TR-S-0025

A Study of Speech Recognition based on Segmental Feature Model

松田 繁樹 Shigeki Matsuda
クルディップ パリワル Kuldip Paliwal
中村 哲 Satoshi Nakamura

2001. 3.30

We introduce a segmental feature model (SFM) that represents temporal relationships between feature vectors. A feature vector sequence can be divided into most likely periods by using the conventional HMM. In the conventional HMM, temporal relationships between these periods are represented, because the conventional HMM consists of plural states connected temporarily. However, temporal relationships between feature vectors in each period is not modeled. If the temporal relationships between the feature vectors are modeled, it is considered that feature vector sequences can be modeled more efficiently than the conventional HMM.

The SFM calculate a probability of a fixed-dimension segmental feature vector, the segmental feature vector is extracted from a variable-length period that is allocated to each state in the SFM. We propose a segmental feature vector based on average values. The segmental feature vector can calculate temporal covariances. And, we propose a new SFM that has variances in a segment (period), to reduce missmatches between a feature vector sequence and a segmental feature vector.

For the SFM using the segmental feature vector based on average values, we performed speech recognition experiments of a phoneme classification and a continuous phoneme recognition. The SFMs achieved higher recognition rates than conventional HMMs in the phoneme classification experiments. However, in the continuous phoneme classification experiments, the SFMs got lower recognition rates than conventional HMMs. It is considered that the SFM does not estimate phoneme boundaries rightly.

Contents

Chapter 1

Introduction

An HMM is a model that represents non-stationary signals such as acoustic feature vector sequences by changing states, each state has one stationary distribution. Figure 1.1 is an example of a feature vector sequence modeled by using three stationary distributions. As equation 1.1, we can obtain a likelihood $P(O|Q,\lambda)$ of the feature vector sequence by calculating a product of probabilities $P(o_t|s_i)$. In the equation, let λ be parameters of an HMM. Let Q be a state sequence $\{(b_1,e_1),(b_2,e_2),\cdots\}$, where (b_i,e_i) represents that a state s_i covers a period b_i to e_i in a whole feature vector sequence O. Let o_t be a feature vector observed at time t. A likelihood outputted from an HMM depends on the state sequence. The feature vector sequence can be divided into most likely periods by finding a optimal state sequence that obtains a maximum likelihood. The optimal state sequence can be obtained simply by using a Viterbi algorithm.

$$P(O|Q,\lambda) = \prod_i \prod_{t=b_i}^{e_i} P(o_t|s_i) \tag{1.1}$$

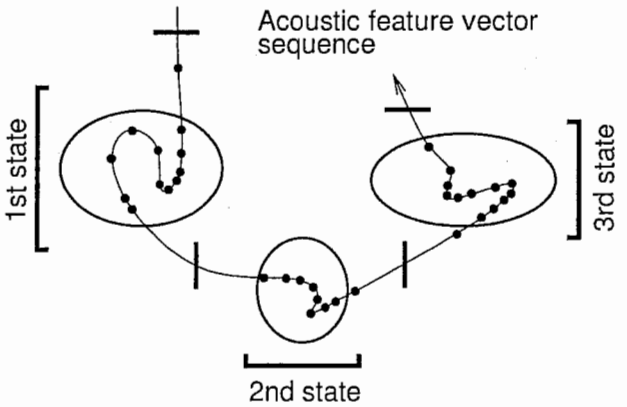

Figure 1.1: An example of a feature vector sequence modeled by three states.

A conventional HMM can calculate temporal relationships between the individual periods, because a conventional HMM is a model that consists of plural states connected temporarily. However, temporal relationships between feature vectors allocated to a state are not modeled, since a probability outputted from the state is obtained by calculating a product of individual probabilities $P(o_t|s)$. If individual states are allocated for individual feature vectors respectively, the temporal relationships can be also modeled by a conventional HMM. However, it is not a realistic solution, it has a problem of an increase of parameters.

A segment model (SM) [Ostendorf 96] was proposed as a technique to model temporal relationships between feature vectors. The SM can represent the temporal relationships efficiently without using a large number of parameters. A conventional HMM is a model assuming that a mean value of a distribution is invariable in a state. However, the SM assumes that the mean value varies with time. As equation 1.2, a probability outputted from a state depends on a temporal order of feature vectors allocated to the state. Therefore, it is considered that a likelihood outputted from the SM reflects temporal relationships between feature vectors.

$$\mathcal{N}(o_t, \mu(0), \sigma)\mathcal{N}(o_{t+1}, \mu(1), \sigma) \neq \mathcal{N}(o_{t+1}, \mu(0), \sigma)\mathcal{N}(o_t, \mu(1), \sigma) \tag{1.2}$$

In this report, we introduce a segmental feature model (SFM)[Ostendorf 96] as a technique to represent temporal relationships between feature vectors efficiently. As equation 1.3, each state in the SFM calculates a probability of a segmental feature vector. Where f represents a function for extracting fixed-dimension segmental feature vector from a variable-length feature vector sequence. It is considered that temporal relationships between feature vectors can be represented efficiently by choosing an appropriate extraction function f.

$$P(O|Q, \lambda) = \prod_i P(f(O_{b_i}^{e_i})|s_i) \tag{1.3}$$

We propose a segmental feature vector based on average values. The segmental feature vector can calculate a temporal covariance. Moreover, we propose a new SFM that has variances in a segment. These SFMs are evaluated by performing speaker-dependent experiments of a phoneme classification and a continuous phoneme recognition.

In section 2.1 and 2.2, we introduce an SM and an SFM as techniques to model temporal relationships between feature vectors. In section 2.3, two new segmental feature vectors based on average values and a KL extraction are proposed. In section 2.4, these SFMs using proposed segmental feature vectors are evaluated in experiments of a phoneme classification and a continuous phoneme recognition. In section 3.1, we propose a new SFM using a likelihood of variances in a segment. The new SFM is evaluated in section 3.2. Section 4 is a conclusion.

Chapter 2

Segmental Feature Model

In this section, we describe two techniques, a segment model (SM) and a segmental feature model (SFM), to model temporal relationships between feature vectors. Especially, we discuss the SFM that can model the temporal relationships more efficiently than the SM.

2.1 Modeling of Temporal Relationships

We introduce a segment model (SM) and a segmental feature model (SFM) as a technique to model temporal relationships.

2.1.1 Segment Model

The SM is a model that calculates temporal relationships between feature vectors efficiently. A mean value of a distribution in the SM varies with time, though the mean value is invariable in a conventional HMM. Figure 2.1 illustrates a concept of the SM.

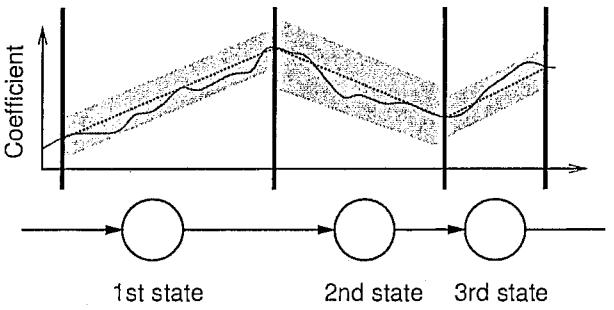

Figure 2.1: A concept of the segment model.

A linear regression and a polynomial function are utilized as a function of the mean value that varies with time, each state has the time-depended mean value individually. A probability outputted from a state in the SM is calculated as equation 2.1. When

a trajectory of a time-depended mean value is similar to a behavior of a feature vector sequence allocated to a state, a large likelihood is outputted from the state. The function $\mu_i(L, \tau)$ calculates a mean value at a relative time τ, the L means a duration length of a period allocated to a state.

$$P(O|Q,\lambda) = \prod_i \prod_{t=b_i}^{e_i} P(o_t|\mu_i(e_i - b_i + 1, t - b_i), s_i) \tag{2.1}$$

A probability outputted from a state in the SM depends on a temporal order of feature vectors allocated to the state, since the state has a distribution with a time-depended mean value. Therefore, it is considered that a likelihood outputted from the SM reflects temporal relationships between feature vectors. Moreover, temporal relationships between feature vectors are modeled efficiently by using a few number of parameters without using lots of states in a conventional HMM.

2.1.2 Segmental Feature Model

A state in the SFM outputs a probability of a segmental feature vector, the segmental feature vector is extracted from a feature vector sequence allocated to the state. Figure 2.2 illustrates a structure of the SFM. As equation 2.2, a likelihood outputted from the SFM is calculated. The f denotes a function for extracting a fixed-dimension segmental feature vector from a variable-length feature vector sequence.

$$P(O|Q,\lambda) = \prod_i P(f(O_{b_i}^{e_i})|s_i) \tag{2.2}$$

In a difference between the SFM and the SM, a state in the SFM generates one fixed-dimension segmental feature vector, though a variable-length feature vector sequence is generated from a state in the SM. It is considered that the SFM models temporal relationships between feature vectors more efficiently than the SM by choosing an appropriate function f.

2.2 Segmental Feature Model

A structure of the SFM is similar to a conventional HMM with a duration control [Ferguson 80]. Figure 2.3 illustrates a structure of a left-to-right SFM. A likelihood outputted from the SFM is calculated by using equation 2.3.

$$P(O|Q,\lambda) = \prod_{i=1}^{N} d_i(e_i - b_i + 1) b_i(f(O_{b_i}^{e_i})) \tag{2.3}$$

A likelihood outputted from the SFM depends on a number of states, because a state in the SFM outputs one probability only. However the likelihood need to depend on time. We can obtain a likelihood that depends on time by calculating a power of τ for a probability outputted from a state.

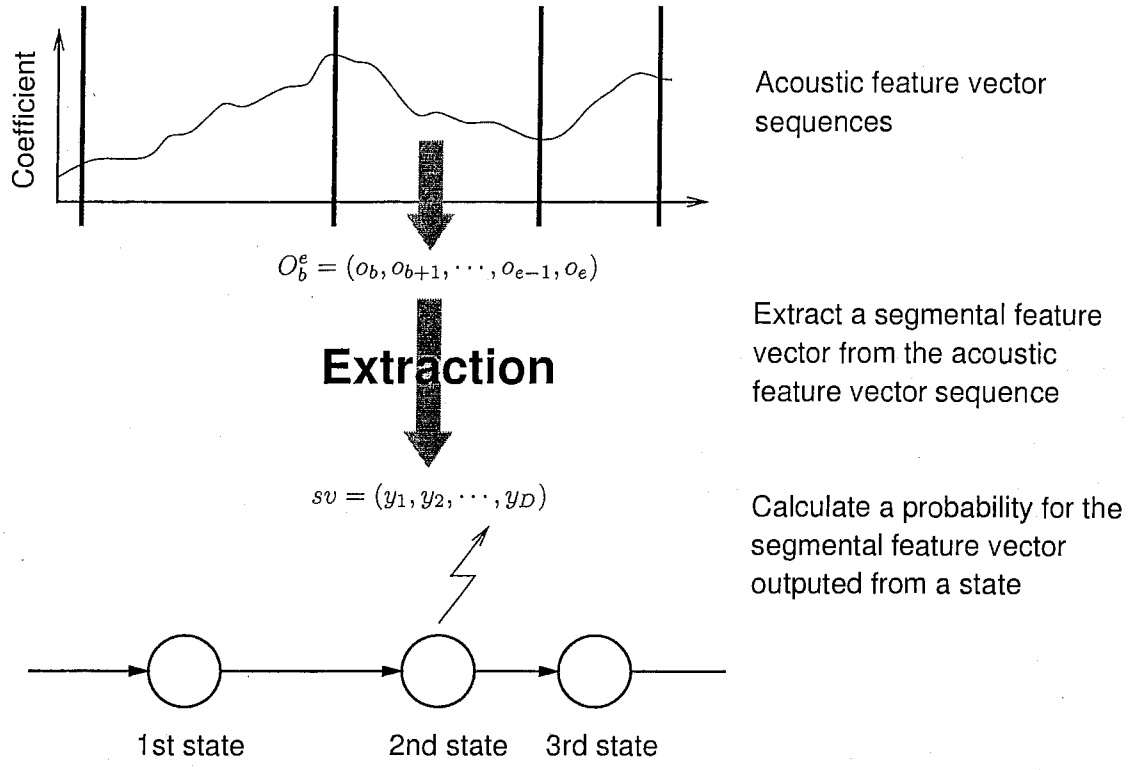

Figure 2.2: A structure of a segmental feature model.

$\lambda = \{S, D, B\}$	Model parameters of an SFM
$S = \{s_i\}$	Set of states $(0 \leq i \leq N-1)$
$D = \{d_i(\tau)\}$	Set of duration controls
$B = \{b_i(sv)\}$	Set of distributions $(sv \in Y)$

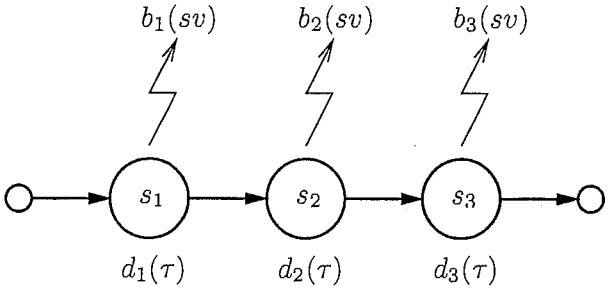

Figure 2.3: A structure of the segmental feature model

Therefor, a probability of a duration control $d_i(\tau)$ and a distribution $b_i(sv)$ are calculated as equation 2.5 and equation ??. In this report, one dimensional gaussian distribution is used for a duration control. And multi-dimension gaussian distribution is done for a distribution.

$$d_i(\tau) = \mathcal{N}(\tau, \mu_d(i), \sigma_d(i))^\tau \tag{2.4}$$

$$b_i(sv) = \mathcal{N}(sv, \mu_b(i), \sigma_b(i))^\tau \tag{2.5}$$

If a distribution is represented by using mixture gaussian distributions, there are two calculation ways as equation 2.6, 1)SMIX (Segment MIXture) and 2)FMIX (Frame MIXture). Figure 2.4 illustrates concepts of the SMIX and the FMIX. In case of the SMIX, a power of τ for each probability of a gaussian distribution is calculated. Then, a sum of these probabilities is done. In case of the FMIX, a sum of probabilities of each mixture component is calculated. Then a power of τ for the sum is calculated. In this report, these calculation ways are evaluated in speech recognition experiments.

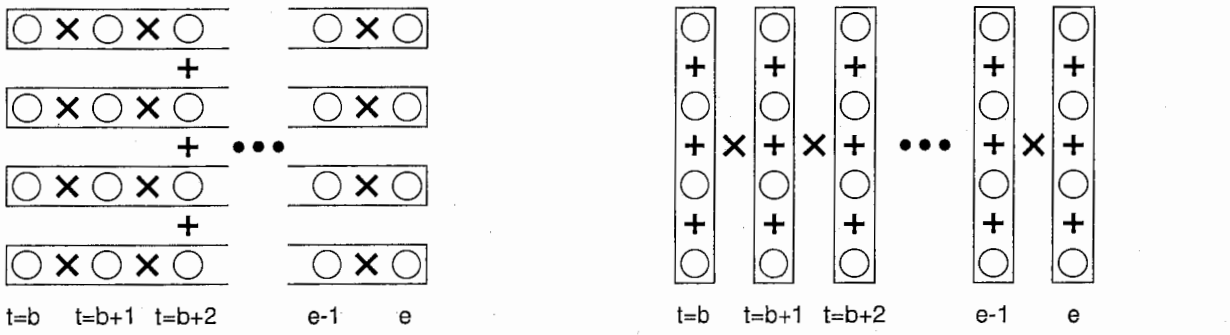

Figure 2.4: Two mixture calculation ways

$$b_i(sv) = \begin{cases} \sum_{m=1}^{M} c_m(i)\mathcal{N}(sv, \mu_m(i), \sigma_m(i))^\tau & (SMIX : Segment\ MIXture) \\ \prod_{t=0}^{\tau} \sum_{m=1}^{M} c_m(i)\mathcal{N}(sv, \mu_m(i), \sigma_m(i)) & (FMIX : Frame\ MIXture) \end{cases} \tag{2.6}$$

2.2.1 Optimal State Sequence

To utilize the SFM for an actual speech recognition, we need to calculate an optimal state sequence. An algorithm obtaining the optimal state sequence is similar to Viterbi algorithm for a conventional HMM with a duration control. The optimal state sequence Q is obtained by following processes. Where λ represents parameters of an SFM. And, let O be a feature vector sequence.

- Searching:
  - Initialization: $(1 \leq t \leq T)$

$$\begin{aligned} \delta_1(t) &= d_1(t)b_1(f(O_1^t)) \\ \psi_1(t) &= 1 \end{aligned}$$

  - Recursion: $(2 \leq i \leq N)$, $(1 \leq t \leq T)$

$$\begin{aligned} \delta_i(t) &= \max_{\tau} \delta_{i-1}(t-\tau)d_i(\tau)b_i(f(O_{t-\tau}^t)) \\ \psi_i(t) &= \arg\max_{\tau} \delta_{i-1}(t-\tau)d_i(\tau)b_i(f(O_{t-\tau}^t)) \end{aligned}$$

- Backtracking:
  - Initialization:

$$\begin{aligned} b_N &= \psi_N(T) \\ e_N &= T \end{aligned}$$

  - Recursion: $(i = N-1, \cdots, 1)$

$$\begin{aligned} b_i &= \psi_i(b_{i+1} - 1) \\ e_i &= b_{i+1} - 1 \end{aligned}$$

2.2.2 Parameter Estimation

Parameters of the SFM can be estimated by the EM algorithm [Dempster 77] as well as a conventional HMM. In this report, we simply introduce a Viterbi training. Where R represents a number of feature vector sequences for training. Let $b_i(r)$ and $e_i(r)$ be start frame number and end frame number of a period that is allocated to a state i for an feature vector sequence r.

- E step:
  In the E step, for a current model parameter λ, optimal state sequences Q is calculated by the Viterbi algorithm.

- M step:
  In the M step, a new model parameter $\hat{\lambda}$ is estimated by using the Q.

  1. Estimation for duration controls D.

$$\begin{aligned} \hat{\mu_d}(i) &= \frac{\sum_{r=1}^{R}(e_i(r) - b_i(r) + 1)(e_i(r) - b_i(r) + 1)}{\sum_{r=1}^{R} e_i(r) - b_i(r) + 1} \\ \hat{\sigma_d}(i) &= \frac{\sum_{r=1}^{R}(\hat{\mu_d}(i) - (e_i(r) - b_i(r) + 1))^2(e_i(r) - b_i(r) + 1)}{\sum_{r=1}^{R} e_i(r) - b_i(r) + 1} \end{aligned}$$

2. Estimation for distributions B.

$$\hat{\mu_b}(i) = \frac{\sum_{r=1}^{R} \zeta_m(i) f(O_{b_i}^{e_i}(r))(e_i - b_i + 1)}{\sum_{r=1}^{R} \zeta_m(i)(e_i(r) - b_i(r) + 1)} \tag{2.7}$$

$$\hat{\sigma_b}(i) = \frac{\sum_{r=1}^{R} \zeta_m(i)(\hat{\mu_b}(i) - f(O_{b_i}^{e_i}(r)))^2(e_i - b_i + 1)}{\sum_{r=1}^{R} \zeta_m(i)(e_i(r) - b_i(r) + 1)} \tag{2.8}$$

- Segment Mixture (SMIX)

$$\zeta_m(i) = \frac{c_m(i)\mathcal{N}(f(O_{b_i}^{e_i}(r)), \mu_m(i), \sigma_m(i))^\tau}{\sum_{m=1}^{M} c_m(i)\mathcal{N}(f(O_{b_i}^{e_i}(r)), \mu_m(i), \sigma_m(i))^\tau}$$

$$\hat{c}_m(i) = \frac{\sum_{r=1}^{R} \zeta_m(i)}{R}$$

- Frame Mixture (FMIX)

$$\zeta_m(i) = \frac{c_m(i)\mathcal{N}(f(O_{b_i}^{e_i}(r)), \mu_m(i), \sigma_m(i))}{\sum_{m=1}^{M} c_m(i)\mathcal{N}(f(O_{b_i}^{e_i}(r)), \mu_m(i), \sigma_m(i))}$$

$$\hat{c}_m(i) = \frac{\sum_{r=1}^{R} \zeta_m(i)(e_i(r) - b_i(r) + 1)}{\sum_{r=1}^{R}(e_i(r) - b_i(r) + 1)}$$

Parameters of the SFM can be estimated by performing the E step and the M step repeatedly.

2.3 Segmental Feature Vector

We propose a segmental feature vector based on average values and a segmental feature vector based on a KL extraction.

2.3.1 Segmental Feature Vector based on Average Values

We propose a segmental feature vector based on average values. The segmental feature vector consists of average values, each average value is calculated from a sub-period p_i. Individual sub-periods are obtained by splitting a period b to e at equal intervals, the period is a feature vector sequence allocated to a state. Figure 2.5 illustrates a concept of the segmental feature vector. In this figure, a split resolution used $N = 3$, and a number of dimensions of a feature vector sequence was $D = 2$. By using the way, the two-dimensional feature vector sequence is converted to the six-dimensional segmental feature vector. A length of the feature vector sequence is variable.

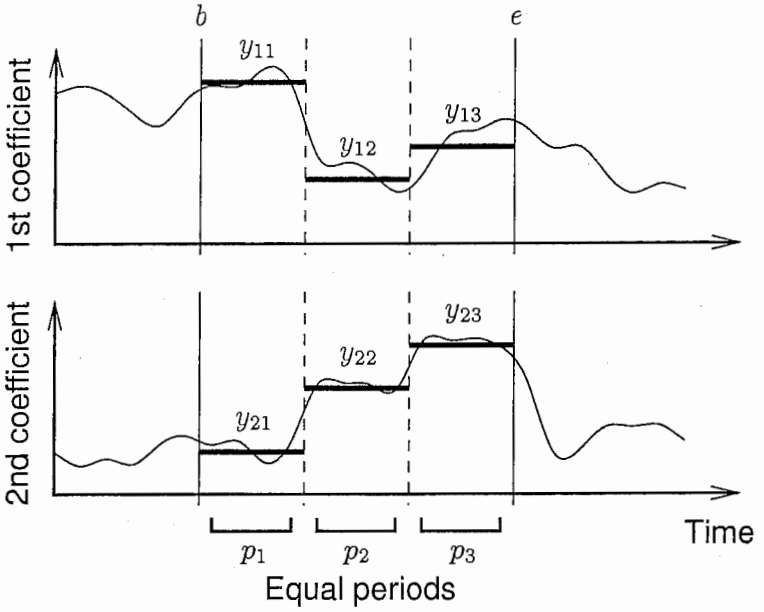

Figure 2.5: Segmental feature using average values

$$(y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23}) = f(O_b^e) \tag{2.9}$$

This segmental feature vector can calculate temporal covariance. As equation 2.10, the temporal covariance can be calculated by using a beltlike covariance matrix, since a variable-length feature vector sequence is converted to $N \times D$-dimensional segmental feature vector.

$$C = \begin{bmatrix} c_{11,11} & c_{11,12} & c_{11,13} & 0 & 0 & 0 \\ c_{12,11} & c_{12,12} & c_{12,13} & 0 & 0 & 0 \\ c_{13,11} & c_{13,12} & c_{13,13} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{21,21} & c_{21,22} & c_{21,23} \\ 0 & 0 & 0 & c_{22,21} & c_{22,22} & c_{22,23} \\ 0 & 0 & 0 & c_{23,21} & c_{23,22} & c_{23,23} \end{bmatrix} \tag{2.10}$$

2.3.2 Segmental Feature Vector based on a KL Extraction

In case of above segmental feature vector based on average values, whenever a time resolution N increases, number of dimensions of the segmental feature vector increases also. The augmentation of number of dimensions leads a degradation of an estimation accuracy of a model. We propose a segmental feature vector based on a KL extraction.

The segmental feature vector is extracted from a segmental feature based on average values by using base vectors, the base vectors are obtained by performing KL extraction.

In this report, we evaluate speech recognition performance of SFMs using the segmental feature vector based on average values.

2.4 Speech Recognition Experiments

We have performed phoneme classification and continuous phoneme recognition experiments to evaluate performance of a speech recognition using SFM.

The ATR word speech database of Japanese important 5240 words uttered by a male speaker (MHT) is used. The half of odd-numbered 2620 words were used for training, and quarter of the even-numbered 655 words were used for testing. The phoneme categories for a recognition were /N, a, b, ch, d, e, f, g, h, i, j, k, m, n, o, p, Q, r, s, sh, t, ts, u, w, j, z/. 12 MFCCs, 12 ΔMFCCs, log-power and Δlog-power extracted with 5ms frame period and 25ms frame length were used as an acoustic feature vector.

We utilize the segmental feature vector based on average values. Each state in the SFM has a partial covariance matrix or a diagonal covariance matrix, a state with the partial covariance matrix can represent a temporal covariance. A probability outputted from state is calculated by using one, two and four mixture distribution.

2.4.1 Phoneme Classification Experiments

The TAUs in table 2.1 are results of the phoneme classification experiment. The ONE mentions later.

Table 2.1: Speaker-Dependent Phoneme Classification Experiments

		Diagonal Covariance				Partial Covariance			
	Conventional	FMIX		SMIX		FMIX		SMIX	
Mixtures	HMM	ONE	TAU	ONE	TAU	ONE	TAU	ONE	TAU
1	83.5%	85.3%	82.9%	85.3%	82.9%	87.6%	86.5%	87.6%	86.4%
2	89.9%	89.4%	88.8%	89.5%	89.2%	91.6%	90.9%	91.1%	90.7%
4	92.5%	92.9%	92.4%	92.8%	91.9%	92.6%	92.7%	92.3%	92.0%
8	94.2%	95.4%	94.8%	95.2%	94.6%	93.4%	92.8%	93.6%	92.6%

From this results, the FMIXs and SMIXs generally got the same recognition rates. In case of a distribution with partial covariances, one, two and three mixture SFMs got higher recognition rates than the conventional HMMs. And, In case of a distribution with diagonal covariances, four mixture SFMs got higher recognition rates than the conventional HMMs.

A counting may of a segmental feature vector in the ONE differ to that in the TAU. Number of counts in the TAU uses a weight of $e - b + 1$ like equation 2.7 and equation 2.8. In other hand, a weight of 1.0 is used in the ONE. The ONE can be calculated by

using equation 2.11 and equation 2.12. The ONE is not based on the maximum likelihood criterion.

SFMs trained by the ONE got higher recognition rate than the TAU. About this reason, it is under examination now.

$$\hat{\mu_b}(i) = \frac{\sum_{r=1}^{R} \zeta_m(i) f(O_{b_i}^{e_i}(r))}{\sum_{r=1}^{R} \zeta_m(i)} \tag{2.11}$$

$$\hat{\sigma_b}(i) = \frac{\sum_{r=1}^{R} \zeta_m(i) (\hat{\mu_b}(i) - f(O_{b_i}^{e_i}(r)))^2}{\sum_{r=1}^{R} \zeta_m(i)} \tag{2.12}$$

2.4.2 Continuous Phoneme Recognition Experiments

We have only to evaluate SFMs trained by the ONE, since the SFMs trained by the ONE got higher recognition rate than the TAU in the phoneme classification experiments. And, one and two mixture SFMs are evaluated.

Figure 2.6 illustrates results of speaker-dependent continuous phoneme recognitions. The One-Pass Viterbi algorithm [Bridle 82] was used for decoding. In the "Label training", each recognition rate of an SFM is trained by using label informations. On other hand, the "Embedded training" means recognition rates of SFMs that are trained by using embedded-extimation.

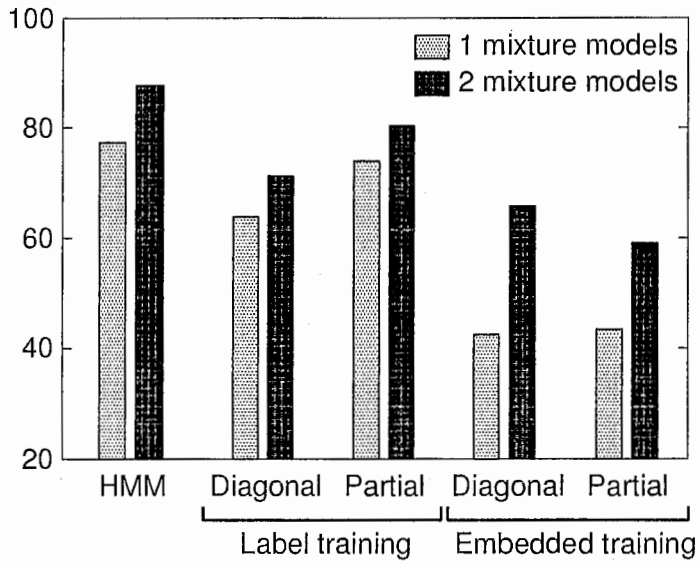

Figure 2.6: Experimental results in continuous phoneme recognitions.

The SFMs got lower recognition rates than the conventional HMMs. And, the hand-labeled SFMs got higher recognition rates than the embedded-estimated SFMs. It is considered that phoneme boundaries are not estimated rightly. Each variance value of a

duration control of a phoneme model is illustrated in figure 2.7. The variances got larger values than label informations.

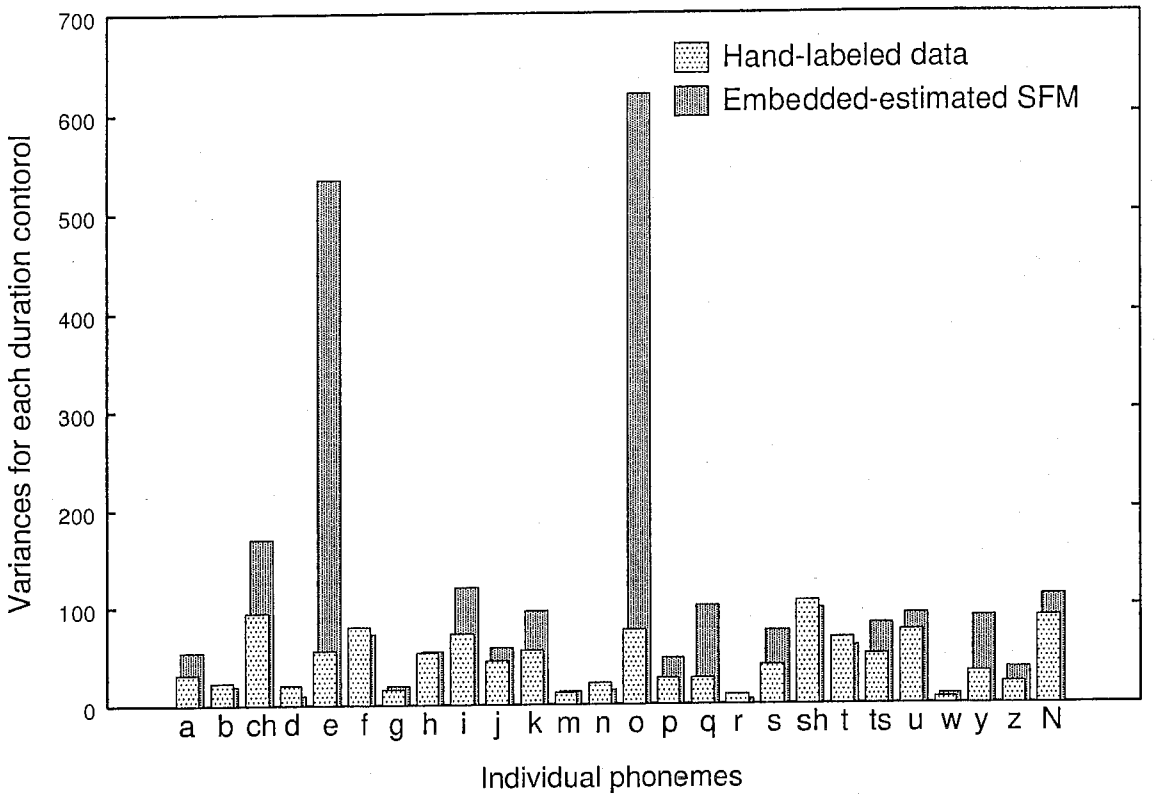

Figure 2.7: Each variance of a duration control.

Chapter 3

Segmental Feature Model with Variances in a Segment

In section 2, we proposed an SFM using a segmental feature vector based on average values. The SFMs got high recognition rates compared with conventional HMMs in phoneme classification experiments. However, phoneme boundaries were not estimated rightly in continuous phoneme recognitions.

In this section, we propose an SFM with variances in a segment, to improve an estimation accuracy of phoneme boundaries. In section 3.1, we study a cause which phoneme boundaries can not be estimated rightly. Then, we propose the SFM with variances in a segment as technique to solve the problem. The SFM is evaluated by speech recognition experiments in section 3.2.

3.1 Variances in a Segment

A segmental feature vector extracted from a feature vector sequence does not represent a trajectory of the feature vector sequence rightly. In other words, the segmental feature vector does not represent nice behavior of the feature vector sequence, since an average value of a sub-period is used. In figure 3.1, a behavior of the feature vector sequence differ to a trajectory of the segmental feature vector extracted from the feature vector sequence.

To reduce the mismatches, we propose an SFM with variances in a segment. The variances in a segment means a penalty for the mismatches between a feature vector sequence and a segmental feature vector. The variances can be calculated by equation 3.1. Let $sv = (y_{1,1}, y_{1,2}, \cdots, y_{D,N})$ be a segmental feature vector, the D is a number of dimensions of a feature vector and the N means a time resolution of the segmental feature vector. Then, let p_i be a frame numbers $\{b, b+1, b+2, \cdots, e\}$ of a feature vector sequence allocated to a state i. The μ_s and σ_s represents mean and variance values for segmental feature vector, and the σ_f means a variance in a segment. The γ is a weight for the penalty.

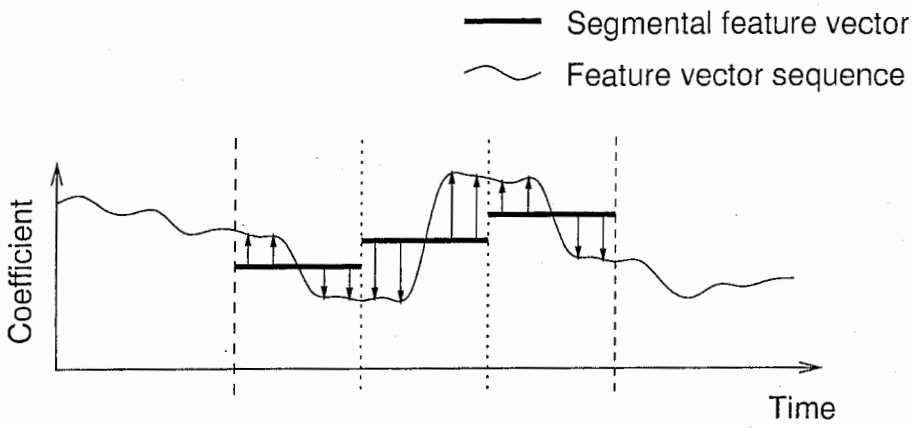

Figure 3.1: Mismatch among trajectories for a feature vector sequence and segmental feature vector

$$b(sv) = \mathcal{N}(sv, \mu_s, \sigma_s)^{b-e+1}(\prod_{d=1}^{D}\prod_{i=1}^{N}\prod_{t\in p_i}\mathcal{N}(y_{d,i}, \mu_s(d,i), \sigma_f(d,i)))^{\gamma} \tag{3.1}$$

By using the equation, we can calculate a penalty for the mismatch between a feature vector sequence and a segmental feature vector.

3.2 Speech Recognition Experiments

We performed experiments of a phoneme classification and a continuous phoneme recognition for the SFM with variances in a segment.

3.2.1 Phoneme Classification Experiments

Table 3.1 shows results of phoneme classification experiments. The same experimental conditions described in section 2.4.1 are used. $\gamma = 2.0$ was used from a preliminary experiment.

These SFM with variances in a segment got higher recognition rates than the conventional HMM. Moreover, these SFMs got higher recognition rates than an SFM without variances in a segment. It is considered that speech signals are modeled efficiently by using variances in a segment to an SFM.

3.2.2 Continuous Phoneme Recognition Experiments

For the SFM with variances in a segment, continuous phoneme recognition experiments were performed. Figure 3.2 shows the results. The SFM with variances in a segment got higher recognition rates than an SFM without the variances. However, the SFM with variances in a segment go lower recognition rates than the conventional HMM.

Table 3.1: Speaker-Dependent Phoneme Classification Experiments

Mixtures	Conventional HMM	Diagonal Covariance				Partial Covariance			
		FMIX		SMIX		FMIX		SMIX	
		ONE	TAU	ONE	TAU	ONE	TAU	ONE	TAU
1	83.5%	85.7%	83.2%	85.6%	83.1%	87.6%	85.7%	87.6%	85.7%
2	89.9%	90.3%	89.8%	90.0%	89.5%	91.3%	91.3%	91.8%	91.3%
4	92.5%	93.4%	93.8%	92.8%	93.7%	94.0%	93.2%	92.9%	93.0%
8	94.2%	95.5%	95.1%	95.6%	95.2%	95.1%	95.0%	95.1%	94.5%

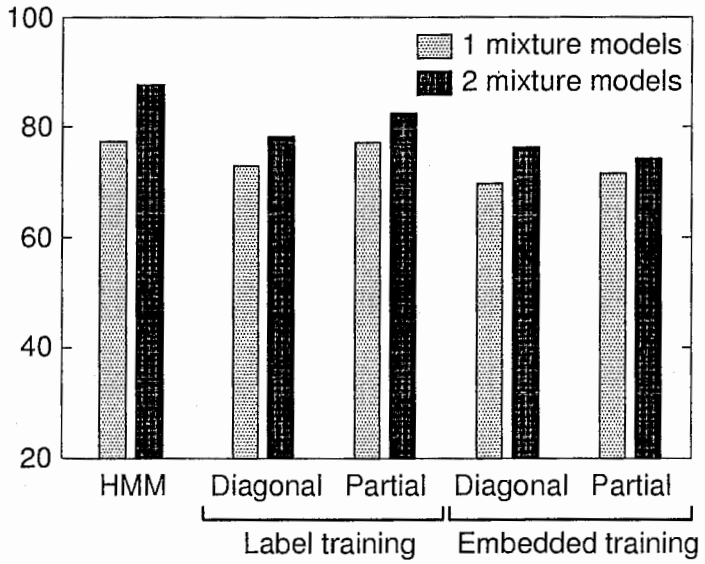

Figure 3.2: Results of continuous phoneme recognition.

Figure 3.3 shows variance values, each box represents a variance value of a duration control in the SFM. Like this figure, the SFM with variances in a segment got smaller variances than the SFM without the variances. However, the variance values are larger than the conventional HMM.

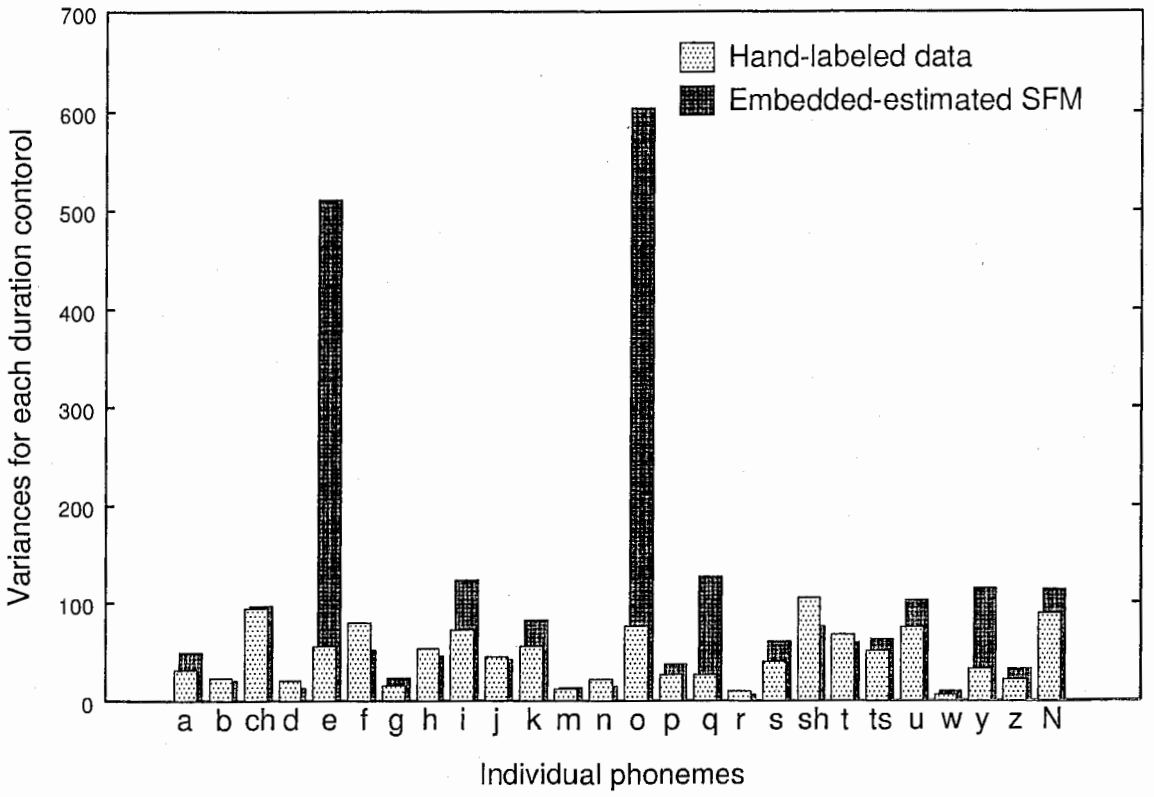

Figure 3.3: Variances for each duration control

Chapter 4

Conclusion

In this report, we proposed a segmental feature vector based on average values. To evaluate performance of a speech recognition, we performed experiments of a phoneme classification and a phoneme continuous recognition. In the phoneme classification experiments, the SFM got higher recognition rates than the conventional HMM. In the continuous phoneme, the SFM got lower recognition rates than the conventional HMM. It is considered that an estimation performance of phoneme boundaries is lower than the conventional HMM.

To solve the problem, we proposed an SFM with variances in a segment. The SFM got higher recognition rates than the SFM without the variances. However, the SFM with the variances got lower recognition rates than the conventional HMM also.

Future works include improving a estimation performance of phoneme boundaries and evaluating speaker-independent speech recognition experiments.

Bibliography

[Dempster 77] A.P. Dempster, et al, : "Maximum likelihood from incomplete date via the EM algorithm," J.R. Statistical Society, Series B, 39, pp.1-38 (1977)

[Ferguson 80] J.D. Ferguson : "Variable duration models for speech," in Proc., Symp. on Application of hidden Markov models to Text and Speech, ed. J.D. Ferguson, Princeton, pp.143-179 (1980)

[Bridle 82] J. Bridle, M. Brown, R. Chamberlain : "An algorithm for Connected Word Recognition," International Conference on Acoust., Speech, Signal Processing (I-CASSP), pp.899-902, 1982.

[Ostendorf 96] M. Ostendorf, V. Digalakis, and O. Kimball : "Form HMM's to segment models: A unified view of stochastic modeling for speech recognition," IEEE Trans. on Speech and Audio Proc., vol.4, no.5, pp.360-378, Sept. 1996.

[Nakagawa 96] S. Nakagawa and K. Yamamoto : "Speech Recognition by Hidden Markov Model Using Segmental Statistics," EIC(D-II), vol.J79-D-II, no.12, pp.2032-2038, Dec. 1996.