# TREND TYING IN THE SEGMENTAL-FEATURE HMM

*Young-Sun Yun*

Electronic Information and Communication Engineering,
School of Information Technology and Multimedia Engineering,
Hannam University, Taejon, Republic Of Korea
E-mail: ysyun@mail.hannam.ac.kr

**ABSTRACT**

We present the reduction method of number of parameters in segmental-feature HMM (SFHMM). If the SFHMM shows better results than CHMM, the number of parameters is greater than that of CHMM. Therefore, there is a need for new approach that reduces the number of parameters. 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. The proposed method shares the trend part of trajectories by quantization. The experiments are performed on TIMIT corpus to examine the effectiveness of the trend tying. The experimental results show that its performance is the almost same to that of previous studies. To obtain the better results with small amount of parameters, the various conditions for the trajectory components must be considered.

## 1. INTRODUCTION

HMM has been widely used for various areas due to its easy implementation and flexible modeling capability. 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, Yun and Oh presented segmental-feature HMM(SFHMM) in which the input speech signals are modeled by a set of frame features (segmental-feature) and the segmental features are represented by the parametric trajectory approach[7, 8]. The SFHMM can be implemented by fixed variance, which shares common variance for all frames in the segment, or time-varying variance, which applies different variance for each frame. If the SFHMM shows better performance than conventional HMM, the number of free parameters of SFHMM is greater than that of HMM. Therefore, studies to reduce the number of parameters of SFHMM are required.

This paper proposes a trend tied SFHMM that shares the trend of observed trajectories to reduce the parameters. The trajectories, in general, can be separated two parts: trend is corresponding to the type of variation and location is the segment mid-point value. If SFHMM is the linear system, trend represents the slope, and if a quadratic system, the trend shows the parabolic tendency. Since the SFHMM uses the parametric trajectory system, the trend and location can be easily separated and the tying of trend can be considered as one of methods decreasing the number of parameters.

## 2. SEGMENTAL-FEATURE HMM

The relation between the successive acoustic feature vectors of speech signals can be approximately by some form of trajectory through the feature space. This trajectory can be implemented in parametric or non-parametric approaches and can be limited to have the fixed length or not. 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. In these approaches, we adopted Gish's approach to describe the segment in detail because Deng's method is a data-generative type in a state rather than a feature representation of a segment. 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

$$\mathbf{C}_t = \mathbf{Z}\mathbf{B}_t + \mathbf{E}, \tag{1}$$

where $\mathbf{C}_t$ and $\mathbf{B}_t$ 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 $\mathbf{Z}$. Each frame is represented by a $D$ dimensional feature vector, $\mathbf{Z}$ and $\mathbf{B}_t$ are $N \times R$ design matrix and $R \times D$ trajectory coefficient matrix, respectively. $\mathbf{E}$ 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:

$$\begin{aligned} \mathbf{C}_t &= \mathbf{Y}_{t-M}^{t+M} = \begin{bmatrix} \mathbf{c}_{t-M} \\ \vdots \\ \mathbf{c}_t \\ \vdots \\ \mathbf{c}_{t+M} \end{bmatrix} \\ \mathbf{c}_\tau &= [\, y_{\tau,1} \; \ldots \; y_{\tau,D} \,], \;\; t-M \le \tau \le t+M. \end{aligned} \tag{2}$$

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 $\mathbf{Z}$ can be defined as

$$\begin{aligned} \mathbf{Z} &= \begin{bmatrix} 1 & (-\frac{M}{2M}) & \ldots & (-\frac{M}{2M})^{R-1} \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & (\frac{M}{2M}) & \ldots & (\frac{M}{2M})^{R-1} \end{bmatrix} \\ \mathbf{z}_\tau &= \left[ 1 \; \left(\frac{\tau - t}{2M}\right) \; \ldots \; \left(\frac{\tau - t}{2M}\right)^{R-1} \right], \end{aligned} \tag{3}$$

where $\mathbf{z}_\tau$ is a row vector of $\mathbf{Z}$ along $\tau$. Because $\mathbf{Z}$ 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 $\mathbf{B}_t$ is also defined as

$$\begin{aligned} \mathbf{B}_t &= \begin{bmatrix} \mathbf{b}_1^t \\ \vdots \\ \mathbf{b}_R^t \end{bmatrix} \\ \mathbf{b}_i^t &= [\, b_{i,1}^t \; \ldots \; b_{i,D}^t \,], \;\; 1 \le i \le R. \end{aligned} \tag{4}$$

Since errors are supposed to be independent and identically distributed, we obtain the trajectory coefficient matrix $\hat{\mathbf{B}}_t$ by *a linear regression* or the following matrix equation:

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

where $'$ means the matrix transpose.

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

$$\chi_t^2 = \frac{1}{N} \sum_{\tau=t-M}^{t+M} (\mathbf{c}_\tau - \mathbf{z}_\tau \hat{\mathbf{B}}_t)(\mathbf{c}_\tau - \mathbf{z}_\tau \hat{\mathbf{B}}_t)'. \tag{6}$$

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 $\hat{\mathbf{B}}_t$ with $\chi_t^2$.

### 2.2. Segment likelihood

In segmental HMM(SHMM), the observation probability of a given segment is represented as the product of an *extra-* and *intra-segmental* variations. Extra-segmental variations refer to such long-term variabilities as speaker identity and the chosen pronunciation of a speech sound. On the other hand, intra-segmental variations show short-term variabilities that occur within a segment because of the continuous articulation process and other random fluctuations [6]. However, 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 $\mathbf{C}_t$ of SFHMM are represented as their unique trajectory $\mathbf{Z}\mathbf{B}_t$ at time $t$, the observation probability of $\mathbf{C}_t$ occurring at state $s_i$ of model $\lambda$ is specified by the equation

$$P(\mathbf{C}_t | s_i, \lambda) \;=\; P(\mathbf{Z}\hat{\mathbf{B}}_t | s_i, \lambda) P(\mathbf{C}_t | \mathbf{Z}\hat{\mathbf{B}}_t, s_i, \lambda). \tag{7}$$

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

$$b_j(\mathbf{C}_t) = P(\mathbf{C}_t | s_j, \lambda) = P(\mathbf{Z}\hat{\mathbf{B}}_t | \mathbf{Z}\boldsymbol{\mathcal{B}}_j, \boldsymbol{\Sigma}_j) P(\mathbf{C}_t | \mathbf{Z}\hat{\mathbf{B}}_t), \tag{8}$$

where $\boldsymbol{\mathcal{B}}_j$ and $\boldsymbol{\Sigma}_j$ are the trajectory model corresponding to state $j$. In this equation, the extra-segmental probability and intra-segmental variation are defined as

$$\begin{aligned} P(\mathbf{Z}\hat{\mathbf{B}}_t | \mathbf{Z}\boldsymbol{\mathcal{B}}_i, \boldsymbol{\Sigma}_i) &= \prod_{\tau=t-M}^{t+M} \frac{1}{(2\pi)^{D/2} |\Sigma_{\tau-t,i}|^{1/2}} \cdot \\ & \exp\left\{ -\frac{1}{2} \{\mathbf{z}_\tau(\hat{\mathbf{B}}_t - \boldsymbol{\mathcal{B}}_i)\} \Sigma_{\tau-t,i}^{-1} \{\mathbf{z}_\tau(\hat{\mathbf{B}}_t - \boldsymbol{\mathcal{B}}_i)\}' \right\} \end{aligned} \tag{9}$$

$$P(\mathbf{C}_t | \mathbf{Z}\hat{\mathbf{B}}_t) = \exp\{-\frac{1}{2}\chi_t^2\} \tag{10}$$

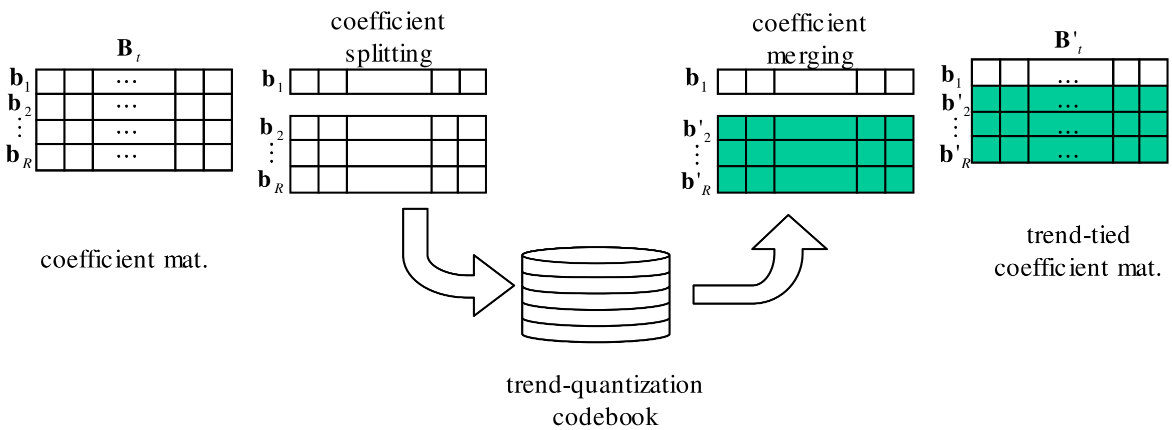


Figure 1: The flow of trend tying : new trajectory coefficient matrix is merged by the original location and quantized trend.

where $\mathbf{\Sigma}_j$ is the sequence of frame variance or common variance by the applied variance approach (time-varying or fixed variance).

## 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.

### 3.1. Separation of trend and location

The trajectory can be rewritten by the linear regression function. For each feature dimension, the following polynomial is considered:

$$y_{\tau,i} = b_{1,i}z_{\tau,1} + b_{2,i}z_{\tau,2} + b_{3,i}z_{\tau,3} + \cdots + b_{R,i}z_{\tau,R}, \quad 1 \leq i \leq D, \tag{11}$$

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 $\left(\frac{\tau-t}{2M}\right)^{r-1}$.

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 $\mathbf{b}_1$ 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 $\mathbf{T}_t$ for the trend can be defined as follows:

$$\mathbf{T}_t = \begin{bmatrix} \mathbf{b}_2^t \\ \vdots \\ \mathbf{b}_R^t \end{bmatrix}. \tag{12}$$

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 $\hat{\mathbf{T}}_t$ from codebook and location from $\mathbf{b}_1$ 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:

$$D(\mathbf{T}_i, \mathbf{T}_j) = \frac{1}{N}\sum_{\tau=1}^{N} \{\tilde{\mathbf{z}}_\tau(\mathbf{T}_i - \mathbf{T}_j)\}\{\tilde{\mathbf{z}}_\tau(\mathbf{T}_i - \mathbf{T}_j)\}', \tag{13}$$

where $\tilde{\mathbf{z}}_\tau$ is the row vector of design matrix which excludes the first column value, $\mathbf{T}_i, \mathbf{T}_j$ are trend coefficient matrices.

Table 1: Comparison of classification rate with various segmental condition. $M$ means the number of mixtures and the quantization level for trend tying is 256.

| type | condition | $M = 1$ | $M = 2$ |
|---|---|---|---|
| CHMM | - | 52.09 | 54.45 |
| SFHMM (Fixed Variance) | $N = 3, R = 2$ | 53.33 | 55.51 |
| | $N = 3, R = 2$ | 53.32 | 55.53 |
| | $N = 5, R = 2$ | 54.22 | 56.31 |
| | $N = 5, R = 3$ | 54.03 | 56.44 |
| SFHMM (Trend Tying) | $N = 3, R = 2$ | 53.25 | 54.95 |
| | $N = 3, R = 3$ | 52.90 | 54.26 |
| | $N = 5, R = 2$ | 53.32 | 54.44 |
| | $N = 5, R = 3$ | 53.06 | 55.01 |

## 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 */iy, ih, ey, eh, ae, aa, ah, ao, ow, uw, uh, ux, er/* and three diphthongs */ay, oy, aw/*. 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 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 and regression order to compare the performance of SFHMM when the trend tying is used. For the trend quantization, 256-level codebook is used. The recognition results are shown in Table 1.

From the experimental result, we found that the proposed system did not outperform CHMM. If the single mixture is used, the performance is better than that of CHMM. On the other hand, the performance is almost the same when the two mixtures are used. It may be caused by the growing of the number of free parameters in system. In CHMM, if the number of mixtures increases, the number of parameters also increases as times as an increment of the number of mixtures. The same effect is also occurred in SFHMM. However, if SFHMM apply the trend tying, the number of parameters for the location increases, whereas that of the trend coefficients is fixed. It is considered another reason that the weighting rate between the trend and location differs. The trend is calculated on $N - 1$ frames, but the location is only obtained at single frame. Therefore, if the location has the same proportion to the trend, the performance may be better.

## 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 parametric trajectories are used for segmental features that are corresponding to set of frames features. The trajectory 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. Furthermore, if the similar ratio of trend to location is used or the multiple trend codebooks are used, the performance may be increase.

## 6. REFERENCES


[1] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," Signal Processing, vol. 27, pp. 65–78, 1992

[2] H. Gish and K. Ng, "A segmental speech model with application to word spotting," In *Proc. of Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. II-447–450, 1993

[3] M.J.F. Gales and S.J. Young, "Segmental Hidden Markov Models," In *Proc. of European Conf. on Speech Comm. and Tech.*, pp. 1579–1582, 1993

[4] H. Gish and K. Ng, "Parametric trajectory models for speech recognition," In *Proc. of Int. Conf. on Spoken Lang. Proc.*, pp. I-466–469, 1996

[5] M. Ostendorf, V.Digalakis, and O.A. Kimball, "From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *IEEE Tr. on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996

[6] W.J. Holmes and M.J. Russell, "Probabilistic trajectory segmental HMMs," *Computer Speech and Language*, vol. 13, pp. 3–37, 1999

[7] Y.-S. Yun and Y.-H. Oh, "A Segmental-Feature HMM for Speech Pattern Modeling," *IEEE Signal Processing Letters*, vol. 7, no. 6, pp. 135–137, 2000

[8] Y.-S. Yun and Y.-H. Oh, "A Segmental-Feature HMM for Continuous Speech Recognition Based On a Parametric Trajectory Model," *Speech Communication*, (to appear)