INVITED PAPER *Special Issue on the 2000 IEICE Excellent Paper Award*

Multi-Space Probability Distribution HMM[**]

Keiichi TOKUDA[†], **Takashi MASUKO**[††], **Noboru MIYAZAKI**[††*], ***and*** **Takao KOBAYASHI**[††], ***Regular Members***

SUMMARY This paper proposes a new kind of hidden Markov model (HMM) based on multi-space probability distribution, and derives a parameter estimation algorithm for the extended HMM. HMMs are widely used statistical models for characterizing sequences of speech spectra, and have been successfully applied to speech recognition systems. HMMs are categorized into discrete HMMs and continuous HMMs, which can model sequences of discrete symbols and continuous vectors, respectively. However, we cannot apply both the conventional discrete and continuous HMMs to observation sequences which consist of continuous values and discrete symbols: F0 pattern modeling of speech is a good illustration. The proposed HMM includes discrete HMM and continuous HMM as special cases, and furthermore, can model sequences which consist of observation vectors with variable dimensionality and discrete symbols.

key words: *hidden Markov model, text-to-speech, F0, multi-space probability distribution*

1. Introduction

The hidden Markov models (HMMs) are widely used statistical models, and have been successfully applied to modeling sequences of speech spectra in speech recognition systems. The performance of HMM-based speech recognition systems has been improved by techniques which utilize the flexibility of HMMs: context-dependent modeling [1], dynamic feature parameters [2], mixtures of Gaussian densities [3], tying techniques (e.g., [4]), and speaker/environment adaptation techniques (e.g., [5]).

HMMs are categorized into discrete and continuous HMMs, which can model sequences of discrete symbols and continuous vectors, respectively. However, we cannot apply both the conventional discrete and continuous HMMs to observations which consist of continuous values and discrete symbols. Modeling the fundumental frequency (F0) pattern of speech is a good illustration: we cannot directly apply both the conventional discrete and continuous HMMs to F0 pattern modeling since F0 values are not defined in the unvoiced region, i.e., the observation sequence of an F0 pattern is composed of one-dimensional continuous values and discrete symbols which represent "unvoiced." Several methods have been investigated for handling the unvoiced region: i) replacing each "unvoiced" symbol by a random vector generated from a probability density function (pdf) with a large variance and then modeling the random vectors explicitly in the continuous HMMs [6], ii) modeling the "unvoiced" symbols explicitly in the continuous HMMs by replacing each "unvoiced" symbol with 0 and adding an extra pdf for the "unvoiced" symbol to each mixture [7], iii) assuming that F0 values always exist but they cannot be observed in the unvoiced region and applying the EM algorithm [8]. Although approach iii) is appropriate from the viewpoint of statistical modeling, it intends to estimate F0 values which are not existent contradictorily. Approaches i) and ii) are based on heuristic assumptions. As a result, we cannot derive statistical techniques, e.g., context-dependent modeling, speaker/environment adaptation techniques, in a statistically correct manner.

This paper describes a new kind of HMM in which the state output probabilities are defined by multi-space probability distributions, and derives its reestimation formulas. Each space in the multi-space probability distribution has its weight and a continuous pdf whose dimensionality depends on the space. An observation consists of an n-dimensional continuous vector and a set of space indices which specify n-dimensional spaces. As a result, the extended HMM includes both the discrete and continuous mixture HMMs as special cases, and furthermore, can model the sequences of observation vectors with variable dimensionality including zero-dimensional observations, i.e., discrete symbols.

This paper is organized as follows. Multi-space probability distribution and multi-space probability distribution HMM (MSD-HMM) are defined in Sects. 2 and 3, respectively. A reestimation algorithm for MSD-HMMs is derived in Sect. 4. The relation between the conventional and the proposed HMMs, and the application of MSD-HMM to F0 pattern modeling, are discussed in Sect. 5. Concluding remarks and our plans for future work are presented in the final section.

Manuscript received November 6, 2001.

[†] The author is with the Department of Computer Science, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.

[††] The authors are with the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama-shi, 226–8502 Japan.

[*] Presently, with the NTT Communication Science Laboratories.

[**] This article was originally published in IEICE Trans. Inf. & Syst. (Japanese Edition) vol.J83-D-II, no.7, pp.1579–1589, July 2000.

2. Multi-Space Probability Distribution

We consider a sample space Ω shown in Fig. 1, which consists of G spaces:

$$\Omega = \bigcup_{g=1}^{G} \Omega_g, \tag{1}$$

where Ω_g is an n_g-dimensional real space R^{n_g}, specified by space index g. While each space has its own dimensionality, some of them may have the same dimensionality.

Each space Ω_g has its probability w_g, i.e., $P(\Omega_g) = w_g$, where $\sum_{g=1}^{G} w_g = 1$. If $n_g > 0$, each space has a pdf function $\mathcal{N}_g(\boldsymbol{x})$, $\boldsymbol{x} \in R^{n_g}$, where $\int \mathcal{N}_g(\boldsymbol{x}) d\boldsymbol{x} = 1$. We assume that Ω_g contains only one sample point if $n_g = 0$. Accordingly, we have $P(\Omega) = 1$.

Each event E, which will be considered in this paper, is represented by a random vector $\boldsymbol{o}$ which consists of a set of space indices X and a continuous random variable $\boldsymbol{x} \in R^n$, that is,

$$\boldsymbol{o} = (X, \boldsymbol{x}), \tag{2}$$

where all spaces specified by X are n-dimensional. On the other hand, X does not necessarily include all indices which specify n-dimensional spaces (see $\boldsymbol{o}_1$ and $\boldsymbol{o}_2$ in Fig. 1). It is noted that not only the observation vector $\boldsymbol{x}$ but also the space index set X is a random variable, which is determined by an observation device (or feature extractor) at each observation. The observation probability of $\boldsymbol{o}$ is defined by

$$b(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_g \mathcal{N}_g(V(\boldsymbol{o})), \tag{3}$$

where

$$S(\boldsymbol{o}) = X, \qquad V(\boldsymbol{o}) = \boldsymbol{x}. \tag{4}$$

It is noted that, although $\mathcal{N}_g(\boldsymbol{x})$ does not exist for $n_g = 0$ since Ω_g contains only one sample point, for simplicity of notation, we define $\mathcal{N}_g(\boldsymbol{x}) \equiv 1$ for $n_g = 0$.

Some examples of observations are shown in Fig. 1. An observation $\boldsymbol{o}_1$ consists of a set of space indices $X_1 = \{1, 2, G\}$ and a three-dimensional vector $\boldsymbol{x}_1 \in R^3$. Thus the random variable $\boldsymbol{x}$ is drawn from one of three spaces Ω_1, Ω_2, $\Omega_G \in R^3$, and its pdf is given by $w_1\mathcal{N}_1(\boldsymbol{x}) + w_2\mathcal{N}_2(\boldsymbol{x}) + w_G\mathcal{N}_G(\boldsymbol{x})$.

The probability distribution defined above, which will be referred to as *multi-space probability distribution* (MSD) in this paper, is the same as the discrete distribution when $n_g \equiv 0$. Furthermore, if $n_g \equiv m > 0$ and $S(\boldsymbol{o}) \equiv \{1, 2, \ldots, G\}$, the multi-space probability distribution is represented by a G-mixture pdf. Thus the multi-space probability distribution is more general than either discrete or continuous distributions.

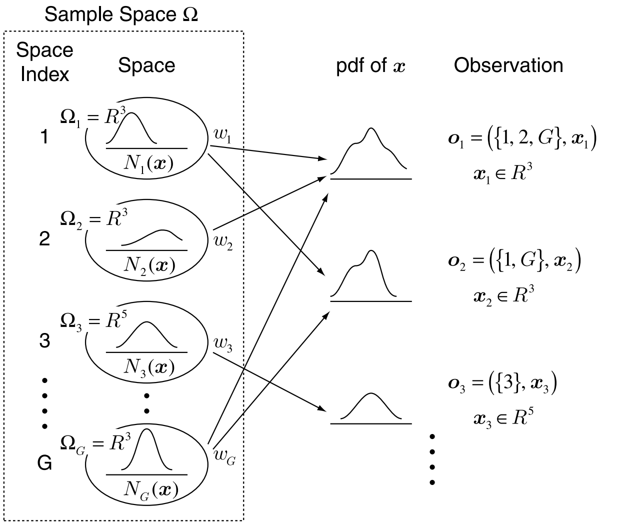

Sample Space Ω
Space Index
Space
pdf of x
Observation
1
$\Omega_1 = R^3$
$N_1(\boldsymbol{x})$
w_1
2
$\Omega_2 = R^3$
$N_2(\boldsymbol{x})$
w_2
3
$\Omega_3 = R^5$
$N_3(\boldsymbol{x})$
w_3
G
$\Omega_G = R^3$
$N_G(\boldsymbol{x})$
w_G
$\boldsymbol{o}_1 = (\{1, 2, G\}, \boldsymbol{x}_1)$
$\boldsymbol{x}_1 \in R^3$
$\boldsymbol{o}_2 = (\{1, G\}, \boldsymbol{x}_2)$
$\boldsymbol{x}_2 \in R^3$
$\boldsymbol{o}_3 = (\{3\}, \boldsymbol{x}_3)$
$\boldsymbol{x}_3 \in R^5$

Fig. 1 Multi-space probability distribution and observations.

The following example shows that the multi-space probability distribution conforms to statistical phenomena in the real world:

> A man is fishing in a pond. There are red fishes, blue fishes, and tortoises in the pond. In addition, some junk articles are in the pond. When he catches a fish, he is interested in the kind of the fish and its size, for example, the length and height. When he catches a tortoise, it is sufficient to measure the diameter if we assume that the tortoise has a circular shape. Furthermore, when he catches a junk article, he takes no interest in its size.

In this case, the sample space consists of four spaces:

Ω_1: two-dimensional space correponding to lengths and heights of red fishes.

Ω_2: two-dimensional space correponding to lengths and heights of blue fishes.

Ω_3: one-dimensional space correponding to diameters of tortoises.

Ω_4: zero-dimensional space correponding to junk articles.

The weights w_1, w_2, w_3, w_4 are determined by the ratio of red fishes, blue fishes, tortoises, and junk articles in the pond. Functions $\mathcal{N}_1(\cdot)$ and $\mathcal{N}_2(\cdot)$ are two-dimensional pdfs for sizes (lengths and heights) of red fishes and blue fishes, respectively. The function $\mathcal{N}_3(\cdot)$ is the one-dimensional pdf for diameters of tortoises. For example, when the man catches a red fish, the observation is given by $\boldsymbol{o} = (\{1\}, \boldsymbol{x})$, where $\boldsymbol{x}$ is a two-dimensional vector which represents the length and height of the red fish. Suppose that he is fishing day and night, and during the night, he cannot distinguish between the colors of fishes, while he can measure their lengths and heights. In this case, the observation of a fish at night is given by $\boldsymbol{o} = (\{1, 2\}, \boldsymbol{x})$.

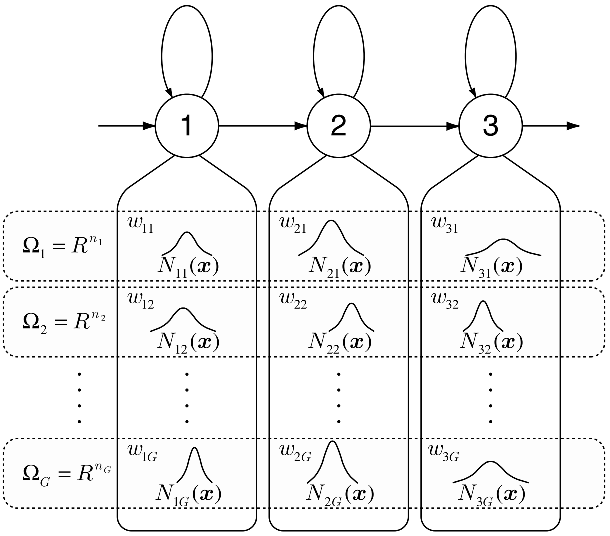


Fig. 2 An HMM based on multi-space probability distribution.

3. HMMs Based on Multi-Space Probability Distribution

By using the multi-space distribution, we define a new kind of HMM. In this paper, we call it multi-space probability distribution HMM (MSD-HMM). The output probability in each state of MSD-HMM is given by the multi-space probability distribution defined in the previous section. An N-state MSD-HMM λ is specified by the initial state probability distribution $\pi = \{\pi_j\}_{j=1}^{N}$, the state transition probability distribution $A = \{a_{ij}\}_{i,j=1}^{N}$, and the state output probability distribution $B = \{b_i(\cdot)\}_{i=1}^{N}$, where

$$b_i(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_{ig} \ \mathcal{N}_{ig}(V(\boldsymbol{o})). \tag{5}$$

As shown in Fig. 2, each state i has G pdfs $\mathcal{N}_{i1}(\cdot)$, $\mathcal{N}_{i2}(\cdot), \dots, \mathcal{N}_{iG}(\cdot)$, and their weights $w_{i1}, w_{i2}, \dots, w_{iG}$, where $\sum_{g=1}^{G} w_{ig} = 1$. The observation probability of $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \dots, \boldsymbol{o}_T\}$ can be written as

$$\begin{aligned}
&P(\boldsymbol{O}|\lambda)\\
&\quad = \sum_{\text{all } \boldsymbol{q}} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{o}_t)\\
&\quad = \sum_{\text{all } \boldsymbol{q}} \prod_{t=1}^{T} a_{q_{t-1}q_t} \sum_{g \in S(\boldsymbol{o}_t)} w_{q_t g} \mathcal{N}_{q_t g}(V(\boldsymbol{o}_t))\\
&\quad = \sum_{\text{all } \boldsymbol{q}} \left[\sum_{g \in S(\boldsymbol{o}_1)} a_{q_0 q_1} w_{q_1 g} \mathcal{N}_{q_1 g}(V(\boldsymbol{o}_1)) \right]\\
&\qquad \left[\sum_{g \in S(\boldsymbol{o}_2)} a_{q_1 q_2} w_{q_1 g} \mathcal{N}_{q_2 g}(V(\boldsymbol{o}_2)) \right]\\
&\qquad \cdots \left[\sum_{g \in S(\boldsymbol{o}_T)} a_{q_{T-1} q_T} w_{q_T g} \mathcal{N}_{q_T g}(V(\boldsymbol{o}_T)) \right]\\
&\quad = \sum_{\text{all } \boldsymbol{q}, \boldsymbol{l}} \prod_{t=1}^{T} a_{q_{t-1}q_t} w_{q_t l_t} \mathcal{N}_{q_t l_t}(V(\boldsymbol{o}_t)),
\end{aligned} \tag{6}$$

where $\boldsymbol{q} = \{q_1, q_2, \dots, q_T\}$ is a possible state sequence, $\boldsymbol{l} = \{l_1, l_2, \dots, l_T\} \in \{S(\boldsymbol{o}_1) \times S(\boldsymbol{o}_2) \times \dots \times S(\boldsymbol{o}_T)\}$ is a sequence of space indices which is possible for the observation sequence $\boldsymbol{O}$, and $a_{q_0 j}$ denotes π_j.

Equation (6) can be calculated efficiently through the forward and backward variables:

$$\alpha_t(i) = P(\boldsymbol{o}_1, \boldsymbol{o}_2, \dots, \boldsymbol{o}_t, q_t = i|\lambda) \tag{7}$$

$$\beta_t(i) = P(\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+2}, \dots, \boldsymbol{o}_T|q_t = i, \lambda), \tag{8}$$

which can be calculated with the forward-backward inductive procedure in a manner similar to conventional HMMs:

1. Initialization:

$$\begin{aligned} &\alpha_1(i) = \pi_i b_i(\boldsymbol{o}_1), \quad 1 \le i \le N\\ &\beta_T(i) = 1, \qquad\qquad\ 1 \le i \le N \end{aligned} \tag{9}$$

2. Recursion:

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^{N} \alpha_t(j) a_{ji} \right] b_i(\boldsymbol{o}_{t+1}),$$

$$1 \le i \le N, \quad t = 1, 2, \dots, T-1 \tag{10}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j),$$

$$1 \le i \le N, \quad t = T-1, 2, \dots, 1. \tag{11}$$

According to the definitions, (6) can be calculated as

$$P(\boldsymbol{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} a_{q_0 i} b_i(\boldsymbol{o}_1) \beta_1(i). \tag{12}$$

The forward and backward variables are also used for calculating the reestimation formulas derived in the next section (i.e., calculation of Eqs. (15) and (16)).

4. Reestimation Algorithm

For a given observation sequence $\boldsymbol{O}$ and a particular choice of MSD-HMM, the objective in maximum likelihood estimation is to maximize the observation likelihood $P(\boldsymbol{O}|\lambda)$ given by Eq. (6), over all parameters in λ. In a manner similar to those reported in [9] and [3], we derive reestimation formulas for the maximum likelihood estimation of MSD-HMM.

4.1 Q-Function

An auxiliary function $Q(\lambda', \lambda)$ of current parameters λ' and new parameters λ is defined as follows:

$$Q(\lambda', \lambda) = \sum_{\text{all } \boldsymbol{q}, \boldsymbol{l}} P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda') \log P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda). \quad (13)$$

In the following, we assume $\mathcal{N}_{ig}(\cdot)$ to be the Gaussian density with mean vector $\boldsymbol{\mu}_{ig}$ and covariance matrix $\boldsymbol{\Sigma}_{ig}$. However, extension to elliptically symmetric densities which satisfy the consistency conditions of Kolmogorov is straightforward. We present the following three theorems:

Theorem 1:

$$Q(\lambda', \lambda) \geq Q(\lambda', \lambda') \rightarrow P(\boldsymbol{O}, \lambda) \geq P(\boldsymbol{O}, \lambda') \quad (14)$$

Theorem 2: If, for each space Ω_g, there are among $V(\boldsymbol{o}_1)$, $V(\boldsymbol{o}_2)$, $\ldots$, $V(\boldsymbol{o}_T)$, $n_g + 1$ observations $g \in S(o_t)$, any n_g of which are linearly independent, $Q(\lambda', \lambda)$ has a unique global maximum as a function of λ, and this maximum is the one and only critical point.

Theorem 3: A parameter set λ is a critical point of the likelihood $P(\boldsymbol{O}|\lambda)$ if and only if it is a critical point of the Q-function.

Theorems 1 and 3 can be proved in a similar manner to the conventional HMM. We have to newly prove Theorem 2, which confirms that the Q-function has a unique global maximum as a function of λ because the proposed HMM has a different state output probability distribution from the conventional descrete or continuous HMMs. The proof of Theorem 2 is given in Appendix.

We define the parameter reestimates to be those which maximize $Q(\lambda', \lambda)$ as a function of λ, λ' being the latest estimates. Because of the above theorems, the sequence of reestimates obtained in this way produces a monotonic increase in the likelihood unless λ is a critical point of the likelihood.

4.2 Maximization of Q-Function

For given observation sequence $\boldsymbol{O}$ and model λ', we derive parameters of λ which maximize $Q(\lambda', \lambda)$.

The posterior probability of being in state i at time t, given the observation sequence $\boldsymbol{O}$ and model λ, is given by

$$\begin{aligned}
&\gamma_t(i, h) \\
&\quad = P(q_t = i, l_t = h | \boldsymbol{O}, \lambda) \\
&\quad = P(q_t = i | \boldsymbol{O}, \lambda) P(l_t = h | q_t = i, \boldsymbol{O}, \lambda) \\
&\quad = \frac{P(q_t = i, \boldsymbol{O} | \lambda)}{P(\boldsymbol{O}|\lambda)} P(l_t = h | q_t = i, \boldsymbol{O}, \lambda) \\
&\quad = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \cdot \frac{w_{ih}\mathcal{N}_{ih}(V(\boldsymbol{o}_t))}{\sum_{g \in S(\boldsymbol{o}_t)} w_{ig}\mathcal{N}_{ig}(V(\boldsymbol{o}_t))}.
\end{aligned} \quad (15)$$

Similarly, the posterior probability of transitions from state i to state j at time $t+1$ is given by

$$\begin{aligned}
\xi_t(i, j) &= P(q_t = i, q_{t+1} = j | \boldsymbol{O}, \lambda) \\
&= \frac{P(q_t = i, q_{t+1} = j, \boldsymbol{O} | \lambda)}{P(\boldsymbol{O}|\lambda)} \\
&= \frac{\alpha_t(i) a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j)}{\sum_{m=1}^{N} \sum_{k=1}^{N} \alpha_t(m) a_{mk} b_k(\boldsymbol{o}_{t+1}) \beta_{t+1}(k)}.
\end{aligned} \quad (16)$$

We define a function $T(\boldsymbol{O}, g)$ which returns a set of time t at which the space index set $S(\boldsymbol{o}_t)$ includes space index g:

$$T(\boldsymbol{O}, g) = \{t | g \in S(\boldsymbol{o}_t)\}. \quad (17)$$

By introducing this fucntion, the following manipulations of the equations can be carried out in a similar manner to the conventional continuous mixture HMMs.

From Eq. (6), $\log P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda)$ can be written as

$$\begin{aligned}
&\log P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda) \\
&\quad = \sum_{t=1}^{T} \left(\log a_{q_{t-1}q_t} + \log w_{q_t l_t} + \log \mathcal{N}_{q_t l_t}(V(\boldsymbol{o}_t)) \right). \quad (18)
\end{aligned}$$

Hence, Q-function Eq. (13) can be written as

$$\begin{aligned}
&Q(\lambda', \lambda) \\
&\quad = \sum_{\text{all } \boldsymbol{q}, \boldsymbol{l}} P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda') \log a_{q_0 q_1} \\
&\qquad + \sum_{\text{all } \boldsymbol{q}, \boldsymbol{l}} P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda') \sum_{t=1}^{T-1} \log a_{q_t q_{t+1}} \\
&\qquad + \sum_{\text{all } \boldsymbol{q}, \boldsymbol{l}} P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda') \sum_{t=1}^{T} \log w_{q_t l_t} \\
&\qquad + \sum_{\text{all } \boldsymbol{q}, \boldsymbol{l}} P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda') \sum_{t=1}^{T} \log \mathcal{N}_{q_t l_t}(V(\boldsymbol{o}_t)).
\end{aligned} \quad (19)$$

The first term of Eq. (19), which is related to $a_{q_0 q_1}$, i.e., π_{q_1}, is given by

$$
\begin{aligned}
&\sum_{\text{all } \boldsymbol{q},\boldsymbol{l}} P(\boldsymbol{O},\boldsymbol{q},\boldsymbol{l}|\lambda') \log a_{q_0 q_1} \\
&\qquad = \sum_{i=1}^{N} \sum_{\text{all } \boldsymbol{l}} P(\boldsymbol{O}, q_1 = i, \boldsymbol{l}|\lambda') \log a_{q_0 i} \\
&\qquad = \sum_{i=1}^{N} P(\boldsymbol{O}, q_1 = i|\lambda') \log \pi_i. \qquad (20)
\end{aligned}
$$

The second term of Eq. (19), which is related to a_{ij}, is given by

$$
\begin{aligned}
&\sum_{\text{all } \boldsymbol{q},\boldsymbol{l}} P(\boldsymbol{O},\boldsymbol{q},\boldsymbol{l}|\lambda') \sum_{t=1}^{T-1} \log a_{q_t q_{t+1}} \\
&\quad = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} \sum_{\text{all } \boldsymbol{l}} P(\boldsymbol{O}, q_t = i, q_{t+1} = j, \boldsymbol{l}|\lambda') \log a_{ij} \\
&\quad = \sum_{i,j=1}^{N} \sum_{t=1}^{T-1} P(\boldsymbol{O}, q_t = i, q_{t+1} = j|\lambda') \log a_{ij}. \qquad (21)
\end{aligned}
$$

The third term of Eq. (19), which is related to w_{ig}, is given by

$$
\begin{aligned}
&\sum_{\text{all } \boldsymbol{q},\boldsymbol{l}} P(\boldsymbol{O},\boldsymbol{q},\boldsymbol{l}|\lambda') \sum_{t=1}^{T} \log w_{q_t l_t} \\
&\quad = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{g \in S(\boldsymbol{o}_t)} P(\boldsymbol{O}, q_t = i, l_t = g|\lambda') \log w_{ig} \\
&\quad = \sum_{i=1}^{N} \sum_{g=1}^{G} \sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g|\lambda') \log w_{ig}. \qquad (22)
\end{aligned}
$$

The fourth term of Eq. (19), which is related to $\mathcal{N}_{ig}(\cdot)$, is given by

$$
\begin{aligned}
&\sum_{\text{all } \boldsymbol{q},\boldsymbol{l}} P(\boldsymbol{O},\boldsymbol{q},\boldsymbol{l}|\lambda') \sum_{t=1}^{T} \log \mathcal{N}_{q_t l_t}(V(\boldsymbol{o}_t)) \\
&\quad = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{g \in S(\boldsymbol{o}_t)} P(\boldsymbol{O}, q_t = i, l_t = g|\lambda') \\
&\quad\quad \cdot \log \mathcal{N}_{ig}(V(\boldsymbol{o}_t)) \\
&\quad = \sum_{i=1}^{N} \sum_{g=1}^{G} \sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g|\lambda') \\
&\quad\quad \cdot \log \mathcal{N}_{ig}(V(\boldsymbol{o}_t)). \qquad (23)
\end{aligned}
$$

Equations (20)–(22) have the form of $\sum_{i=1}^{N} u_i \log y_i$, which attains its unique maximum point

$$
y_i = \frac{u_i}{\sum_{j=1}^{N} u_j} \qquad (24)
$$

under the constraint $\sum_{i=1}^{N} y_i = 1$, $y_i \geq 0$. Therefore, the parameters π_i, a_{ij}, and w_{ig} which maximize Eq. (20), subject to the stochastic constraints $\sum_{i=1}^{N} \pi_i = 1$, $\sum_{j=1}^{N} a_{ij} = 1$, and $\sum_{g=1}^{G} w_g = 1$, respectively, can be derived as

$$
\begin{aligned}
\pi_i &= \frac{P(\boldsymbol{O}, q_1 = i|\lambda')}{\sum_{j=1}^{N} P(\boldsymbol{O}, q_1 = j|\lambda')} = \frac{P(\boldsymbol{O}, q_1 = i|\lambda')}{P(\boldsymbol{O}|\lambda')} \\
&= P(q_1 = i|\boldsymbol{O}, \lambda') \\
&= \sum_{g \in S(\boldsymbol{o}_1)} \gamma'_1(i, g) \qquad (25)
\end{aligned}
$$

$$
\begin{aligned}
a_{ij} &= \frac{\sum_{t=1}^{T-1} P(\boldsymbol{O}, q_t = i, q_{t+1} = j|\lambda')}{\sum_{k=1}^{N} \sum_{t=1}^{T-1} P(\boldsymbol{O}, q_t = i, q_{t+1} = k|\lambda')} \\
&= \frac{\sum_{t=1}^{T-1} P(\boldsymbol{O}, q_t = i, q_{t+1} = j|\lambda')}{\sum_{t=1}^{T-1} P(\boldsymbol{O}, q_t = i|\lambda')} \\
&= \frac{\sum_{t=1}^{T-1} P(q_t = i, q_{t+1} = j|\boldsymbol{O}, \lambda')}{\sum_{t=1}^{T-1} P(q_t = i|\boldsymbol{O}, \lambda')} \\
&= \frac{\sum_{t=1}^{T-1} \xi'_t(i, j)}{\sum_{t=1}^{T-1} \sum_{g \in S(\boldsymbol{o}_t)} \gamma'_t(i, g)} \qquad (26)
\end{aligned}
$$

$$
\begin{aligned}
w_{ig} &= \frac{\sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g|\lambda')}{\sum_{h=1}^{G} \sum_{t \in T(\boldsymbol{O},h)} P(\boldsymbol{O}, q_t = i, l_t = h|\lambda')} \\
&= \frac{\sum_{t \in T(\boldsymbol{O},g)} \gamma'_t(i, g)}{\sum_{h=1}^{G} \sum_{t \in T(\boldsymbol{O},h)} \gamma'_t(i, h)}. \qquad (27)
\end{aligned}
$$

When $\mathcal{N}_{ig}(\cdot)$, $n_g > 0$ is the n_g-dimensional Gaussian density function with mean vector $\boldsymbol{\mu}_{ig}$ and covariance matrix $\boldsymbol{\Sigma}_{ig}$, Eq. (23) is maximized by setting the partial derivatives with respect to $\boldsymbol{\mu}_{ig}$ and ${\boldsymbol{\Sigma}_{ig}}^{-1}$:

$$\frac{\partial}{\partial \boldsymbol{\mu}_{ig}} \sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda') \cdot \log \mathcal{N}_{ig}(V(\boldsymbol{o}_t)) = \boldsymbol{0} \tag{28}$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{ig}^{-1}} \sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda') \cdot \log \mathcal{N}_{ig}(V(\boldsymbol{o}_t)) = \boldsymbol{0}. \tag{29}$$

From

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}_{ig}} \sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda') \log \mathcal{N}_{ig}(V(\boldsymbol{o}_t)) \\
&= \sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda') \cdot \frac{\partial}{\partial \boldsymbol{\mu}_{ig}} \log \mathcal{N}_{ig}(V(\boldsymbol{o}_t)) \\
&= \sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda') \cdot \boldsymbol{\Sigma}_{ig}^{-1} (V(\boldsymbol{o}_t) - \boldsymbol{\mu}_{ig}) \\
&= \boldsymbol{0}
\end{aligned} \tag{30}$$

and

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_{ig}^{-1}} \sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda') \log \mathcal{N}_{ig}(V(\boldsymbol{o}_t)) \\
&= \sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda') \cdot \frac{\partial}{\partial \boldsymbol{\Sigma}_{ig}^{-1}} \log \mathcal{N}_{ig}(V(\boldsymbol{o}_t)) \\
&= \frac{1}{2} \sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda') \cdot \left(\boldsymbol{\Sigma}_{ig} - (V(\boldsymbol{o}_t) - \boldsymbol{\mu}_{ig})(V(\boldsymbol{o}_t) - \boldsymbol{\mu}_{ig})^T\right) \\
&= \boldsymbol{0},
\end{aligned} \tag{31}$$

$\boldsymbol{\mu}_{ig}$ and $\boldsymbol{\Sigma}_{ig}$ are given by

$$\begin{aligned}
\boldsymbol{\mu}_{ig} &= \frac{\displaystyle\sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda') V(\boldsymbol{o}_t)}{\displaystyle\sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda')} \\
&= \frac{\displaystyle\sum_{t \in T(\boldsymbol{O},g)} \gamma'_t(i,g) V(\boldsymbol{o}_t)}{\displaystyle\sum_{t \in T(\boldsymbol{O},g)} \gamma'_t(i,g)}, \quad n_g > 0
\end{aligned} \tag{32}$$

and

$$\begin{aligned}
\boldsymbol{\Sigma}_{ig} &= \frac{\displaystyle\left(\sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda') (V(\boldsymbol{o}_t) - \boldsymbol{\mu}_{ig})(V(\boldsymbol{o}_t) - \boldsymbol{\mu}_{ig})^T\right)}{\displaystyle\sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda')} \\
&= \frac{\displaystyle\left(\sum_{t \in T(\boldsymbol{O},g)} \gamma'_t(i,g) (V(\boldsymbol{o}_t) - \boldsymbol{\mu}_{ig})(V(\boldsymbol{o}_t) - \boldsymbol{\mu}_{ig})^T\right)}{\displaystyle\sum_{t \in T(\boldsymbol{O},g)} \gamma'_t(i,g)}, \\
&\qquad n_g > 0,
\end{aligned} \tag{33}$$

respectively. From the condition mentioned in Theorem 2, it can be shown that each $\boldsymbol{\Sigma}_{ig}$ is positive definite.

From Sect. 4.1, by iterating the following procedure: 1) calculating λ which maximizes $\mathcal{Q}(\lambda', \lambda)$ by Eqs. (25)–(27), (32), and (33), and 2) substituting the obtained λ for λ', we can obtain a critical point of $P(\boldsymbol{O}|\lambda)$.

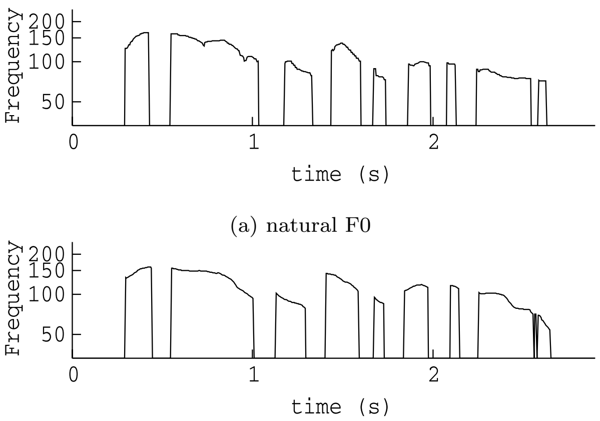
5. Discussion

5.1 Relation to Discrete Distribution HMM and Continuous Distribution HMM

The MSD-HMM includes the discrete HMM and the continuous mixture HMM as special cases since the multi-space probability distribution includes the discrete distribution and the continuous distribution. If $n_g \equiv 0$, the MSD-HMM is the same as the discrete HMM. In the case where $S(\boldsymbol{o}_t)$ specifies one space, i.e., $|S(\boldsymbol{o}_t)| \equiv 1$, the MSD-HMM is exactly the same as the conventional discrete HMM. If $|S(\boldsymbol{o}_t)| \geq 1$, the MSD-HMM is the same as the discrete HMM based on the multi-labeling VQ [10]. If $n_g \equiv m > 0$ and $S(\boldsymbol{o}) \equiv \{1, 2, \ldots, G\}$, the MSD-HMM is the same as the continuous G-mixture HMM. These can also be confirmed by the fact that if $n_g \equiv 0$ and $|S(\boldsymbol{o}_t)| \equiv 1$, the reestimation formulas Eqs. (25)–(27) are the same as those for discrete HMM of codebook size G, and if $n_g \equiv m$ and $S(\boldsymbol{o}_t) \equiv \{1, 2, \ldots, G\}$, the reestimation formulas Eqs. (25)–(33) are the same as those for continuous HMM with m-dimensional G-mixture densities. Accordingly, MSD-HMM includes the discrete and continuous mixture HMMs as special cases, and furthermore, can model the sequence of observation vectors with variable dimensionality including zero-dimensional observations, i.e., discrete symbols.

In addition, multi-channel HMMs [11] are also related to MSD-HMMs. Multi-channel HMMs have a special structure similar to MSD-HMMs. However, they assume that each channel always observes a discrete symbol, and they cannot be applied to the observation sequence composed of continuous vectors with variable dimensionality including zero-dimensional observations, i.e., discrete symbols. On the other hand, MSD-HMM includes the multi-channel HMM which was finally derived in [11] as a special case under the following conditions:

- The sample space consists of zero-dimensional spaces, each of which has a one-to-one correspondence with each symbol used in the multi-channel HMM.
- The observation consists of M space indices, each of which has a one-to-one correspondence with a channel and is drawn from symbols used in the channel.



(a) natural F0

(b) generated F0

Fig. 3 Generated F0 pattern for the sentence "heikiNbairitsuwo sageta keisekiga aru."

5.2 Application to F0 Pattern Modeling

While the observation of F0 has a continuous value in the voiced region, there exist no values for the unvoiced region. We can model this kind of observation sequence assuming that the observed F0 value occurs from one-dimensional spaces and the "unvoiced" symbol occurs from the zero-dimensional space defined in Sect. 2, that is, by setting $n_g = 1$ $(g = 1, 2, \ldots, G-1)$, $n_G = 0$ and

$$S(\boldsymbol{o}_t) = \begin{cases} \{1, 2, \ldots, G-1\}, & \text{(voiced)} \\ \{G\}, & \text{(unvoiced)} \end{cases} \tag{34}$$

the MSD-HMM can cope with F0 patterns including the unvoiced region without heuristic assumptions. In this case, the observed F0 value is assumed to be drawn from a continuous $(G-1)$-mixture pdf.

Experiments reported in [12] have shown that the likelihood of the MSD-HMM for the training data increases monotonically by calculating the reestimation formulas iteratively. From the trained MSD-HMMs, we can generate F0 patterns which approximate those of natural speech by using an algorithm (described in [13] as the "case 1" algorithm) for speech parameter generation from HMMs with dynamic features . An example is shown in Fig. 3, without an explanation of the experimental conditions because of limitations of space[†].

Real world phenomena of time sequences are not necessarily observed as a sequence of discrete symbols or continuous vectors. Accordingly, the proposed HMMs can be applied to not only F0 pattern modeling but also the modeling of various kinds of time sequences which consist of continuous vectors with variable dimensionality including zero-dimensional vectors, i.e., discrete symbols. As a result, MSD-HMM expected to be useful in such research areas as the prediction of human actions and economic forecasting.

Algorithms based on a statistical framework, which have been developed for the conventional HMMs, can be applied or extended to MSD-HMMs since the proposed scheme does not utilize any heuristic assumption or approximation. Actually, in the example shown in Fig. 3, we derived a model clustering scheme based on the minimum description length (MDL) principle, and used it in the model training.

6. Conclusion

A multi-space probability distribution HMM has been proposed and its reestimation formulas are derived. The MSD-HMM includes the discrete HMM and the continuous mixture HMM as special cases, and furthermore, can cope with the sequence of observation vectors with variable dimensionality including zero-dimensional observations, i.e., discrete symbols. As a result, MSD-HMMs can model F0 patterns without heuristic assumptions.

In the near future, we will present a speech synthesis system in which sequences of speech spectra [15], F0 patterns [12] and state durations [16] are modeled by MSD-HMM in a unified framework [14]. Fundamental frequency (F0) pattern modeling based on MSD-HMM may also be useful for enhancing the performance of speech recognition systems.

Acknowledgment

This work was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research (B), 10555125, and the Research Foundation for the Electrotechnology of Chubu, Grant for Research. The authors are grateful to Mr. Takayoshi Yoshimura, PhD student, Nagoya Institute of Technology, for carrying out experiments.

[†] Detailed experimental conditions can be found in [14].

References

[1] S. Schwartz, Y-L. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, “Context-dependent modeling for acoustic-phonetic of continuous speech,” Proc. ICASSP85, pp.1205–1208, March 1985.

[2] S. Furui, “Speaker independent isolated word rcognition using dynamic features of speech spectrum,” IEEE Trans. Acoust., Speech & Signal Process., vol.34, no.1, pp.52–59, Feb. 1986.

[3] B.-H. Juang, “Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains,” AT&T Technical Journal, vol.64, no.6, pp.1235–1249, 1985.

[4] J.J. Odell, “The use of context in large vocabulary speech recognition,” PhD Thesis, Cambridge University, March 1995.

[5] C.-H. Lee, C.H. Lin, and B.H. Juang, “A study on speaker adaptation of the parameters of continuous density hidden Markov models,” IEEE Trans. Acoust., Speech & Signal Process., vol.39, no.4, pp.806–814, April 1992.

[6] G.J. Freij and F. Fallside, “Lexical stress recognition using hidden Markov models,” Proc. ICASSP88, pp.135–138, April 1988.

[7] U. Jensen, R.K. Moore, P. Dalsgaard, and B. Lindberg, “Modeling intonation contours at the phrase level using continuous density hidden Markov models,” Computer Speech and Language, vol.8, no.3, pp.247–260, July 1994.

[8] K. Ross and M. Ostendorf, “A dynamical system model for generating F_0 for synthesis,” Proc. ESCA/IEEE Workshop on Speech Synthesis, pp.131–134, Sept. 1994.

[9] L.A. Liporace, “Maximum likelihood estimation for multivariate observations of Markov sources,” IEEE Trans. Inf. Theory, vol.28, no.5, pp.729–734, Sept. 1982.

[10] M. Nishimura and K. Toshioka, “HMM-based speech recognition using multi-dimensional multi-labeling,” Proc. ICASSP87, pp.1163–1166, May 1987.

[11] D. Xu, C. Fancourt, and C.Wang, “Multi-channel HMM,” Proc. ICASSP96, pp.841–844, May 1996.

[12] N. Miyazaki, K. Tokuda, T. Masuko, and T. Kobayashi, “A study on pitch pattern generation using HMMs based on multi-space probability distributions,” IEICE Technical Report, SP98-12, 1998.

[13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” ICASSP2000, pp.1315–1318, June 2000.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” Proc. EUROSPEECH99, pp.2347–2350, Sept. 1999.

[15] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, “Speech synthesis from HMMs using dynamic features,” Proc. ICASSP96, pp.389–392, May 1996.

[16] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Duration modeling for HMM-based speech synthesis,” Proc. ICSLP98, pp.29–32, Nov.–Dec. 1998.

Appendix: Proof of Theorem 2

The proof consists of the following three parts:

a) The second derivative of the Q-function along any direction in the parameter space is strictly negative at a critical point. This implies that any critical point is a relative maximum and that if there are more than one they are isolated.

b) $\mathcal{Q}(\lambda', \lambda) \to -\infty$ as λ approaches the finite boundary of the parameter space or the point at ∞. This property implies that the global maximum is a critical point.

c) The critical point is unique.

Proof (a)

From Sect. 4.2, the Q-function can be written as

$$\begin{aligned}
&\mathcal{Q}(\lambda', \lambda)\\
&= \sum_{i=1}^{N} P(\boldsymbol{O}, q_1 = i|\lambda') \log \pi_i\\
&\quad + \sum_{i,j=1}^{N} \sum_{t=1}^{T-1} P(\boldsymbol{O}, q_t = i, q_{t+1} = j|\lambda') \log a_{ij}\\
&\quad + \sum_{i=1}^{N} \sum_{g=1}^{G} \sum_{t \in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g|\lambda')\\
&\quad \cdot \left(\log w_{ig} - \frac{n_g}{2} \log(2\pi) + \frac{1}{2} \log |\boldsymbol{C}_{ig}| \right.\\
&\quad \left. - \frac{1}{2} (V(\boldsymbol{o}_t) - \boldsymbol{\mu}_{ig})^T \boldsymbol{C}_{ig} (V(\boldsymbol{o}_t) - \boldsymbol{\mu}_{ig}) \right),
\end{aligned} \tag{A$\cdot$1}$$

where $\boldsymbol{C}_{ig} = \boldsymbol{\Sigma}_{ig}^{-1}$. From the condition on observations $\boldsymbol{o}_t$, described in Theorem 2, $\boldsymbol{\Sigma}_{ig}$ and $\boldsymbol{\Sigma}_{ig}^{-1}$ are positive definite if $\boldsymbol{\Sigma}_{ig}$ is calculated by Eq. (33).

Let us express λ as a linear combination of two arbitrary points: $\lambda = \theta\lambda^{(1)} + (1-\theta)\lambda^{(2)}$, where $0 \le \theta \le 1$. That is,

$$\pi_i = \theta\pi_i^{(1)} + (1-\theta)\pi_i^{(2)} \tag{A$\cdot$2}$$

$$a_{ij} = \theta a_{ij}^{(1)} + (1-\theta) a_{ij}^{(2)} \tag{A$\cdot$3}$$

$$w_{ig} = \theta w_{ig}^{(1)} + (1-\theta) w_{ig}^{(2)} \tag{A$\cdot$4}$$

$$\boldsymbol{C}_{ig} = \theta \boldsymbol{C}_{ig}^{(1)} + (1-\theta) \boldsymbol{C}_{ig}^{(2)} \tag{A$\cdot$5}$$

$$\boldsymbol{\mu}_{ig} = \theta \boldsymbol{\mu}_{ig}^{(1)} + (1-\theta) \boldsymbol{\mu}_{ig}^{(2)}. \tag{A$\cdot$6}$$

Substituting these equations for Eq. (A$\cdot$1) and taking the second derivative with respect to θ, we obtain

$$\begin{aligned}
&\frac{\partial^2 \mathcal{Q}}{\partial \theta^2}\\
&= \sum_{i=1}^{N} P(\boldsymbol{O}, q_1 = i|\lambda') \frac{-(\pi_i^{(1)} - \pi_i^{(2)})^2}{(\theta\pi_i^{(1)} + (1-\theta)\pi_i^{(2)})^2}\\
&\quad + \sum_{i,j=1}^{N} P(\boldsymbol{O}, q_t = i, q_{t+1} = j|\lambda')
\end{aligned}$$

$$
\begin{aligned}
&\cdot\frac{-(a_{ij}^{(1)}-a_{ij}^{(2)})^2}{(\theta a_{ij}^{(1)}+(1-\theta)a_{ij}^{(2)})^2}\\
&+\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t\in T(\boldsymbol{O},g)} P(\boldsymbol{O},q_t=i,l_t=g|\lambda')\\
&\cdot\Bigg(\frac{-(w_{ig}^{(1)}-w_{ig}^{(2)})^2}{(\theta w_{ig}^{(1)}+(1-\theta)w_{ig}^{(2)})^2}\\
&+\frac{1}{2}\sum_{k=1}^{n_g}\frac{-(x_{igk}^{(1)}-x_{igk}^{(2)})^2}{(\theta x_{igk}^{(1)}+(1-\theta)x_{igk}^{(2)})^2}\\
&-(\boldsymbol{\mu}_{ig}^{(1)}-\boldsymbol{\mu}_{ig}^{(2)})^T\\
&\cdot\left(\theta\boldsymbol{C}_{ig}^{(1)}+(1-\theta)\boldsymbol{C}_{ig}^{(2)}\right)(\boldsymbol{\mu}_{ig}^{(1)}-\boldsymbol{\mu}_{ig}^{(2)})\\
&+2(\boldsymbol{\mu}_{ig}^{(1)}-\boldsymbol{\mu}_{ig}^{(2)})^T(\boldsymbol{C}_{ig}^{(1)}-\boldsymbol{C}_{ig}^{(2)})\\
&\cdot[V(\boldsymbol{o}_t)-(\theta\boldsymbol{\mu}_{ig}^{(1)}+(1-\theta)\boldsymbol{\mu}_{ig}^{(2)})]\Bigg), \qquad \text{(A$\cdot$ 7)}
\end{aligned}
$$

where $x_{igk}^{(1)}$ and $x_{igk}^{(2)}$ satisfy $x_{igk}=\theta x_{igk}^{(1)}+(1-\theta)x_{igk}^{(2)}$ for x_{igk} which are the diagonal entries of $\boldsymbol{U}_{ig}\boldsymbol{C}_{ig}\boldsymbol{U}_{ig}^{-1}$, and the orthogonal matrix $\boldsymbol{U}_{ig}$ diagonalizes $\boldsymbol{C}_{ig}$.

At a critical point, from the relation

$$
\begin{aligned}
&\left.\frac{\partial\mathcal{Q}}{\partial\boldsymbol{\mu}_{ig}}\right|_{\boldsymbol{\mu}_{ig}=\theta\boldsymbol{\mu}_{ig}^{(1)}+(1-\theta)\boldsymbol{\mu}_{ig}^{(2)}}\\
&=(\theta\boldsymbol{C}_{ig}^{(1)}+(1-\theta)\boldsymbol{C}_{ig}^{(2)})\\
&\quad\cdot\Bigg(\sum_{t\in T(\boldsymbol{O},g)}P(\boldsymbol{O},q_t=i,l_t=g|\lambda')\\
&\quad\cdot\left(V(\boldsymbol{o}_t)-(\theta\boldsymbol{\mu}_{ig}^{(1)}+(1-\theta)\boldsymbol{\mu}_{ig}^{(2)})\right)\Bigg)\\
&=\boldsymbol{0}, \qquad \text{(A$\cdot$ 8)}
\end{aligned}
$$

the sum involving the term bracketed by [ ] in Eq. (A· 7) vanishes. All of the remaining terms have negative values. Therefore, independent of $\lambda^{(1)}$ and $\lambda^{(2)}$,

$$\frac{\partial^2\mathcal{Q}}{\partial\theta^2}\le 0 \qquad \text{(A$\cdot$ 9)}$$

along any direction.

Proof (b)

The Q-function $Q(\lambda',\lambda)$ can be rewritten as

$$
\begin{aligned}
&\mathcal{Q}(\lambda',\lambda)\\
&=\sum_{i=1}^{N}P(\boldsymbol{O},q_1=i|\lambda')\log\pi_i\\
&+\sum_{i,j=1}^{N}\sum_{t=1}^{T-1}P(\boldsymbol{O},q_t=i,q_{t+1}=j|\lambda')\log a_{ij}\\
&+\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t\in T(\boldsymbol{O},g)}P(\boldsymbol{O},q_t=i,l_t=g|\lambda')\\
&\cdot\Bigg(\log w_{ig}-\frac{n_g}{2}\log(2\pi)\\
&+\frac{1}{2}\sum_{s=1}^{n_g}\log y_{igs}-\frac{1}{2}\sum_{s=1}^{n_g}y_{igs}z_{tigs}^2\Bigg), \qquad \text{(A$\cdot$ 10)}
\end{aligned}
$$

where y_{igs}, $s=1,2,\ldots,n_g$ are are eigenvalues of $\boldsymbol{C}_{ig}$, $\boldsymbol{e}_{igs}$, $s=1,2,\ldots,n_g$ are an orthonormal set of eigenvectors of $\boldsymbol{C}_{ig}$, and $z_{tigs}=(V(\boldsymbol{o}_t)-\boldsymbol{\mu}_{ig})^T\boldsymbol{e}_{igs}$.

When λ approaches ∞ or the boundary of the parameter space, one of the following conditions holds.

1) $\pi_i\to 0$
2) $a_{ij}\to 0$
3) $z_{tigs}^2\to\infty$
4) $w_{ig}\to 0$
5) $y_{igs}\to 0$
6) $y_{igs}\to\infty$

When one of the conditions 1)–5) holds, it is obvious that $Q(\lambda',\lambda)\to-\infty$ because one of the terms in Eq. (A· 10) approaches $-\infty$. In the case where $y_{igs}\to\infty$, from the condition on observations $\boldsymbol{o}_t$, described in Theorem 2, z_{tigs}^2 has a nonzero positive value at some t. Thus,

$$\log y_{igs}-z_{tigs}^2y_{igs}\to-\infty. \qquad \text{(A$\cdot$ 11)}$$

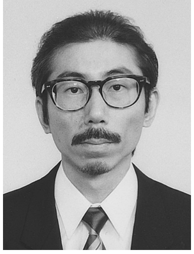
As a result, $Q(\lambda',\lambda)\to-\infty$, as λ approaches the finite boundary of the parameter space or the point at ∞.

Proof (c)

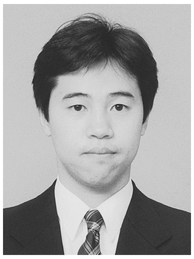
From Proof (a), if there are multiple critical points, they are isolated. Assume that $\boldsymbol{C}_{ig}=\boldsymbol{\tau}_{ig}^T\boldsymbol{\tau}_{ig}$, where $\boldsymbol{\tau}_{ig}$'s are triangular and positive definite. We can rewrite Eq. (A· 1) as

$$
\begin{aligned}
&\mathcal{Q}(\lambda',\lambda)\\
&=\sum_{i=1}^{N}P(\boldsymbol{O},q_1=i|\lambda')\log\pi_i\\
&+\sum_{i,j=1}^{N}\sum_{t=1}^{T-1}P(\boldsymbol{O},q_t=i,q_{t+1}=j|\lambda')\log a_{ij}\\
&+\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t\in T(\boldsymbol{O},g)}P(\boldsymbol{O},q_t=i,l_t=g|\lambda')\\
&\cdot\Bigg(\log w_{ig}-\frac{n_g}{2}\log(2\pi)+\log|\boldsymbol{\tau}_{ig}|\\
&-\frac{1}{2}||\boldsymbol{\tau}_{ig}(V(\boldsymbol{o}_t)-\boldsymbol{\mu}_{ig})||^2\Bigg). \qquad \text{(A$\cdot$ 12)}
\end{aligned}
$$

The change of variables $\{\pi_i,\ a_{ij},\ w_{ig},\ \boldsymbol{\mu}_{ig},\ \boldsymbol{C}_{ig}\} \rightarrow \{\pi_i,\ a_{ij},\ w_{ig},\ \boldsymbol{\mu}_{ig},\ \boldsymbol{\tau}_{ig}\}$, which is a diffeomorphism, maps critical points onto critical points. Therefore, the global maximum is the unique ciritical point since Eq. (A· 12) is convex with respect to $\pi_i, a_{ij}, w_{ig}, \boldsymbol{\mu}_{ig}, \boldsymbol{\tau}_{ig}$.



Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, in 1984, and the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1986 and 1989, respectively. From 1989 to 1996 he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. Since 1996 he has been with the Department of Computer Science, Nagoya Institute of Technology as Associate Professor. He is a co-recipient of both the Best Paper Award and the Inose Award from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. His research interests include speech coding, speech synthesis and recognition and multimodal signal processing. He is a member of IEEE, ISCA, IPSJ, ASJ and JSAI.



Takashi Masuko received the B.E. degree in computer science, and the M.E. degree in intelligence science from Tokyo Institute of Technology, Tokyo, Japan, in 1993 and 1995, respectively. In 1995, he joined the Precision and Intelligence Laboratory, Tokyo Institute of Technology, as a Research Associate. He is currently a Research Associate at the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. He is a co-recipient of both the Best Paper Award and the Inose Award from the IEICE in 2001. His research interests include speech synthesis, speech recognition, speech coding, and multimodal interface. He is a member of IEEE, ISCA and ASJ.

Noboru Miyazaki was born in Ehime, Japan, in 1972. He received the B.E. degree in information engineering and the M.E. degree in intelligence science from Tokyo Institute of Technology, Tokyo, Japan, in 1995 and 1997, respectively. In 1997, he joined the Basic Research Laboratories, Nippon Telegraph and Telephone Corporation (NTT), Japan as a researcher. He is currently a researcher at the NTT Communication Science Laboratories. He is a co-recipient of both the Paper Award and the Inose Award from the IEICE in 2001. His research interests include speech recognition and spoken dialogue processing. He is a member of ASJ.

Takao Kobayashi received the B.E. degree in electrical engineering, the M.E. and Dr.Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 1977, 1979, and 1982, respectively. In 1982, he joined the Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology as a Research Associate. He became an Associate Professor at the same Laboratory in 1989. He is currently a Professor at the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. He is a co-recipient of both the Best Paper Award and the Inose Award from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. His research interests include speech analysis and synthesis, speech coding, speech recognition, and multimodal interface. He is a member of IEEE, ISCA, IPSJ and ASJ.