ELSEVIER

Speech Communication 38 (2002) 115–130

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A segmental-feature HMM for continuous speech recognition based on a parametric trajectory model

Young-Sun Yun [a,*], Yung-Hwan Oh [b]

[a] Electronic Information and Communication Engineering, School of Information Technology and Multimedia Engineering, Hannam University, 133 Ojung-dong, Daeduk-gu, Taejon 306-791, Republic of Korea

[b] Division of Computer Science, Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, 373-1 Kusong-dong, Yusong-gu, Taejon 305-701, Republic of Korea

Received 5 September 2000; received in revised form 15 May 2001; accepted 24 May 2001

Abstract

In this paper, we propose a new acoustic model for characterizing segmental features and an algorithm based upon a general framework of hidden Markov models (HMMs). The segmental features are represented as a trajectory of observed vector sequences by a polynomial regression function. To obtain the polynomial trajectory from speech segments, we modify the design matrix to include transitional information for contiguous frames. We also propose methods for estimating the likelihood of a given segment and trajectory parameters. The observation probability of a given segment is represented as the relation between the segment likelihood and the estimation error of the trajectories. The estimation error of a trajectory is considered the weight of the likelihood of a given segment in a state. This weight represents the probability of how well the corresponding trajectory characterizes the segment. The proposed model can be regarded as a generalization of a conventional HMM and a parametric trajectory model. We conducted several experiments to establish the effectiveness of the proposed method and the characteristics of the segmental features. The recognition results on the TIMIT database demonstrate that the performance of segmental-feature HMM (SFHMM) is better than that of a conventional HMM.

Keywords: Parametric trajectory model; Segmental HMM; Segmental-feature HMM; Speech Recognition

1. Introduction

Hidden Markov models (HMMs) are widely used for automatic speech recognition. HMMs provide a framework that is broadly appropriate for modeling speech patterns by considering time-scale and short-term spectral variation. However, most approaches based on HMMs assume that the sequence of observations corresponding to a state are conditionally independent of each other and come from the same output distribution. These assumptions of state-conditioned stationarity and independence of observation events imply that each state is a stationary source that generates independent and identically distributed (i.i.d.) observation events. The stationary-state assumption appears to be reasonable when a state is intended to represent a short segment of sonorant or fricative speech sounds. However, for longer segments of these

* Corresponding author.

E-mail addresses: ysyun@mail.hannam.ac.kr (Y.-S. Yun), yhoh@cs.kaist.ac.kr (Y.-H. Oh).

0167-6393/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(01)00047-4

sounds and for all types of plosive sounds, such an assumption is inadequate (Deng et al., 1994).

This paper proposes a segmental-feature HMM (SFHMM) to overcome the weaknesses in the assumptions of conventional HMMs and to take acoustic context information between neighborhood frames into consideration. The SFHMM uses parametric trajectories for input features and each trajectory is modeled in a given segment to include information from adjacent frames whose range is determined by a design matrix for the fitting problem. The trajectory is referred to as a "segmental feature" because it is obtained from a segment (a set of frames).

To consider the neighborhood frame information, the design matrix is modified to be symmetric with respect to the current observation vector and to use a time-normalized matrix so as not to depend on the segment length. In contrast to previous work (Gish and Ng, 1993, 1996), we use fixed segment lengths so that the mean trajectory can be obtained from the expected averages of feature vectors, to reduce the estimation time of the trajectory. The estimation and evaluation problems of SFHMM are similar to those of a conventional HMM, because the HMM frame features are replaced with the segmental features (trajectories) and the problems can be expanded easily by considering a set of frame features.

The remainder of the paper is organized as follows. In Section 2, we introduce segment modeling using the parametric trajectory approach and present a new design matrix to reflect neighborhood events. We also describe the basic formulae associated with the SFHMM. Section 3 proposes a mathematical form for the SFHMM to evaluate likelihood in a given state and to estimate parameters. We then consider specializations of SFHMM that work as conventional HMMs or as parametric trajectory models. Experimental results using the TIMIT database and discussions are reported in Section 4, and we conclude in Section 5 with a summary of the main results.

2. Speech segment modeling

The effect of independence assumption can be reduced by the modification of the data to reflect speech dynamics or of the model to be appropriate for speech signals. The popular way of the former approach is to use an acoustic feature vector which includes information over a time span of several frames. This is usually achieved by considering the time derivatives, such as the first or second order delta coefficients of the static features (Furui, 1986). The assumptions of HMMs are related to their frame-based characteristic, whereby states are associated with single acoustic feature vectors. The latter approach incorporates the concept of modeling frame sequences rather than individual frames to provide an accurate model for the speech dynamics.

Considering the above concept, many studies have been presented to mitigate the impact of the inappropriate assumptions of HMMs. One approach tried to relax the assumptions of HMMs by representing the duration in a given state as a regression function of the frame index. The frame index represents the position from the beginning of a given state (Deng, 1992) or sojourn time (Deng et al., 1994). Even though these studies modeled the state duration effectively, they could not escape from the independence assumption of HMM, since their work was based on frame features.

Some work has been done to overcome the shortcomings of HMMs by addressing the relationship between successive acoustic feature vectors. Segmental acoustic models were studied to exploit acoustic features that are apparent at the segmental, but not at the frame level. The term "segment" is taken as referring to any sequence of frames representing some linguistically meaningful speech unit. Segmental HMM (SHMM) was presented to specify the relationships in a segment as probabilistic functions of a state (Gales and Young, 1993; Russell, 1993) or of some form of trajectory (Holmes and Russell, 1995b, 1999). The earlier work described the output probability distribution of a state by *inter-segmental distribution* (the distribution of the segment mean) and *intra-segmental distribution* (the observation probabilities given that mean). The later work represents the segment as a stochastic process in which the mean changes as a function of time according to the parameters of the trajectory. This stochastic process is considered as a constant distribution over time (Russell, 1993) or a linear trajectory of the variable duration (Holmes and Russell, 1995b, 1999). Thus, if the assumption of stochastic distribution changes, the parameters in a given segment must be redefined.

In another study of segmental models, observations in the segment were defined as the linear sampling of a normalized trajectory (parametric segment model or parametric trajectory model) (Gish and Ng, 1993, 1996). To model a given segment as a normalized trajectory, the boundary of the segment must be known, because the trajectory represents the tendency of successive feature vectors as a normalized feature in the range of $[0, 1]$.

If multiple observation vectors are used to train the regression function or trajectory coefficients in a non-stationary HMM or parametric segmental model, both approaches use the concatenated observation vector for the estimation of the relevant parameters, because the segment lengths are variable and different from each other (Gish and Ng, 1993, 1996; Deng et al., 1994). The estimation time of the parameters therefore increases exponentially. Thus, there is a requirement to study acoustic modeling techniques that improve the performance of speech recognition systems and to reduce the parameter re-estimation and evaluation time. In this section, we present a new acoustic model based on the segmental feature with restriction of the segmental models.

2.1. Parametric segmental feature modeling

The relation between the successive acoustic feature vectors of a speech signal can be approximated by some form of trajectory through the feature space. This idea has formed the basis for a number of segmental models, and it has been implemented in parametric and non-parametric approaches. In the parametric approach, the polynomial trajectory is estimated from a specific region, and distributions of the region are represented by points along the trajectory. On the other hand, the non-parametric trajectory model has distribution parameters that are estimated for each model region (Goldenthal and Glass, 1993). We use the parametric approach for segment modeling, because this method is well-motivated by the smooth trajectories in many speech units (Ostendorf et al., 1996).

In their parametric trajectory model, Gish and Ng (1993) modeled a speech segment $\boldsymbol{C}$ with a duration of N frames, which is $N \times D$ matrix, as

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

where each frame is represented by a D-dimensional feature vector. $\boldsymbol{Z}$ is an $N \times R$ design matrix that specifies the type of window to be used, $\boldsymbol{B}$ is an $R \times D$ trajectory coefficient matrix, and $\boldsymbol{E}$ denotes residual errors. R is the regression order in the trajectory model: $R = 1$ for a constant trajectory, $R = 2$ for a linear trajectory, and $R = 3$ for a quadratic trajectory. In this study, a segment is normalized so that its frames are distributed uniformly between times 0 and 1 inclusively (Gish and Ng, 1993) (called a "complete" trajectory model because each trajectory represents the entire input data corresponding to the speech unit as the normalized form). This normalization is handled in the design matrix $\boldsymbol{Z}$.

In a complete trajectory model (Gish and Ng, 1993, 1996; Fukada et al., 1997), a segment might correspond to a phone-sized unit and the boundaries of a segment must be known, because the segment is normalized to the range of $[0, 1]$. If the segment boundaries are unknown, distribution mapping between a segment model and variable length observations is very important. However, the mapping takes a long time and it is difficult to find the correct mappings.

Thus, in our system, segmental features are extracted from successive frame features using the design matrix $\boldsymbol{Z}$. This matrix determines the range of frames that are reflected to the segmental feature and the weights of each frame feature in a segment. If we limit the lengths of segments to a fixed value, the design matrices can be the same for most segments except segments at the ends of the speech signal. At the ends of the signal, the segments whose length is less than the given fixed value cannot be a consideration or their design matrices must shrink from the top or bottom. Since the lengths of segments are set to a fixed value, the distribution mapping problem can be avoided. We represent the segmental features as polynomial trajectories that are obtained by Eq. (1). If an

HMM-like method is used and each acoustic model is trained for a phone-sized unit, the boundary alignment of phones or words is easily performed in the evaluation phase using the Viterbi algorithm, because we can consider that the segmental features are replaced with the frame features used in conventional HMM. In the estimation phase, the mean trajectory parameters are easily obtained by the expected average of observation feature vectors because the design matrices for most segments have the same form.

If the normalized form of the design matrix is adopted, we may use the design matrix to represent relative time, to make the trajectory of a segment independent of the segment length. We can also extend our approach to a variable segmental feature, which is obtained from segments with variable lengths. However, we cannot apply the normalized time regression matrix directly to our approach, as Gish and Ng (1993) suggested, because the segment boundary must be known and the trajectory is characterized by the complete trajectory. Therefore, we suggest a new design matrix in which the bases are symmetric with respect to the current frame (time) in each column and are normalized by the segment length. The range of normalized bases is set to $[-0.5, 0.5]$ to focus on the center of a segment.

To reflect the above concept, we make some modifications to the parametric representation of Eq. (1). In previous work, a segment was represented by a phone-sized unit. However, in our approach, a segment contains a fixed number of frames and the segment can be moved by one or more frames (less than the segmental length to avoid discontinuity) for analysis. For a given speech segment of $N = 2M + 1$ frames, we consider the current frame feature (the observation vector at time t) on the center of a segment as follows:

$$\boldsymbol{C}_t = \boldsymbol{Y}_{t-M}^{t+M} = \begin{bmatrix} y_{t-M,1} & \cdots & y_{t-M,D} \\ \vdots & \vdots & \vdots \\ y_{t,1} & \cdots & y_{t,D} \\ \vdots & \vdots & \vdots \\ y_{t+M,1} & \cdots & y_{t+M,D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_{t-M} \\ \vdots \\ \boldsymbol{c}_t \\ \vdots \\ \boldsymbol{c}_{t+M} \end{bmatrix},$$

$$\boldsymbol{c}_\tau = [y_{\tau,1} \ldots y_{\tau,D}], \quad t - M \leqslant \tau \leqslant t + M. \tag{2}$$

Because the observation vector at time t ($\boldsymbol{c}_t$) is on the center of a segment and the segment is moved by time index t in consecutive frames, the beginning and end of the speech segment may be overlapped with the neighboring segments at time $t-1$ or $t+1$.

Thus, there is a strong requirement to describe a new design matrix that can imply the transitional information between the adjacent segments and can show that the current feature vector is positioned at the center of a segment. Since a design matrix centers on the current observation vector and we consider the segments overlapping, each row shows the relative position from the current time. Each element in the first-half of the row has a negative value, while those in the last half have positive values.

To satisfy the above conditions, the design matrix $\boldsymbol{Z}$ can be defined as

$$\boldsymbol{Z} = \begin{bmatrix} 1 & \left(-\frac{M}{2M}\right) & \left(-\frac{M}{2M}\right)^2 & \ldots & \left(-\frac{M}{2M}\right)^{R-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \left(-\frac{m}{2M}\right) & \left(-\frac{m}{2M}\right)^2 & \ldots & \left(-\frac{m}{2M}\right)^{R-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \left(\frac{m}{2M}\right) & \left(\frac{m}{2M}\right)^2 & \ldots & \left(\frac{m}{2M}\right)^{R-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \left(\frac{M}{2M}\right) & \left(\frac{M}{2M}\right)^2 & \ldots & \left(\frac{M}{2M}\right)^{R-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}_{t-M} \\ \vdots \\ \boldsymbol{z}_{t-m} \\ \vdots \\ \boldsymbol{z}_t \\ \vdots \\ \boldsymbol{z}_{t+m} \\ \vdots \\ \boldsymbol{z}_{t+M} \end{bmatrix},$$

$$\boldsymbol{z}_\tau = \left[1 \quad \left(\frac{\tau-t}{2M}\right) \quad \left(\frac{\tau-t}{2M}\right)^2 \quad \ldots \quad \left(\frac{\tau-t}{2M}\right)^{R-1}\right]. \tag{3}$$

The base is represented as time normalized by the segment length in each column, so that it is not dependent on the segment length. Because $\boldsymbol{Z}$ is represented by the relative position from the current observation vector, the previous and following acoustic features of the current observation can be included in the estimation of the trajectory. Thus, even though acoustic models are represented as context-independent (CI) models, the estimated trajectory can contain the partial context information of neighboring acoustic events. The base $((\tau - t)/2M)$ of each column in the design matrix has a normalized range from -0.5 to 0.5 because τ varies from $t - M$ to $t + M$. The trajectory coefficient matrix $\boldsymbol{B}_t$ is also defined as

$$\boldsymbol{B}_t = \begin{bmatrix} b_{1,1} & \cdots & b_{1,D} \\ \vdots & \vdots & \vdots \\ b_{R,1} & \cdots & b_{R,D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_R \end{bmatrix},$$
$$\boldsymbol{b}_i = [b_{i,1} \dots b_{i,D}], \quad 1 \leqslant i \leqslant R. \tag{4}$$

Fig. 1 shows the relation between the observation vectors, the design matrix, the trajectories and their coefficient matrices.

Given the segment model as Eqs. (1)–(4), the next step is to estimate the model parameters. Assuming that errors are independent and identically distributed, we obtain the trajectory coefficient matrix $\hat{\boldsymbol{B}}_t$ by *a linear regression*. For each feature dimension, the following polynomial is considered:

$$y_{\tau,i} = b_{1,i} z_{\tau,1} + b_{2,i} z_{\tau,2} + b_{3,i} z_{\tau,3} + \cdots + b_{R,i} z_{\tau,R}, \quad 1 \leqslant i \leqslant D, \tag{5}$$

where $z_{\tau,r} = ((\tau - t)/2M)^{r-1}$. This linear regression equation is easily solved by singular value decomposition (SVD). If the trajectory model is an over-determined system ($N > R$), SVD produces a solution that is the best approximation in the least-squares sense (Press et al., 1992). Otherwise, the coefficient matrix $\hat{\boldsymbol{B}}$ can be obtained by the following matrix equation:

$$\hat{\boldsymbol{B}}_t = \left[\boldsymbol{Z}'\boldsymbol{Z}\right]^{-1} \boldsymbol{Z}'\boldsymbol{C}_t, \tag{6}$$

where $'$ means the matrix transpose. The above methods must be applied carefully at the beginning and end of a speech signal, where the conditions

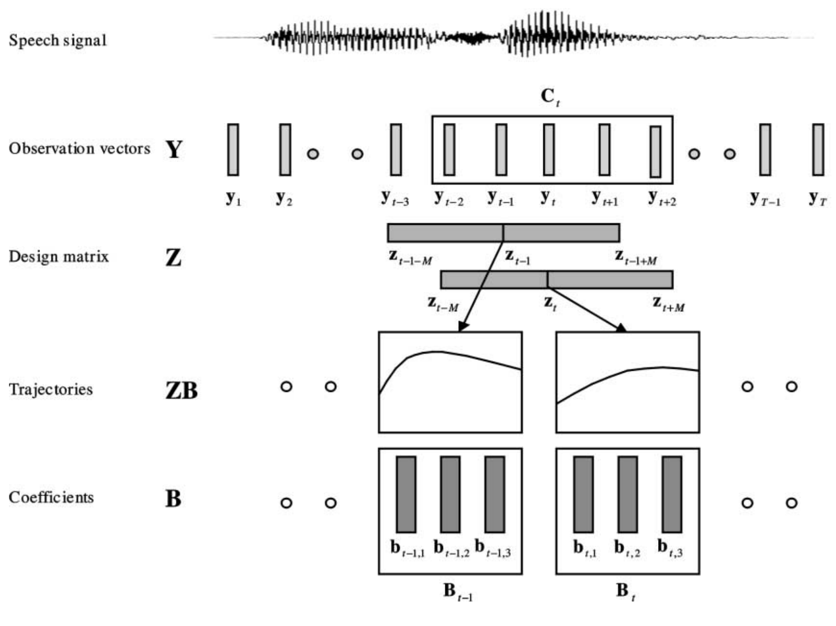

Fig. 1. The relation between the observation vectors, the design matrix, the trajectories and their coefficients.

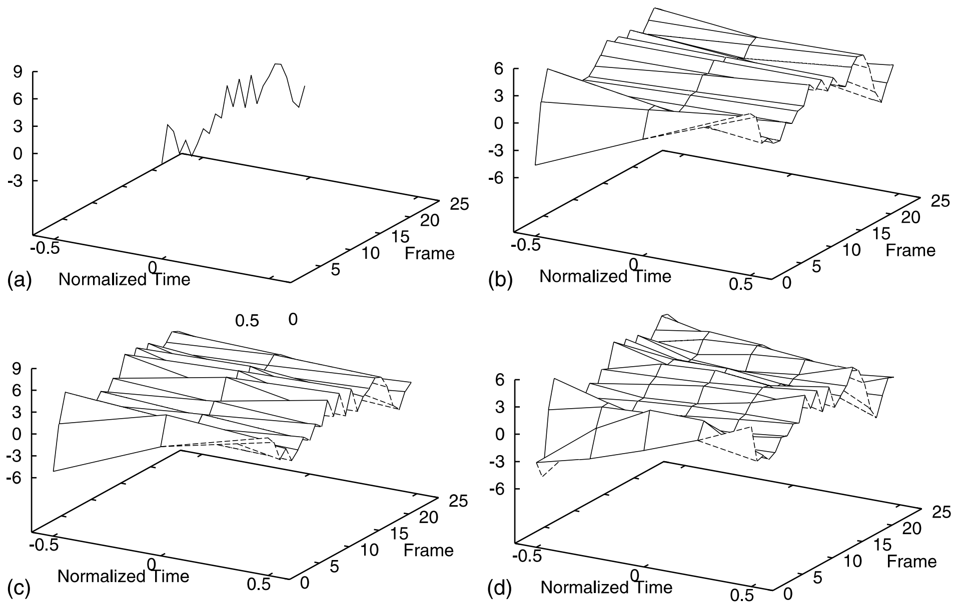

Fig. 2. Temporal variation of the zeroth MFCC coefficient for the vowel segment /oy/, and the corresponding segmental features with $N = 3$ and 5, respectively. N means the segment length and R the regression order for trajectories: (a) original; (b) $N = 3,\ R = 2$; (c) $N = 3,\ R = 3$; (d) $N = 5,\ R = 3$.

$\tau < 0$ or $\tau \geqslant T$ (T is the total number of observations in a speech signal) occur.

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

$$\chi_t^2 = \frac{\sum_{\tau=t-M}^{t+M} (\boldsymbol{c}_\tau - \boldsymbol{z}_\tau \hat{\boldsymbol{B}}_t)(\boldsymbol{c}_\tau - \boldsymbol{z}_\tau \hat{\boldsymbol{B}}_t)'}{N}, \tag{7}$$

where N is the number of frames in a segment. Smaller values of χ^2 indicate better data fitting; if χ^2 is zero, the fit is perfect. After the parameter estimation, each segment is replaced by its trajectory $\boldsymbol{Z}\hat{\boldsymbol{B}}_t$ with χ_t^2.

When speech signals are entered, they are grouped into a fixed number of successive frames (segments). Each segment can be moved by one or more frames for speech analysis and it is represented by its trajectory and estimation error. The observation vector at time t is represented by a D-dimensional point (frame feature) in conventional speech processing, but the proposed segmental modeling considers the observation vectors of a segment as sampled points along the corresponding trajectory (segmental feature).

Fig. 2 illustrates the difference between the temporal variation of the zeroth MFCC coefficient in conventional speech processing and its corresponding segmental feature (trajectories $\boldsymbol{ZB}_t$) with a different segment length and regression order. In this figure, the features in the conventional speech processing are represented by simple points (frame features) at each frame index, while the segmental features are represented as the polynomial trajectories (the points along linear, quadratic or higher-order regression functions).

2.2. Segmental HMM

An SHMM provides an efficient representation of a set of possible underlying trajectories for a speech sound by modeling variability in the trajectories (Holmes and Russell, 1999). In this model, the observation probability of a given segment $Y = y_1, \ldots, y_T$ occurring at state s_i of model λ is specified by the equation

$$P(Y \mid s_i, \lambda) = \int P(f_m \mid s_i, \lambda) P(Y \mid f_m, s_i, \lambda)\, \mathrm{d}f_m, \tag{8}$$

where f_m is the available trajectory. $P(f_m \mid s_i, \lambda)$ is the probability that f_m is a valid trajectory for the segment corresponding to the given state and it is referred to as an *extra-segmental* variation. $P(Y \mid f_m, s_i, \lambda)$ is the probability that the observation segment Y is a valid instance of trajectory f_m, and it accounts for the *intra-segmental* variation in the realization of a particular trajectory. 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 represent short-term variabilities that occur within a segment as a result of the continuous articulation process and other random fluctuations (Holmes and Russell, 1995a). Both segmental variations are based on the assumption that all observations are independent of a given state, but conditioned on the segment (Gales and Young, 1993; Russell, 1993; Holmes and Russell, 1995b).

In our approach, a segment is a fixed number of frames and the segment length is $N = 2M + 1$. Thus, the segment $\boldsymbol{Y}_{t-M}^{t+M}$ can be expressed by $\boldsymbol{C}_t$. Because the segment $\boldsymbol{C}_t$ is represented by its unique trajectory $\boldsymbol{ZB}_t$ at time t, we can rewrite Eq. (8) as follows:

$$P(\boldsymbol{C}_t \mid s_i, \lambda) = P(\boldsymbol{ZB}_t \mid s_i, \lambda) P(\boldsymbol{C}_t \mid \boldsymbol{ZB}_t, s_i, \lambda). \tag{9}$$

The observation probability of segment $\boldsymbol{C}_t$, for a given state, can be obtained by summing up all mixture probabilities in the state if we use multiple mixture models. Then, Eq. (9) can be rewritten as

$$\begin{aligned} P(\boldsymbol{C}_t \mid s_i, \lambda) &= \sum_{k=0}^{K} c_{ik} P(\boldsymbol{C}_t \mid s_i, m_k, \lambda) \\ &= \sum_{k=0}^{K} c_{ik} P(\boldsymbol{ZB}_t \mid s_i, m_k, \lambda) \\ &\quad \times P(\boldsymbol{C}_t \mid \boldsymbol{ZB}_t, s_i, m_k, \lambda), \end{aligned} \tag{10}$$

where c_{ik} shows the weight of mixture m_k in state s_i. In the previous SHMM, the probability of the segment given the state was represented by integrating over the set of all trajectory parameters. However, in our approach, we instead use a summation, to be formulated like the mixture model of conventional HMMs.

However, if we want to use the most likely trajectory, with parameters that maximize the joint probability of observations and the state, the above equation can be rewritten as follows:

$$\begin{aligned} P(\boldsymbol{C}_t \mid s_i, \lambda) &= \max_k P(\boldsymbol{C}_t \mid s_i, m_k, \lambda) \\ &= \max_k P(\boldsymbol{ZB}_t \mid s_i, m_k, \lambda) \\ &\quad \times P(\boldsymbol{C}_t \mid \boldsymbol{ZB}_t, s_i, m_k, \lambda). \end{aligned} \tag{11}$$

3. Segmental-feature HMM

A segment has a variable length in previous segmental models (Gales and Young, 1993; Russell, 1993; Gish and Ng, 1993). However, in SFHMM, the segment length is restricted to a predetermined fixed value. Since each segment is represented by its polynomial trajectory $\boldsymbol{ZB}_t$ and estimation error, after the estimation of the coefficients matrix $\boldsymbol{B}_t$ and the goodness-of-fit χ^2, these parameters can be used to evaluate the likelihood of speech segments. In this section, we propose a new segment likelihood for a given model, and a parameter estimation algorithm within the framework of the conventional HMM.

3.1. Segment likelihood

In previous SHMMs (Russell, 1993; Holmes and Russell, 1995b), the extra-segmental variance is corresponding to the variability of the trajectories which are represented as constant or linear and is obtained by the variance of the trajectory parameters themselves. Intra-segmental distributions are assumed Gaussian with a diagonal covariance. However, we can suppose extra-segmental and intra-segmental variations as follows. Extra-segmental variations are represented by Gaussian distributions with mean trajectories and their variances, while intra-segmental variations are also defined as the estimation error of the trajectory in a segment. This means that intra-segmental variations may act as weights to extra-segmental variations, which show the observation probability of a given trajectory in the state. The weight is the probability of how well the given trajectory reflects the characteristics of the segment.

The extra-segmental probability of the trajectory $\boldsymbol{ZB}_t$, given state s_i and model λ, can be defined as

$$P(\boldsymbol{ZB}_t \mid s_i, \lambda) = P(\boldsymbol{ZB}_t \mid \boldsymbol{Z}\mathscr{B}_i, \boldsymbol{\Sigma}_i) = \prod_{\tau=t-M}^{t+M} \frac{1}{(2\pi)^{D/2} |\Sigma_{\tau-t,i}|^{1/2}} \times \exp\left\{ -\frac{1}{2} \{\boldsymbol{z}_\tau(\boldsymbol{B}_t - \mathscr{B}_i)\} \times \Sigma_{\tau-t,i}^{-1} \{\boldsymbol{z}_\tau(\boldsymbol{B}_t - \mathscr{B}_i)\}' \right\}, \tag{12}$$

where $\mathscr{B}_i$ is the mean trajectory coefficient matrix, $\boldsymbol{\Sigma}_i$ is the variance of the trajectories, and $\boldsymbol{\Sigma}_i = \{\Sigma_{-M,i}, \ldots, \Sigma_{0,i}, \ldots, \Sigma_{M,i}\}$, $\Sigma_{n,i}, -M \leqslant n \leqslant M$ is the variance of points at frame index n along trajectories in state s_i. For given segments occurring in a state, the expression of variance can be considered as a time-varying quantity reflecting temporal dependencies of frames in segments or a common variance for all frames in the segments. The time-varying variances can represent the temporal dynamics of the segment and $\Sigma_{\tau-t,i}$ is obtained by the variance of a frame feature along the relative time $\tau - t$ in Eq. (12). However, the common variance can be referred to as a fixed variance because the temporal variabilities in the segment are constant, and each element of $\boldsymbol{\Sigma}_i$ is the same for all frame indexes $\tau - t$ in the segment, such that $\Sigma_{\tau-t,i} = \Sigma_i, \Sigma_i$ is the common variance for all frames in a segment.

Since intra-segmental variation is independent of state s_i and is interpreted as a trajectory estimation error, the variation can be defined with χ^2 as follows:

$$P(\boldsymbol{C}_t \mid \boldsymbol{ZB}_t, s_i, \lambda) = P(\boldsymbol{C}_t \mid \boldsymbol{ZB}_t) = \exp\left\{ -\frac{1}{2}\chi_t^2 \right\} = \exp\left\{ -\frac{1}{2N} \sum_{\tau=t-M}^{t+M} (\boldsymbol{c}_\tau - \boldsymbol{z}_\tau \boldsymbol{B}_t) \times (\boldsymbol{c}_\tau - \boldsymbol{z}_\tau \boldsymbol{B}_t)' \right\}. \tag{13}$$

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

$$\begin{aligned} b_j(\boldsymbol{C}_t, \boldsymbol{ZB}_t) = P(\boldsymbol{C}_t \mid s_j, \lambda) &= \sum_{k=0}^{K} c_{jk} b_{jk}(\boldsymbol{C}_t, \boldsymbol{ZB}_t) \\ &= \sum_{k=0}^{K} c_{jk} P(\boldsymbol{C}_t \mid s_j, m_k, \lambda) \\ &= \sum_{k=0}^{K} c_{jk} P(\boldsymbol{ZB}_t \mid \boldsymbol{Z}\mathscr{B}_{jk}, \boldsymbol{\Sigma}_{jk}) P(\boldsymbol{C}_t \mid \boldsymbol{ZB}_t) \\ &= P(\boldsymbol{C}_t \mid \boldsymbol{ZB}_t) \times \sum_{k=0}^{K} c_{jk} P(\boldsymbol{ZB}_t \mid \boldsymbol{Z}\mathscr{B}_{jk}, \boldsymbol{\Sigma}_{jk}), \end{aligned} \tag{14}$$

where $\mathscr{B}_{jk}$ and $\boldsymbol{\Sigma}_{jk}$ are the trajectory model (coefficient matrix of mean trajectory and covariance) corresponding to mixture m_k for state j.

Since $P(\boldsymbol{C}_t \mid \boldsymbol{ZB}_t)$ is independent of state j and denotes the estimation error of the trajectory for the segment, the intra-segmental variation is considered as the time varying weight of the contribution to the extra-segmental probability for a given state.

3.2. Parameter estimation

A Baum–Welch-like parameter re-estimation can be derived for SFHMM. Let $\gamma_t(j)$ and $\xi_t(j,k)$ be respectively the posterior probabilities of being in state j and of being in the kth mixture of state j at time t, given the observation sequence $\boldsymbol{C}_t$ and the model λ. $\xi_t(j,k)$ can be defined as

$$\xi_t(j,k) = P(s_t = j, k_t = k \mid \boldsymbol{C}_t, \lambda) = \frac{\sum_i \alpha_{t-1}(i) a_{ij} c_{jk} b_{jk}(\boldsymbol{C}_t, \boldsymbol{ZB}_t) \beta_t(j)}{\sum_{i \in S_F} \alpha_T(i)}, \tag{15}$$

where a_{ij} is the transition probability from state i to state j, S_F the set of final states, and $\alpha_t(i)$ the forward probability of state i at time t. If $\gamma_t(j)$ is the same as that of the conventional HMM, after the estimation of $\gamma_t(i)$ and $\xi_t(j,k)$, we can estimate the weight for the kth mixture in state j as follows:

$$\bar{c}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k)}{\sum_{t=1}^{T} \gamma_t(j)}. \tag{16}$$

The mean trajectory of a mixture in a state of SFHMM can be obtained by the ratio of the expected mean trajectories of the kth density over the expected number of times of being in state j for all segments,

$$\boldsymbol{Z}\bar{\mathscr{B}}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k)\boldsymbol{Z}\boldsymbol{B}_t}{\sum_{t=1}^{T} \xi_t(j,k)}. \tag{17}$$

As SFHMM analyzes a speech segment with a fixed length, i.e. it uses the same form of the design matrix, we can commonly omit the design matrix $\boldsymbol{Z}$ in Eq. (17). Thus, the trajectory coefficient matrix can be estimated by

$$\bar{\mathscr{B}}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k)\boldsymbol{B}_t}{\sum_{t=1}^{T} \xi_t(j,k)}. \tag{18}$$

After estimating $\bar{\mathscr{B}}_{jk}$, we can also estimate the covariance by the difference between the input segmental feature and the mean trajectory $\boldsymbol{Z}\bar{\mathscr{B}}_{jk}$. For the covariance estimation, we may consider the time-varying variance that reflects the temporal dependence of frames in segments, and the fixed variance, which is commonly applied to all frames in a segment, occurring in a state.

3.2.1. Time-varying variance

Variances are obtained by the difference between the input vectors and the mean vectors. These features are important to the probability distribution of segmental features with mean vectors. In our approach, the variance in a state can be calculated by the difference between the estimated trajectories of the input speech segment and the mean trajectory of a given state, because features for a segment are represented by a polynomial trajectory. We first assume that variances of a state can be obtained by a sequence of variances of frames of segments that are observed in the state. Since each segment has N frames (equal to the segment length), the variance of the state has a sequence of variances of frames (N variances), and an individual variance is computed for each frame of the segments. Thus, we refer to this approach as a time-varying variance because variances are ordered by the relative position of frames (or relative time) in the segment. As the segment is represented by its unique trajectory in our approach, the variance can be obtained by the distance between the estimated trajectory and the mean trajectory. To calculate the variance at each frame, the distance between both trajectories is achieved by accumulating the distance between points along each trajectory.

Hence, the time-varying variance can be estimated by

$$\bar{\boldsymbol{\Sigma}}_{jk} = \left\{ \bar{\Sigma}_{-M,jk}, \ldots, \bar{\Sigma}_{0,jk}, \ldots, \bar{\Sigma}_{M,jk} \right\},$$
$$\bar{\Sigma}_{n,jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k)\{\boldsymbol{z}_n\boldsymbol{B}_t - \boldsymbol{z}_n\bar{\mathscr{B}}_{jk}\}'\{\boldsymbol{z}_n\boldsymbol{B}_t - \boldsymbol{z}_n\bar{\mathscr{B}}_{jk}\}}{\sum_{t=1}^{T} \xi_t(j,k)}, \tag{19}$$

where n is the relative position in a segment. $\boldsymbol{z}_n\boldsymbol{B}_t$ and $\boldsymbol{z}_n\bar{\mathscr{B}}_{jk}$ are points along the estimated trajectory of a given segment at time t, and along the mean trajectory of the kth mixture in state j, respectively.

3.2.2. Fixed variance

If the time-varying variance approach is used, the trajectory coefficient matrix $\boldsymbol{B}$ and N variances are necessary to represent each state with a single mixture in SFHMM. Thus, if the segment widens or the number of mixtures increases, the number of parameters increases and more computing time is required. We can reduce the number of parameters representing the state by using a common variance for all frames in a segment. In this case, the common variance can be obtained by averaging variances of all frames in the segment. It means that the variances of all frames are assumed the same and the temporal dependencies of variances in a segment are ignored. This approach is called "fixed variance" because variances of frames in the segments are fixed in contrast to the time-varying variance method.

The fixed variance is calculated as follows:

$$\bar{\Sigma}_{jk} = \frac{1}{N} \sum_{n=-M}^{M} \bar{\Sigma}_{n,jk}. \tag{20}$$

This equation can be expanded as

$$\bar{\Sigma}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k)\{(\boldsymbol{Z}\boldsymbol{B}_t - \boldsymbol{Z}\bar{\mathscr{B}}_{jk})'(\boldsymbol{Z}\boldsymbol{B}_t - \boldsymbol{Z}\bar{\mathscr{B}}_{jk})\}}{N\sum_{t=1}^{T} \xi_t(j,k)}, \tag{21}$$

where $\boldsymbol{Z}\boldsymbol{B}_t$ and $\boldsymbol{Z}\bar{\mathscr{B}}_{jk}$ are the set of feature vectors representing the segment at time t and of the kth mixture in state j, respectively.

If the fixed variance approach is used for SFHMM, the output probability of the kth mixture in state j can be simplified as

$$P(\boldsymbol{ZB}_t \mid s_j, m_k, \lambda) = P(\boldsymbol{ZB}_t \mid \boldsymbol{Z}\mathscr{B}_{jk}, \boldsymbol{\Sigma}_{jk}) = \frac{1}{(2\pi)^{ND/2}|\Sigma_{jk}|^{N/2}} \cdot \exp\left\{-\frac{1}{2}\mathrm{Tr}\left\{(\boldsymbol{ZB}_t - \boldsymbol{Z}\mathscr{B}_{jk}) \times \Sigma_{jk}^{-1}(\boldsymbol{ZB}_t\boldsymbol{Z}\mathscr{B}_{jk})'\right\}\right\}. \quad (22)$$

The above approaches to estimating variance are equivalent to that of conventional HMM if the segment length is one, i.e. a single frame is used to represent the segment. However, if the segment length is greater than one, these approaches have different variances and the probability of extra-segmental variation $P(\boldsymbol{ZB}_t \mid s_i, \lambda)$ in Eq. (12) is also different.

3.3. Special cases of SFHMM

There are two special cases of the SFHMM where it can be reduced to models described in the literature.

(1) If the segment length $N = 1$ and the regression order $R = 1$, the trajectory $\boldsymbol{ZB}_t$ of a segment at time t is identical to the single frame feature $\boldsymbol{c}_t$. Thus, the trajectory estimation error or goodness-of-fit χ^2 is zero and the probability of intra-segmental variation $P(\boldsymbol{C}_t \mid \boldsymbol{ZB}_t, s_i, \lambda)$ becomes one. The extra-segmental variation $P(\boldsymbol{ZB}_t \mid s_i, \lambda)$ in Eq. (12) is altered as follows:

$$P(\boldsymbol{ZB}_t \mid s_i, \lambda) = P(\boldsymbol{ZB}_t \mid \boldsymbol{Z}\mathscr{B}_t, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_{0,i}|^{1/2}} \cdot \exp\left\{-\frac{1}{2}(z_0\boldsymbol{B}_t - z_0\mathscr{B}_i) \times \Sigma_{0,i}^{-1}(z_0\boldsymbol{B}_t - z_0\mathscr{B}_i)'\right\}, \quad (23)$$

when $\tau = t$. In the above equation, $z_0\boldsymbol{B}_t$ shows the input vector $\boldsymbol{c}_t$ and $z_0\mathscr{B}_i$ are the mean feature for a state s_i. Thus, this equation is identical to the observation probability of a state in conventional HMM. Therefore, the SFHMM is identical to the continuous HMM.

(2) If the segment length N is equal to the observation length T of the input speech signal for a given model, each acoustic model can be represented as one state (or segment) with the variable length approach. To deal with input speech with variable length, the design matrix $\boldsymbol{Z}$ must consider the variable length of a segment by expanding the number of rows to the observation length T. Then, Eq. (6) is modified to reflect the segment dependent design matrix $\boldsymbol{Z}_k$, for kth segment, as follows:

$$\hat{\boldsymbol{B}}_k = [\boldsymbol{Z}_k'\boldsymbol{Z}_k]^{-1}\boldsymbol{Z}_k\boldsymbol{C}_k. \quad (24)$$

Since each model is represented by a single state, the estimation of the mean trajectory can be identical to that of the polynomial segmental model (Gish and Ng, 1996),

$$\hat{\mathscr{B}} = \left[\sum_{k=1}^{K}\boldsymbol{Z}_k'\boldsymbol{Z}_k\right]^{-1}\left[\sum_{k=1}^{K}\boldsymbol{Z}_k'\boldsymbol{Z}_k\hat{\boldsymbol{B}}_k\right], \quad (25)$$

where $\hat{\boldsymbol{B}}_k$ is the estimate of the trajectory parameters obtained from the kth segment with the different design matrix $\boldsymbol{Z}_k$. If the mean trajectory coefficient $\mathscr{B}$ and its estimation process are adjusted as in the above equation, SFHMM resembles the parametric trajectory model.

4. Experiments

SFHMMs were evaluated on two tasks in automatic speech recognition: phoneme recognition (39 phone classes), and vowel classification (16 vowels). In the phoneme recognition task, we examined the properties of segmental features by comparing the proposed method with conventional HMM in a combination of stationary and dynamic features. We conducted another experiment to find the characteristics of segment length and regression order of SFHMM on the same task. In another task, we checked the different representations of variance in a state.

The same front-end was applied to both tasks. For signal processing, waveforms were preemphasized and blocked into 20-ms frames at every 10-ms interval. For feature extraction, frames were Hamming-windowed and transformed by a Fourier transform. The output was converted to a mel scale with 24 channels and a cosine transform was applied. The first 12 cosine coefficients together with the normalized log energy value were used to obtain the segmental feature of SFHMM or for inputs to conventional HMMs.

The TIMIT database was used for all experiments. From the 630 available speakers, 462 speakers were used for training and the remaining 168 speakers for tests. Thus, there were a total of 4620 training sentences and 1,680 test sentences (10 sentences from each speaker including SA sentences).

4.1. Phoneme recognition

In the phoneme recognition task, the basic inventory of model units was defined as the set of the 48 symbols that are used in the time-aligned phonetic transcriptions provided with TIMIT. After recognition, the 48-symbol set was folded into 39 categories, as in (Lee and Hon, 1989). Experiments were carried out with context-independent ("monophone") models, and a phone bigram language model was also used.

4.1.1. Properties of segmental features

The aim of these experiments was to compare the performance of models using segmental features with that of conventional HMM using frame features. The performance of SFHMM was compared with that of the baseline system in terms of various feature combinations. The baseline system is an SFHMM that has unit segment length (single frame) and unit regression order. As mentioned in the previous section, the SFHMM with $N = 1$ and $R = 1$ is identical to the standard HMM. Thus, we compared the performance of systems with non-unit segmental unit and regression order with that of the baseline system. In these experiments, the segmental feature is represented by the trajectory when the segmental length is greater than one and the regression order is also greater than one. The trajectory is obtained from the regression functions of basic features.

We carried out two experiments with different feature sets. First, the baseline system used stationary (12-MFCC and log energy) and dynamic features (the first derivatives of the stationary features), while the SFHMM used segmental features based on the stationary features. Table 1 shows the performances (accuracies) of the baseline and SFHMM systems according to different features.

Table 1
The phoneme recognition accuracy I of SFHMM and conventional HMM

Systems	Single mixture	Two mixtures
Baseline	52.8	56.1
$N = 3,\ R = 2$	51.0	54.0
$N = 3,\ R = 3$	51.5	54.3
$N = 5,\ R = 2$	51.6	54.6
$N = 5,\ R = 3$	52.9	55.8
$N = 5,\ R = 4$	53.1	56.3
$N = 5,\ R = 5$	53.2	56.4

The baseline system uses stationary features (12-MFCC and log energy) and their first derivatives, while the SFHMM uses trajectories based on the stationary features.

In these experiments, if the segment length and the regression order are equal, then goodness-of-fit (estimation error of trajectory) is zero. When the segment length is five and the regression order is greater than three, the SFHMM outperforms the conventional HMM.

We conducted an additional experiment, expanding the basic feature set for the baseline system and for the trajectory features of SFHMM. The baseline system used second derivatives of the stationary features, and the trajectories of the SFHMM were calculated by including the first derivatives of the stationary features. The recognition results are shown in Table 2.

The proposed method always outperforms the baseline system in these experiments. From the experimental results, we believe that the proposed

Table 2
The phoneme recognition accuracy II of SFHMM and of conventional HMM

Systems	Single mixture	Two mixtures
Baseline	52.6	57.0
$N = 3,\ R = 2$	54.4	58.1
$N = 3,\ R = 3$	54.7	58.5
$N = 5,\ R = 2$	54.6	58.7
$N = 5,\ R = 3$	55.6	59.9
$N = 5,\ R = 4$	55.6	60.1
$N = 5,\ R = 5$	55.6	60.1

The baseline system uses stationary features and their first and second derivatives, while the trajectory for the SFHMM is obtained from the stationary features and their first derivatives.

approach, using trajectories based on stationary features, can represent the dynamics of a speech signal as effectively as a method using dynamic features. When we considered the trajectory on the first derivatives of the stationary features, we obtained better results than the baseline system including the second derivatives of the stationary features. Therefore, in the following experiment, SFHMM will use the trajectory based on both stationary and dynamic features.

4.1.2. Characteristics of segment length and regression order

We found that the segmental features based on both stationary and dynamic features could represent speech patterns effectively in the previous experiment. In this section, we will examine the characteristics of segment length and regression order in representing the trajectories.

The baseline systems used 26 feature vectors (12-MFCC, log energy and their first derivatives) and these vectors derived the segmental features (polynomial trajectories). As mentioned in the previous section, "baseline system" refers to the SFHMM with $N = 1$ and $R = 1$. In this condition, the baseline system is identical to conventional HMM. The other systems are the SFHMM with different segment lengths (N) and regression orders (R).

Several experiments were performed to find the effect of different segment lengths and regression orders. The experimental environments were the same as the previous experiment and the recognition results are reported in Tables 3 and 4.

Better performance was achieved when the segment length was longer and the regression order was higher in SFHMM. It may be inferred from these results that the different condition of the segment length or regression order affects the performance. The percent correct rose and the substitution error fell if the regression order increased for the same segment length. This means that discrimination increased by higher regression order in the same segment length. However, if the segment length increased with the same regression order, the insertion and substitution errors decreased, but the deletion error increased. In this case, better accuracies can be obtained with the increase in deletion error because the sum of reduced errors is greater than the increased error. It is therefore considered that the amount of transi-

Table 3
Overall recognition results with *single* mixture for different segment lengths N and regression orders R

SFHMM	%Corr.	%Acc.	%Subs.	%Del.	%Ins.	%Err.
Baseline	58.0	52.8	29.5	12.5	5.2	47.2
$N = 3,\ R = 2$	59.4	54.4	28.5	12.1	4.9	45.6
$N = 3,\ R = 3$	59.6	54.7	28.3	12.1	4.9	45.3
$N = 5,\ R = 2$	58.9	54.6	28.0	13.1	4.3	45.4
$N = 5,\ R = 3$	60.0	55.6	27.5	12.5	4.4	44.4
$N = 5,\ R = 4$	60.1	55.6	27.4	12.5	4.5	44.4
$N = 5,\ R = 5$	60.1	55.6	27.4	12.6	4.5	44.4

Table 4
Overall recognition results with *two* mixtures for different segment lengths N and regression orders R

SFHMM	%Corr.	%Acc.	%Subs.	%Del.	%Ins.	%Err.
Baseline	62.5	56.1	27.3	10.2	6.4	43.9
$N = 3,\ R = 2$	63.7	58.1	26.3	10.0	5.6	41.9
$N = 3,\ R = 3$	64.1	58.5	26.1	9.9	5.6	41.5
$N = 5,\ R = 2$	63.8	58.7	25.7	10.5	5.1	41.3
$N = 5,\ R = 3$	65.0	59.9	24.9	10.1	5.1	40.1
$N = 5,\ R = 4$	65.3	60.1	24.8	9.9	5.2	39.9
$N = 5,\ R = 5$	65.2	60.1	24.8	10.0	5.1	39.9

tional information grows if the segment is wider for the same regression order because of a reduction in insertion and substitution errors.

Relative to the baseline system, the reduction in error rate ranges from 3.3% to 5.9% for a linear trajectory system ($R = 2$), and from 4.0% to 8.7% with a quadratic trajectory system ($R = 3$). From the above results, if the segment is long, its regression order is high, and the number of mixtures for state modeling is small, the performance does not improve any further. This result indicates that if a long segment is modeled by a higher regression order, its variability increases and more mixtures are necessary for appropriate modeling. In Table 3, when the segment length is five and the regression order is greater than two, accuracies are the same. However, if we increase the number of mixtures, the accuracies are the same when the regression order is greater than three in Table 4. Hence, if we use a large number of mixtures, the performance of SFHMM with a different regression order may be different.

4.2. Vowel classification

In this section, we describe experiments for examining the effect of different variance estimation techniques. To compare the performances of SFHMM with different variance representations, we chose the vowel classification task so that the experimental environment is similar to that of previous work (Gish and Ng, 1996; Fukada et al., 1997). For the vowel classification task, 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. In this experiment, we used the segmental features based on a 26-dimensional basic feature (12-MFCC, log energy and their first derivatives). After the tokens were extracted from training and testing sentences of the TIMIT corpus, 41,429 tokens were employed for training and 15,119 tokens were used for testing.

We conducted the experiments by changing the number of mixtures, segment length and regression order for SFHMM. If the time-varying variance approach was adopted for SFHMM, variance was calculated for each frame of segments occurring in a state. However, if the SFHMM used the fixed variance approach, the variance of a segment was calculated by averaging the variances of frames in it. Table 5 shows the performance of SFHMM with time-varying variance and the recognition results of SFHMM with fixed variance are shown in Table 6.

In these tables, we obtained 61.1% correct performance if two mixtures were used. This result is less than those of the previous work (Gish and Ng (1996) reported 66.06% for best performance and Fukada et al. (1997)'s result is 66.2%). Even if our system shows low performance, these results can be relatively comparable because the experimental environments, such as the data set, the number of mixtures, and the regression order of the mean trajectories, are mismatched. Gish and Ng (1996) used full covariance matrix and eight mixtures, and Fukada et al. (1997) employed up to nine mixtures for their best performance.

Table 5
Vowel classification performance of SFHMM with time-varying variance

Systems	Single mixture	Two mixtures
Baseline	55.23	58.44
$N = 3,\ R = 2$	56.86	59.56
$N = 3,\ R = 3$	57.21	59.64
$N = 5,\ R = 2$	58.77	60.43
$N = 5,\ R = 3$	58.79	60.84
$N = 5,\ R = 4$	58.76	60.88
$N = 5,\ R = 5$	58.74	60.86

Table 6
Vowel classification performance of SFHMM with fixed variance

Systems	Single mixture	Two mixtures
Baseline	55.23	58.44
$N = 3,\ R = 2$	56.12	59.90
$N = 3,\ R = 3$	56.12	59.87
$N = 5,\ R = 2$	56.99	59.90
$N = 5,\ R = 3$	56.86	60.14
$N = 5,\ R = 4$	56.70	61.10
$N = 5,\ R = 5$	56.68	60.94

When the time-varying variances were applied to SFHMM with a single mixture, its performance was improved by varying the segment length and regression order. If the SFHMM took the fixed variance approach with a single mixture, the performance improvement was not as high as when the time-varying variance approach was used. However, if we increased the number of mixtures to describe SFHMMs, the difference of performance between the two approaches was very small. The recognition results of the SFHMM with fixed variance were even higher than those of the SFHMM with time-varying variance were.

This result is in contrast to results of previous work. We think that this phenomenon was caused by small segment lengths. In the previous work, the complete trajectory was used to model the segment and the segment corresponded to an acoustic model such as a phone. Thus, the segment length was longer than that of SFHMM. If the segment is wider, fixed variance cannot represent the variability of a segment. However, if the segment is as narrow as that of SFHMM, common variance can express its variability when more mixtures are used. As a result, we examined whether, if a segment length is small (smaller units are used) and multiple mixtures are used to represent SFHMM, the fixed variance approach could be applied to SFHMM without loss of performance.

4.3. Discussion

The experiments described in this section have demonstrated that SFHMMs can perform better than HMMs both for phoneme recognition and for vowel classification. We performed experiments to find the characteristics of segmental features by comparison between conventional HMM and SFHMM. Conventional HMM uses stationary and dynamic features together, while SFHMM uses segmental features based on the stationary features. Comparison shows that SFHMM with segmental features performs better than conventional HMM when longer segmental lengths and higher regression orders are used. If SFHMM uses the polynomial segmental feature based on both stationary and dynamic features, the performance improvement is high. Thus, we consider that the segmental features are effective in speech processing.

We also conducted experiments on the phoneme recognition task to examine the characteristics of SFHMM by changing the segment length and regression order. From the experiments, we obtained some interesting results: discrimination power increases if the regression order increases for the same segment length, and the amount of transitional information grows if the segment length is long for the same regression order in SFHMM.

Finally, we tried to reduce the number of free parameters without loss of performance. In this experiment, we compared performance using different variance estimation techniques. The time-varying variance approach represents the variance of frames of segments occurring in a state individually. In contrast, the fixed variance approach considers that the variance of a state can be obtained by averaging all variances of frames in segments. The two approaches were compared using SFHMM. In the experimental results, if the single mixture was used to model the state of SFHMM, the performance with time-varying variance was better than that with fixed variance. However, when multiple mixtures were used, the difference in performance was small. Hence, we can argue that the fixed variance approach is effective when the segment length is small.

Even though improvements in recognition performance have been demonstrated with the proposed method, one may suspect that the performance improvement is due to more free parameters. Comparisons of the numbers of free parameters are shown in Table 7.

Table 7
Comparison of the number of free parameters in a state for conventional HMM and SFHMMs (D = feature dimension, R = regression order, N = segment length, V = variance dimension: $D \times D$ or $1 \times D$ by full or diagonal covariance)

Systems	For mean	For variance
Conventional HMM	$1 \times D$	V
SFHMM (time-varying variance)	$R \times D$	$N \times V$
SFHMM (fixed variance)	$R \times D$	V

In this table, we do not consider the design matrix because it has a fixed number of rows and columns and each element is automatically determined if segment length and regression order are defined. The number of free parameters of SFHMM is greater than that of conventional HMM when the fixed variance or time-varying variance approaches are used. However, the segmental feature can include more information than stationary and dynamic features because the dynamic features represent the average variabilities for a set of frames, while the segmental features consider the trend of features in a segment.

We can generally set the number of parameters of conventional HMM to be equal to that of SFHMM by adjusting the dimensionality of the feature vectors. If the HMMs use the first-time derivative features and the SFHMMs use only static features with the regression order of two and fixed variance, the number of parameters in the models is the same. In this condition, we can compare the relative performance of the models without performing additional experiments. From the previous experiments, we found that the performance of fixed variance model is similar to that of time-varying variance system when two mixtures are used. Thus, we can guess the performance of the fixed variance SFHMM in Table 1. As a result, the performance of the SFHMM is worse than that of the HMM in case of the same number of parameters. This result can be explained by the variability of the mean trajectory. The time-derivative feature is expressed as a single term indicating the average variability of a segment and the segmental feature is represented by the polynomial trajectory. This means that the variability of trajectories is represented by the variabilities of multiple frame features and the variability of mean trajectories in each mixture of SFHMM is higher than that of mean vectors in mixtures of HMM. The difference between the performances of both models can be decreased by increasing the number of mixtures or model units to reduce the average variability in each mixture of SFHMMs. Therefore, though the worse results are currently obtained with the same number of free parameters, studies for new models that have more information are necessary and they can be meaningful. From this point of view, we think that the proposed method is valuable even if the number of parameters is greater than that of conventional HMM.

5. Conclusions

We have proposed a new modeling method for continuous speech recognition that uses segmental features represented by polynomial trajectories. By selecting a design matrix that is symmetric with respect to the current frame, our approach can represent the transitional information between adjacent units. For SFHMM, we present a basic mathematical analysis and a Baum-Welch-like parameter re-estimation formula within HMM concepts. Thus, this method can be regarded as an extension or a generalization of a conventional HMM and a parametric trajectory model.

We conducted several experiments to find the effectiveness of SFHMM and the characteristics of the segmental features based on polynomial regression functions. We first compared the recognition results of conventional HMM using stationary and dynamic features with the performance of SFHMM using segmental features based on the stationary features. The experimental results show that the proposed segmental features are as effective as the combination of stationary and dynamic features.

To obtain the characteristics of the variability of a segment, experiments using the phoneme recognition task were performed using changes of the segment length and the regression order in SFHMM. The recognition results showed that discrimination power increases if the regression order is higher for the same segment length and that the amount of transitional information grows if the segment is wider for the same regression order in SFHMM. The proposed method gave consistently better performances than conventional HMM, if the segmental window was wide and the regression order was high. Thus, we think that SFHMM can represent the temporal correlation between speech frames and partial context-dependent (CD) information, i.e. transitional information, if explicit CD modeling is not used.

In the last experiments, we tried to reduce the number of parameters without loss of performance. To reduce the number of parameters for representing a segment, the fixed variance approach is considered to represent all variances of frames of segments occurring in a state as an average (or common) variance, in contrast to the time-varying variance approach in which each frame in a segment has an individual variance. The experimental results with different variance estimations show that the approach reflecting the temporal variability of frames has better performance than that of constant variance with single mixture. However, if the number of mixtures is increased, the difference of recognition results between the two approaches is small. Thus, we believe that the fixed variance approach is effective on small segment units.

Acknowledgements

We would like to thank the reviewers for their valuable remarks and suggestions.

References

Deng, L., 1992. A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal. Signal Process. 27, 65–78.

Deng, L., Aksmanovic, M., Sun, D., Wu, J., 1994. Speech recognition using hidden Markov models with polynomial regression functions as non-stationary states. IEEE Trans. Speech Audio Proc. 2 (4), 507–520.

Fukada, T., Sagisaka, Y., Paliwal, K., April 1997. Model parameter estimation for mixture density polynomial segment models. In: Internat. Conf. on Acoust. Speech Signal Process. 1997, Munich, Germany.

Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. Acoust. Speech Signal Process. 34 (1), 52–59.

Gales, M., Young, S., June 1993. The theory of segmental hidden Markov models. CUED/F-INFENG/TR 133, Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, England.

Gish, H., Ng, K., 1993. A segmental speech model with application to word spotting. In: Internat. Conf. on Acoust. Speech Signal Process. 1993, Minneapolis, MN.

Gish, H., Ng, K., October 1996. Parametric trajectory models for speech recognition. In: Internat. Conf. on Spoken Lang. Process. 1996, Philadelphia, PA.

Goldenthal, W., Glass, J., 1993. Modeling spectral dynamics for vowel classification. In: Europ. Conf. on Speech Commun. Technol. 1993, Berlin, Germany.

Holmes, W., Russell, M., 1995a. Experimental evaluation of segmental HMMs. In: Internat. Conf. on Acoust. Speech Signal Process. 1995, Detroit, MI.

Holmes, W., Russell, M., 1995b. Speech recognition using a linear dynamic segmental HMM. In: Internat. Conf. on Acoust. Speech Signal Process. 1995, Detroit, MI.

Holmes, W., Russell, M., 1999. Probabilistic-trajectory segmental HMMs. Comput. Speech Lang. 13, 3–37.

Lee, K., Hon, H., 1989. Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. 37 (11), 1648–1661.

Ostendorf, M., Digalakis, V., Kimball, O., 1996. From HMMs to segment models: a unified view of stochastic modeling for speech recognition. IEEE Trans. Speech Audio Process. 4 (5), 360–378.

Press, W., Teukolsky, A., Vetterling, W., Flannery, B., 1992. Numerical Recipes in C, 2nd Ed. Cambridge University Press, Cambridge.

Russell, M., 1993. A segmental HMM for speech pattern modeling. In: Internat. Conf. on Acoust. Speech Signal Process. 1993, Minneapolis, MN.