λ through the expectation-maximization algorithm. The expectation function of new estimate λ' given current estimate λ is determined by

$$\begin{aligned}Q(\lambda' \mid \lambda) &= \sum_S p(S \mid Y, \lambda) \log p(S, Y \mid \lambda') = \sum_S p(S \mid Y, \lambda) \\ &\cdot \left\{\left[\log \pi_i'^{\mathrm{r}} + \sum_{m=1}^{M} \log \pi_i'^{f_m}\right] + \left[\sum_{t=1}^{T}\left(\log a_{ij}'^{\mathrm{r}} + \sum_{m=1}^{M} \log a_{ij}'^{f_m}\right)\right]\right. \\ &\left. + \left[\sum_{t=1}^{T}\left(\log p'(\mathbf{r} \mid s_t^{\mathrm{r}}) + \sum_{m=1}^{M} \log p'(f_m \mid s_t^{f_m})\right)\right]\right\}.\end{aligned} \tag{12}$$

By maximizing $Q(\lambda' \mid \lambda)$ with respect to λ', we can find the closed-form solutions given below

$$\mu_{ik}^{f_m} = \frac{\sum_{t=1}^{T} \gamma_t^{f_m}(i,k) f_{m,t}}{\sum_{t=1}^{T} \gamma_t^{f_m}(i,k)}, \quad m = 1, \cdots, M, \quad \boldsymbol{\mu}_{ik}^{\mathrm{r}} = \frac{\sum_{t=1}^{T} \gamma_t^{\mathrm{r}}(i,k)\mathbf{r}_t}{\sum_{t=1}^{T} \gamma_t^{\mathrm{r}}(i,k)} \tag{13}$$

$$(\sigma_{ik}^{f_m})^2 = \frac{\sum_{t=1}^{T} \gamma_t^{f_m}(i,k)(f_{m,t} - \mu_{ik}^{f_m})^2}{\sum_{t=1}^{T} \gamma_t^{f_m}(i,k)}, \quad m = 1, \cdots, M,$$

$$\boldsymbol{\Sigma}_{ik}^{\mathrm{r}} = \frac{\sum_{t=1}^{T} \gamma_t^{\mathrm{r}}(i,k)(\mathbf{r}_t - \boldsymbol{\mu}_{ik}^{\mathrm{r}})(\mathbf{r}_t - \boldsymbol{\mu}_{ik}^{\mathrm{r}})^T}{\sum_{t=1}^{T} \gamma_t^{\mathrm{r}}(i,k)} \tag{14}$$

where $\gamma_t^{f_m}(i,k)$ and $\gamma_t^{\mathbf{r}}(i,k)$ are the occupation probabilities for state $s_t = i$ and mixture k at different stream m and time t.

At each time moment, we accumulate log-likelihood score for each stream. A simple approach is to set equivalent stream weight. But, in FA model, one common factor shall represent more than one feature component. The stream weight should be adaptive. In this work, we calculate the stream weights from the columns of the rotated factor loading matrix $\mathbf{H} = \{h_{ij}\}$. The weight function ω_j in stream j is empirically determined by taking absolute values of entries $\{h_{ij}\}$ and computing the ratio of those values in column j over all columns.

3.3. Implementation Procedure

Typically, using the proposed FASHMM, we are able to incorporate multiple Markov chains to model a multivariate input sequence of features for speech recognition. We are not only extracting the salient common factors but also modeling the dynamics of common factors containing highly correlated features. In training procedure, similar to standard HMM, we first collect acoustic feature vectors corresponding to different words through Viterbi decoding algorithm and then estimate the word dependent factor loading matrix $\mathbf{W}$. After finding $\mathbf{W}$, the common factors $\{f_m\}$ and residual vector $\mathbf{r}$ are estimated and viewed as the inputs for estimation of HMM parameters at each stream. Because more than one stream is involved, the computational complexity of FASHMM is higher than that of standard HMM. This process is engaged in training phase as well as test phase. In test phase, a new decoding algorithm is developed for fulfilling FASHMM and shown in Figure 3. The additional computation is spent on finding the state transition boundaries for each stream and each word. Each stream has its own Markov chain. Let the number of states in stream j be denoted by M_j. The total

amount of states in a word is given by $\sum_{j=1}^{M+1} M_j$. Notably, before calculating log-likelihood $l(t,w,s,j)$ and accumulated log-likelihood $\delta(t,w,s)$, we should perform FA procedure of extracting common factors and residual factors using word-dependent transformation matrix. Also, the search of optimal states in different streams is much more complicated compared to that only considering one stream in standard HMM.

1. Let $\delta(t,w,s)$ denote the accumulated log-likelihood at time t and state s of word w and $l(t,w,s,j)$ denote the log-likelihood at jth stream.
2. for $t = 2,\cdots,T$
   for $w = 1,\cdots,W$

$$\delta(t,w,1) = \sum_{j=1}^{M+1} l(t,w,1,j) + \max\left\{\sum_{j=1}^{M+1} l(t-1,w,1,j), \sum_{j=1}^{M+1} l(t-1,w^*,s(w^*),j)\right\}$$

   where $s(w^*)$ denotes the maximum number of state combinations occurring during word w^*, $w^* = 1,\cdots W$. State $s = 1$ means that all streams are at their first state in word w.
   for $s = 2,\cdots,s(w)$

$$\delta(t,w,s) = \sum_{j=1}^{M+1} \omega_j l(t,w,s,j) + \max\left\{l(t-1,w,s,j), \max_{s\in s^-}[\delta(t-1,w,s^-)]\right\}$$

   where s^- is the number of all possible states emitting to state s, ω_j is the jth stream weight.
   end
   end
   end
3. Back trace the best path with the maximum accumulated log-likelihood $\delta(t,w,s)$.

Figure 3: FASHMM decoding algorithm.

4. EXPERIMENTS

4.1. Experimental Setup

In the experiments, we carried out the proposed FASHMM for connected digit recognition using TIDIGIT database. The vocabulary had 11 words containing digits from "1" to "9" and digit "0" with two pronunciations "zero" and "oh". Each speaker uttered isolated digits and connected digit strings with up to 7 digits. We used adult utterances from 111 males and 114 females. There were 1700 training utterances and 8000 test utterances. All utterances were down sampled to 8 kHz with 16 bit resolution. We extracted 39 dimensional feature vectors consisting of 12 MFCCs and one log-energy, and their first and second derivatives. In standard HMM, we fixed 16 states for each digit and one shared state for all silence segments. Number of mixture components in a HMM state was four at most. In the evaluation, we changed the number of states for different streams. In FA procedure, we used maximum likelihood approach to estimate factor loading matrix and then find common factors and residual factors [11]. Considering the computational efficiency, we simplified the model topology to a two-streamed FASHMM. The first stream corresponded to the common factors **f** and the other stream corresponded to the residual vector **r**. In this way, the first stream weight was determined by integrating the stream weights of M common factors into one shared weight.

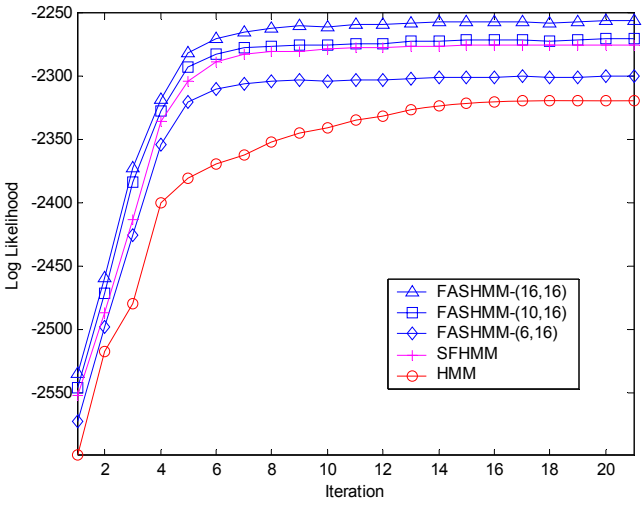

Figure 4: Evaluation of log likelihood function.

4.2. Evaluation of Likelihood Function

In what follows, we calculate the accumulated log-likelihood to evaluate the goodness-of-fit performance of training data using different methods. We compare the standard HMM, the streamed FHMM (SFHMM) [6] and the proposed FASHMM. FAHMM is not focused on the streamed modeling so that we don't include it for comparison. In SFHMM, we used three features of log-energy, delta log-energy and delta delta log-energy in the first stream. The remaining 36 features of 12 MFCCs and their first and second derivatives were collected in the second stream. In the implementation of FASHMM, we also used 3 and 36 acoustic features in the first and the second streams, respectively. The assignment of acoustic features to two streams was determined automatically by FA principle. Number of states was varied at two streams. In Figure 4, we fixed 16 states in the second stream and used 6, 10 and 16 states in the first stream. We can see that SFHMM-(16, 16) and FASHMM attain higher likelihood score that standard HMM. In case of using 10 and 16 states in the first stream, FASHMM has better likelihood score compared to SFHMM.

4.3. Recognition Results

In evaluation of speech recognition, we compare the word error rate (WER) of using standard HMM, SFHMM, and different realizations of FASHMM in Table 2. Error rate reduction is reported in the last column. In realizations of FASHMM, we change the number of features and the number of states at each stream. The streamed HMMs using SFHMM and FASHMM outperform the conventional HMM without streaming process. Also, different realizations of FASHMM do decrease the word error in comparison with SFHMM. The best performance is obtained by setting 5 and 34 features and 6 and 16 states for the first stream and the second stream, respectively. Error rate reduction is 35.8% at most. Attractively, this realization does not only attain the lowest word error but also involve the smallest model complexity among the streamed HMM methods. Further, the number of parameters is almost the same as the conventional HMM in case of 9 and 30 features and 12 and 9 states for 1st stream and 2nd stream, respectively. In this case, the error rate reduction is 13.6% which is still significant.

Table 2: Comparison of word error rates (%) using different methods. * and ** denote that FASHMM has significant improvement compared to HMM and SFHMM, respectively, through the *t* test at a significance level $\alpha = 0.05$.

	No. of features at each stream	No. of states at each stream	WER (%)	Error rate reduction
HMM	39	16	1.702	--
SFHMM	(3, 36)	(16, 16)	1.539	9.6 %
FASHMM	(3, 36)	(16, 16)	1.130	33.6 % * **
		(10, 16)	1.130	33.6 % * **
		(6, 16)	1.153	32.2 % * **
	(5, 34)	(16, 16)	1.122	34.0 % * **
		(10, 16)	1.092	35.8 % * **
		(6, 16)	1.137	33.2 % * **
	(7, 32)	(5, 11)	1.389	18.4 % * **
	(9, 30)	(12, 9)	1.470	13.6 % *
	(10, 29)	(16, 16)	1.266	25.6 % * **
		(10, 16)	1.221	28.2 % * **
		(6, 16)	1.328	22.0 % * **

5. CONCLUSION

We have presented the FA approach to extract the common factors and the residual factors in acoustic features and separate the Markov chains for these factors. Accordingly, we were able to represent the sophisticated dynamics in stochastic process of speech signal. A new topology of FA streamed HMM was proposed. By maximizing the likelihood of training data, we estimated the FASHMM parameters using those features in FA space. From the experimental results, we obtained the desirable recognition performance on the connected digit recognition using TIDIGIT database. In the future, we are investigating the issue of model complexity, namely the optimal selection of the numbers of stream, state and mixture component. We are applying the proposed FASHMM framework for large vocabulary continuous speech recognition.

6. REFERENCES

[1] A. Betkowska, K. Shinoda, and S. Furui, "FHMM for robust speech recognition in home environment", *International Symposium on Large-Scale Knowledge Resources (LKR2006)*, pp.129-132, 2006.

[2] J.-T. Chien and C.-W. Ting, "Subspace modeling and selection for noisy speech recognition", *Proc. of International Conference on Spoken Language Processing (INTERSPEECH)*, pp. 789-792, 2006.

[3] A. N. Deoras and M. H. Johnson, "A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel", *Proc. of International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, vol. 1, pp.861-864, 2004.

[4] S. Dupont and H. Bourlard, "Using multiple time scales in a multi-stream speech recognition system", *Proc. of 5th European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 1, pp. 3-6, 1997.

[5] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models", *Machine Learning*, 29, pp. 245-275, 1997.

[6] B. Logan and P. Moreno, "Factorial HMMs for acoustic modeling", *Proc. of International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pp.813-816, 1998.

[7] A. Nadas, D. Nahamoo and M. A. Picheny, "Speech recognition using noise-adaptive prototypes", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 37, no. 10, pp. 1495-1503, 1989.

[8] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall, 1993.

[9] A.-V. I. Rosti, M. J. F. Gales, "Factor analyzed hidden Markov models for speech recognition", *Computer Speech and Language*, vol. 18, no. 2, pp. 181-200, 2004.

[10] L. K. Saul and M. G. Rahim, "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition", *IEEE Transaction on Speech and Audio Processing*, vol. 8, no. 2, pp. 115-125, 2000.

[11] M. S. Srivastava, *Methods of Multivariate Statistics*, John Wiley & Sons, 2002.

[12] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space", *Proc. of International Conference on Spoken Language Processing (INTERSPEECH)*, pp.89-92, 2006.