ABSTRACT

In this paper, the reduction method of number of parameters in the segmental-feature HMM (SFHMM) can be considered. It is reported that the SFHMM shows better results than conventional HMM in the previous studies. However, its number of parameters is greater than that of HMM. Therefore, there is a need for new approach that reduces the number of parameters. The trajectories are used for the acoustic features of the SFHMM. In general, trajectory can be separated by the trend and location. Since the trend means the variation of segmental features and occupies the large portion of SFHMM, if the trend is shared, the number of parameters of SFHMM maybe decreases. We consider the trend tying of segmental features by the quantization. The experiments are performed on TIMIT corpus to examine the effectiveness of the trend tying.

1. INTRODUCTION

HMM has been widely used in various areas for a long time due to its easy implementation, flexible modeling capability, and high performance. However, it is reported that the HMM does not effectively represent the temporal dependency of speech signals because of weakness of its assumptions. Various studies have been done to mitigate the weakness by adopting segmental models[3, 5] or trajectory approaches [4, 6]. These models use the segmental feature rather than frame feature or the regression function of frame features. In previous work, segmental-feature HMM (SFHMM) was presented in which the input speech signals are modeled based on the fixed length segment to take some advantages on noisy environment and to easily be implemented. In the SFHMM, the input speech signals, which are transformed to segmental feature based on well-known speech features, are transferred to the classification module. Thus, in this section, we describe the segmental feature and likelihood which are the core of the classification module.

2. SEGMENTAL-FEATURE HMM

In the previous work, the proposed SFHMM adopted the parametric method and is modeled based on the fixed length segment to take some advantages on noisy environment and to easily be implemented. In the SFHMM, the input speech signals, which are transformed to segmental feature based on well-known speech features, are transferred to the classification module. Thus, in this section, we describe the segmental feature and likelihood which are the core of the classification module.

2.1. Segmental feature

Deng [1] proposed a parametric approach for a non-stationary state HMM where polynomial trend functions are used as time-varying means of the output Gaussian distributions in the HMM states. In another trajectory method, Gish and Ng [2] modeled each feature dimension of a speech segment as a polynomial regression function. These features are called segmental features because the features are extracted from segments that correspond to the set of frames. In contrast to Gish’s approach, the features of SFHMM are based on the fixed length segment. In SFHMM, to express the fixed segment with time indexes, the segmental features can be expressed as

\[ C_t = ZB_t + E, \]  

where \( C_t \) and \( B_t \) are the speech segment and trajectory coefficients at time \( t \). In this equation, the segmental feature is extracted from the successive frame features using the
design matrix \( Z \). Each frame is represented by a \( D \) dimensional feature vector, \( Z \) and \( B_t \) are \( N \times R \) design matrix and \( R \times D \) trajectory coefficient matrix, respectively. \( E \) finally denotes residual error which is assumed to be independent and identically distributed.

For a given speech segment of \( N = 2M + 1 \) frames, the observed features are represented by the following matrix:

\[
C_t = Y_{t+M}^{-M} = \begin{bmatrix}
c_{t-M} \\
\vdots \\
c_t \\
c_{t+M}
\end{bmatrix}
\]

\[
c_t = [y_{t,1} \ldots y_{t,D}], \quad t-M \leq t \leq t+M. \tag{2}
\]

Because we consider the current frame feature on the center of a segment at time \( t \), the beginning and end of the speech segment may be overlapped with the neighboring segment at time \( t-1 \) or \( t+1 \). To represent the speech segment properly, the design matrix \( Z \) can be defined as

\[
Z = \begin{bmatrix}
1 & (-M/2M) & \ldots & (-M/2M)^{R-1} \\
\vdots & \vdots & \ddots & \vdots \\
1 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1 & (M/2M) & \ldots & (M/2M)^{R-1}
\end{bmatrix}
\]

\[
z_t = \begin{bmatrix}
1 \\
(\tau - t/2M) \\
\vdots \\
(\tau - t/2M)^{R-1}
\end{bmatrix}, \tag{3}
\]

where \( z_t \) is a row vector of \( Z \) along \( \tau \). Because \( Z \) represents the relative position from the current observation vector which is normalized by the segment length, the previous and following acoustic features of the current observation can be reflected in the trajectory. The trajectory coefficient matrix \( B_t \) is also defined as

\[
B_t = \begin{bmatrix}
b_{t,1}^t \\
\vdots \\
b_{t,R}^t
\end{bmatrix}
\]

\[
b_{t,i}^t = [b_{t,1,i} \ldots b_{t,D,i}], \quad 1 \leq i \leq R. \tag{4}
\]

Since errors are supposed to be independent and identically distributed, we obtain the trajectory coefficient matrix \( B_t \) by a linear regression or the following matrix equation:

\[
\hat{B}_t = [Z'Z]^{-1}Z'C_t, \tag{5}
\]

where ' means the matrix transpose.

With \( \hat{B}_t \) estimated, a goodness-of-fit measure \( \chi^2 \) can be obtained by summing the frame residual error over the segment at time \( t \),

\[
\chi_t^2 = \frac{1}{N} \sum_{\tau = t-M}^{t+M} (c_{\tau} - z_{\tau} \hat{B}_t)(c_{\tau} - z_{\tau} \hat{B}_t)'. \tag{6}
\]

The smaller the value of \( \chi^2 \), the better the data fitting. After the parameter estimation, the segment is represented by its trajectory coefficient matrix \( \hat{B}_t \) with \( \chi^2_t \).

\section{2. Segment likelihood}

It is assumed in SFHMM that extra-segmental variations are represented by Gaussian distributions with mean trajectories and their variances, while intra-segmental variations are defined as the estimation error of the trajectory in a segment.

Since the observation vectors \( C_t \) of SFHMM are represented as their unique trajectory \( ZB_t \) at time \( t \), the observation probability of \( C_t \) occurring at state \( s_i \) of model \( \lambda \) is specified by the equation

\[
P(C_t|s_i, \lambda) = P(Z\hat{B}_t|s_i, \lambda)P(C_t|Z\hat{B}_t, s_i, \lambda). \tag{7}
\]

Therefore, the output probability of a segment at time \( t \) for state \( j \) can be defined as

\[
b_j(C_t) = P(C_t|s_j, \lambda) = P(Z\hat{B}_t|ZB_j, \Sigma_j)P(C_t|Z\hat{B}_t), \tag{8}
\]

where \( \hat{B}_j \) and \( \Sigma_j \) are the trajectory model corresponding to state \( j \). In this equation, the extra-segmental probability and intra-segmental variation are defined as

\[
P(Z\hat{B}_t|ZB_j, \Sigma_j) = \prod_{\tau=t-M}^{t+M} \frac{1}{(2\pi)^{D/2} \Sigma_{\tau-t,i}^{-1}3/2} \exp \left\{ - \frac{1}{2} (z_{\tau}(\hat{B}_t - B_j)) \Sigma_{\tau-t,i}^{-1} (z_{\tau}(\hat{B}_t - B_j))' \right\} \tag{9}
\]

\[
P(C_t|Z\hat{B}_t) = \exp \left\{ - \frac{1}{2} \chi_t^2 \right\}; \tag{10}
\]

where \( \Sigma_j \) is the sequence of frame variance or common variance by the applied variance approach (time-varying or fixed variance).

\section{3. TREND TYING}

In SFHMM, a segment is modeled as a polynomial trajectory of fixed duration. The trajectory is obtained by the sequence of feature vectors of speech signals and can be divided by trend and location. The trend indicates the variation of consequent frame feature vectors, while the location points to the positional difference of trajectories.
The trajectory can be rewritten by the linear regression function. For each feature dimension, the following polynomial function is considered:
\[ y_{\tau,i} = b_{1,i} z_{\tau,1} + b_{2,i} z_{\tau,2} + b_{3,i} z_{\tau,3} + \cdots + b_{R,i} z_{\tau,R}, \]
where \( y_{\tau,i} \) means the cepstral features of \( i \) th dimension of \( \tau \) th frame in a segment, \( b_{r,i} \) is \( r \) th trajectory coefficient, and \( z_{\tau,r} \) is the element of the design matrix and indicates \( (\frac{z_{\tau,r}}{\sum_{M}})^{r-1} \).

From the above equation, we can find that the first column element of the design matrix is one, i.e. \( z_{\tau,1} = 1 \). Therefore, \( b_{1,i} \) means intercept on cepstral feature domain, while the remains of the equation are related to the segmental variation, e.g. trend. Consequently, if we share the remains of the equation in a trajectory representation, its trend can be shared with other trajectories.

In SFHMM, current observation vectors are placed at the center of segment. Therefore, \( b_{1,i} \) indicates the smoothed mid-point. If the above polynomial function is considered by matrix equation, the first row of the trajectory coefficient matrix \( B_1 \) means the \( D \)-dimensional location and the remains of rows are considered as the \((R-1) \times D\)-dimensional trend. To share the trend, at first, coefficient splitting is required for trend quantization. The new coefficient matrix \( \tilde{T}_i \) for the trend can be defined as follows:
\[
\tilde{T}_i = \begin{bmatrix} b_{2,i}^T \\ \vdots \\ b_{R,i}^T \end{bmatrix}. \tag{12}
\]

This trend coefficient is replaced by the nearest codeword in the trend quantization codebook. If the trend coefficient is replaced with trained codeword, new trajectory coefficient, which new trend \( \tilde{T}_i \) from codebook and location from \( B_1 \) are merged, is used for the input feature. In the estimation phase, the mean trend is also selected in trend codebook and the mean trajectory is modified to merge the adjusted trend and location.

Fig. 1 shows the flow of the trend tying process. In the proposed system, the trend, which is used for input feature and for training step, is always adjusted to the nearest codeword.

### 3.2. Trend quantization
Trend quantization algorithm is similar to that of well-known vector quantization. However, the distance measure has to be modified to compare two trends. The Euclidean distance is modified to reflect the trend characteristics as follows:
\[
D(\tilde{T}_i, \tilde{T}_j) = \frac{1}{N} \sum_{\tau=1}^{N} \{ \tilde{z}_r(\tilde{T}_i - \tilde{T}_j) \} \{ \tilde{z}_r(\tilde{T}_i - \tilde{T}_j) \}^T, \tag{13}
\]
where \( \tilde{z}_r \) is the row vector of design matrix which excludes the first column value, \( \tilde{T}_i, \tilde{T}_j \) are trend coefficient matrices.

### 4. EXPERIMENTAL RESULTS
SFHMMs were evaluated on 16-vowel classification task to examine the trend tying effect. The first 12 cosine coefficients together with the normalized log energy value, and their first derivatives were used to obtain the segmental feature of SFHMM or for inputs of conventional HMMs. SFHMM apply the fixed variance approach for segmental likelihood. For the experiment, we extracted the 16 vowels: 13 monophthongs /i/, /ih/, /ey/, /eh/, /ae/, /aa/, /ah/, /ao/, /ow/, /uw/, /uh/, /ux/, /er/ and three diphthongs /lay/, /oy/, /awl/. The vowels were excised, using the given phonetic segmentations, from the TIMIT corpus without any restrictions on the phonetic contexts of the vowels. After the tokens were extracted from
Table 1. Recognition results of the baseline system. (M means the number of mixtures)

<table>
<thead>
<tr>
<th>type</th>
<th>condition</th>
<th>M = 1</th>
<th>M = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>-</td>
<td>52.09</td>
<td>54.45</td>
</tr>
<tr>
<td>SFHMM</td>
<td>N = 3, R = 2</td>
<td>53.33</td>
<td>55.51</td>
</tr>
<tr>
<td></td>
<td>N = 3, R = 3</td>
<td>53.32</td>
<td>55.53</td>
</tr>
<tr>
<td></td>
<td>N = 5, R = 2</td>
<td>54.22</td>
<td>56.31</td>
</tr>
<tr>
<td></td>
<td>N = 5, R = 3</td>
<td>54.03</td>
<td>56.44</td>
</tr>
</tbody>
</table>

Table 2. Classification rates of SFHMM when the trend tying is used. (D represents the size of trend quantization codebook)

<table>
<thead>
<tr>
<th>type</th>
<th>condition</th>
<th>M = 1</th>
<th>M = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFHMM</td>
<td>N = 3, R = 2</td>
<td>52.55</td>
<td>54.78</td>
</tr>
<tr>
<td></td>
<td>N = 3, R = 3</td>
<td>52.27</td>
<td>54.28</td>
</tr>
<tr>
<td></td>
<td>N = 5, R = 2</td>
<td>53.62</td>
<td>54.20</td>
</tr>
<tr>
<td></td>
<td>N = 5, R = 3</td>
<td>51.85</td>
<td>53.60</td>
</tr>
<tr>
<td>SFHMM</td>
<td>N = 3, R = 2</td>
<td>52.73</td>
<td>55.06</td>
</tr>
<tr>
<td></td>
<td>N = 3, R = 3</td>
<td>52.21</td>
<td>54.38</td>
</tr>
<tr>
<td></td>
<td>N = 5, R = 2</td>
<td>53.22</td>
<td>54.88</td>
</tr>
<tr>
<td></td>
<td>N = 5, R = 3</td>
<td>52.60</td>
<td>54.53</td>
</tr>
<tr>
<td>SFHMM</td>
<td>N = 3, R = 2</td>
<td>52.69</td>
<td>54.99</td>
</tr>
<tr>
<td></td>
<td>N = 3, R = 3</td>
<td>52.16</td>
<td>53.69</td>
</tr>
<tr>
<td></td>
<td>N = 5, R = 2</td>
<td>53.22</td>
<td>54.92</td>
</tr>
<tr>
<td></td>
<td>N = 5, R = 3</td>
<td>52.26</td>
<td>53.64</td>
</tr>
</tbody>
</table>

training and testing sentences of TIMIT corpus, 41,429 tokens were employed for training and 11,606 tokens (in complete test corpus) were used for testing.

We conducted the experiments by changing the number of mixtures, segment length, regression order, and size of trend quantization codebook to compare the performance of SFHMM when the trend tying is used. The recognition results of the baseline systems are shown in Table 1, while the results of trend tied SFHMMs are in Table 2.

From the experimental result, we found that the proposed system did not outperform HMM. The performance of trend tied SFHMM shows the best results when the codebook size is 128. Regardless of the number of mixtures, the results of trend tied SFHMM shows better performances than conventional HMM in this case. However, the other cases, i.e., the codebook sizes are 64, 256, the performance is almost the same in both conditions of number of mixtures. It maybe caused by the following thought; if the small codebook is used, the SFHMM using the trend tying cannot reflect sufficiently the variations of speech signals, while the large trend codebook show the tendency of overfitting to the training data.

5. CONCLUSION

We have proposed the reduction method of number of parameters for SFHMM that uses segmental features represented by the polynomial regression function. The trajectory representing segmental feature can be separated by the trend and location; the trend shows the variation of segmental features, and the location indicates the reference point that is corresponding to segment mid-point value. The proposed method shares the trend of trajectories by quantization algorithm which is similar to vector quantization algorithm. To reveal the performance of trend tying in SFHMM, the experiments are done on TIMIT corpus. From the recognition results, the performance of the proposed approach is almost the same to that of conventional HMM. However, if the performance is not distinguishable from previous studies, our method can be regarded as one of parameter reduction method.

6. REFERENCES