PAPER

A Hidden Semi-Markov Model-Based Speech Synthesis System

Heiga ZEN[†a)], ***Nonmember***, **Keiichi TOKUDA**[†b)], **Takashi MASUKO**[††*], **Takao KOBAYASIH**[††c)], ***and*** **Tadashi KITAMURA**[†d)], ***Members***

SUMMARY A statistical speech synthesis system based on the hidden Markov model (HMM) was recently proposed. In this system, spectrum, excitation, and duration of speech are modeled simultaneously by context-dependent HMMs, and speech parameter vector sequences are generated from the HMMs themselves. This system defines a speech synthesis problem in a generative model framework and solves it based on the maximum likelihood (ML) criterion. However, there is an inconsistency: although state duration probability density functions (PDFs) are explicitly used in the synthesis part of the system, they have not been incorporated into its training part. This inconsistency can make the synthesized speech sound less natural. In this paper, we propose a statistical speech synthesis system based on a hidden semi-Markov model (HSMM), which can be viewed as an HMM with explicit state duration PDFs. The use of HSMMs can solve the above inconsistency because we can incorporate the state duration PDFs explicitly into both the synthesis and the training parts of the system. Subjective listening test results show that use of HSMMs improves the reported naturalness of synthesized speech.

key words: *hidden Markov model, hidden semi-Markov model, HMM-based speech synthesis*

1. Introduction

A statistical speech synthesis system based on the hidden Markov model (HMM) [1], [2] was recently developed. In this system, spectrum, excitation and duration of speech are modeled simultaneously by context-dependent HMMs, and speech parameter vector sequences are generated from the HMMs themselves [2]. It can synthesize speech with various voice characteristics by transforming its model parameters. For example, either a speaker adaptation [3], [4], a speaker interpolation [5], or an eigenvoice technique [6] was applied to this system, and it was shown that the system could modify its voice characteristics.

For any text-to-speech (TTS) synthesis system, controlling timing of events in speech signals is one of the most difficult problems, since there are many contextual factors that affect timing (e.g., phone identity, accent, stress, location, and part-of-speech). Furthermore, a number of factors that affect duration interact with each other. Thus, a variety of approaches to controlling timing using statistical models have been proposed [7]–[10].

In the HMM-based speech synthesis system, rhythm and tempo of synthesized speech is controlled by decision tree-clustered state duration models [11]–[13]. One of the major limitations of the HMM is that it does not adequately represent the temporal structure of speech. This is because state duration probability density functions (PDFs) of HMMs are implicitly modeled by their state self-transition probabilities. To overcome this limitation, the HMM-based speech synthesis system represents state durations PDFs explicitly by Gaussian distributions [11]. They are estimated from statistical variables obtained in the last iteration of the expectation-maximization (EM) algorithm [14], and then clustered by phonetic decision trees [15]. They are not re-estimated in the EM iteration. In the synthesis part, a sentence HMM corresponding to a text arbitrarily chosen to be synthesized is constructed by concatenating context-dependent HMMs. Then the speech parameter generation algorithm generates sequences of speech parameter vectors for the given HMM [16]. The state duration PDFs are explicitly used in the speech parameter generation procedure.

This system defines a speech synthesis problem in a generative model framework and solves it using the maximum likelihood (ML) criterion. However, there is an inconsistency: although state duration PDFs are explicitly used to generate speech parameter vector sequences from the HMMs, they have not been incorporated into the expectation step of the EM algorithm. This inconsistency can make the synthesized speech sound less natural. In this paper, we propose a statistical speech synthesis system based on a hidden semi-Markov model (HSMM), which can be viewed as an HMM with explicit state duration PDFs. The use of HSMMs can solve the above inconsistency because we can incorporate the state duration PDFs explicitly into both the synthesis and the training parts of the system.

The rest of this paper is organized as follows. Section 2 reviews the HMM-based speech synthesis system. Section 3 describes the generalized forward-backward algorithm, parameter reestimation formulae, decision tree-based context clustering, and speech parameter generation algorithm for the HSMM. Subjective listening test results are presented in Sect. 4. Concluding remarks and future plans are presented

Manuscript received July 27, 2006.
Manuscript revised December 11, 2006.

[†]The authors are with the Department of Computer Science and Engineering, 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 Corporate Research & Development Center, Toshiba Corporation.
a) E-mail: zen@lavender.ics.nitech.ac.jp
b) E-mail: tokuda@nitech.ac.jp
c) E-mail: takao.kobayashi@ip.titech.ac.jp
d) E-mail: kitamura@nitech.ac.jp
DOI: 10.1093/ietisy/e90–d.5.825

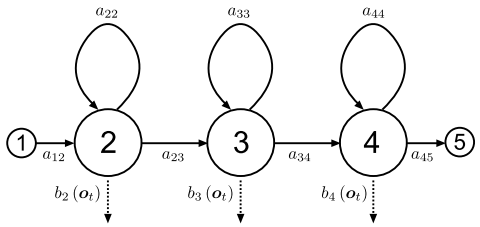


Fig. 1 Example of a five-state, left-to-right, no-skip HMM.

in the final section.

2. HMM-Based Speech Synthesis System

2.1 Forward-Backward Algorithm

An N-state continuous HMM λ is specified by sets of initial state probabilities $\{\pi_i\}_{i=1}^N$, state transition probabilities $\{a_{ij}\}_{i,j=1}^N$, and state output PDFs $\{b_i(\cdot)\}_{i=1}^N$. Here, we assume that the first and N-th states are beginning and ending null states as illustrated in Fig. 1, respectively. Thus, the initial state probabilities $\{\pi_i\}_{i=1}^N$ become $\pi_1 = 1$ and $\pi_2 =, \ldots, \pi_N = 0$.

For the given λ, the output probability of an observation vector sequence $\boldsymbol{o} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$ of length T can be computed efficiently using the forward-backward algorithm [17]. Partial forward probability variables $\alpha_t(\cdot)$ and partial backward probability variables $\beta_t(\cdot)$ are defined as follows:

$$\alpha_0(j) = \begin{cases} 1 & j = 1 \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

$$\alpha_t(j) = P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_t, q_t = j \mid \lambda) \tag{2}$$

$$= \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(\boldsymbol{o}_t) \quad \begin{pmatrix} t = 1, 2, \ldots, T \\ 1 \le j \le N \end{pmatrix}, \tag{3}$$

$$\beta_{T+1}(i) = \begin{cases} 1 & i = N \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

$$\beta_T(i) = a_{iN} \beta_{T+1}(N) \quad (1 \le i \le N), \tag{5}$$

$$\beta_t(i) = P(\boldsymbol{o}_{t+1}, \ldots, \boldsymbol{o}_T \mid q_t = i, \lambda) \tag{6}$$

$$= \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j) \quad \begin{pmatrix} t = T-1, \ldots, 1 \\ 1 \le i \le N \end{pmatrix}, \tag{7}$$

where $a_{11} =, \ldots, = a_{N1} = 0$, and $a_{N2} =, \ldots, = a_{NN} = 0$, $q_t = j$ denotes being the j-th state at time t, and we assume that $b_1(\cdot) = b_N(\cdot) = 1$. From Eqs. (3) and (7), $P(\boldsymbol{o} \mid \lambda)$ is given by

$$P(\boldsymbol{o} \mid \lambda) = \sum_{i=1}^{N} P(\boldsymbol{o}, q_t = i \mid \lambda) \tag{8}$$

$$= \sum_{i=1}^{N} \alpha_t(i) \cdot \beta_t(i) \quad (1 \le t \le T). \tag{9}$$

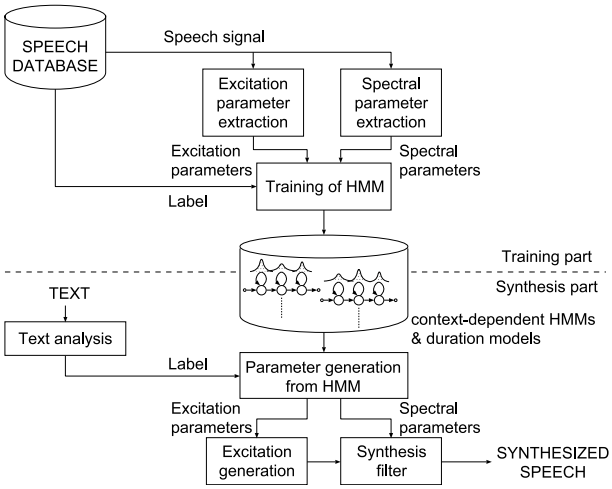


Fig. 2 Overview of a typical HMM-based speech synthesis system.

Generally, computational complexity of the above recursions is on the order of $O(N^2T)$. However, if a simple left-to-right structure illustrated in Fig. 1 is assumed, it reduces to $O(NT)$.

2.2 Duration Handling in an HMM-Based Speech Synthesis System

Figure 2 is an overview of a typical HMM-based speech synthesis system [2]. It consists of training and synthesis parts. In the training part, spectrum (e.g., mel-cepstral coefficients and their dynamic features) and excitation (e.g., $\log F_0$, and its dynamic features) parameters are extracted from a speech database and modeled by context-dependent HMMs. Although sequences of mel-cepstral coefficients can be modeled by continuous HMMs, sequences of $\log F_0$ cannot be modeled using continuous or discrete HMMs without heuristic assumptions since each $\log F_0$ observation can be viewed as consisting of a one-dimensional continuous $\log F_0$ value (voiced regions) or a discrete symbol, which represents an unvoiced frame (unvoiced regions). To model this kind of observation, HMMs based on multi-space probability distributions (MSD-HMMs) have been proposed [18]. An MSD-HMM includes both discrete and continuous HMMs as its special cases and can model the sequences of $\log F_0$ with no heuristic assumptions.

In the synthesis part, first a text to be synthesized is converted to a context-dependent label sequence and then the sentence HMM λ is constructed by concatenating the context-dependent HMMs based on the label sequence. Second, its state durations are determined so as to maximize their probabilities

$$\log P(\boldsymbol{d} \mid \lambda) = \sum_{j=1}^{N} \log p_j(d_j), \tag{10}$$

where $\boldsymbol{d} = \{d_1, d_2, \ldots, d_N\}$ is a set of state durations, d_j is the state duration at the j-th state, N is the number of states in the sentence HMM λ, and $p_j(\cdot)$ denotes the state duration

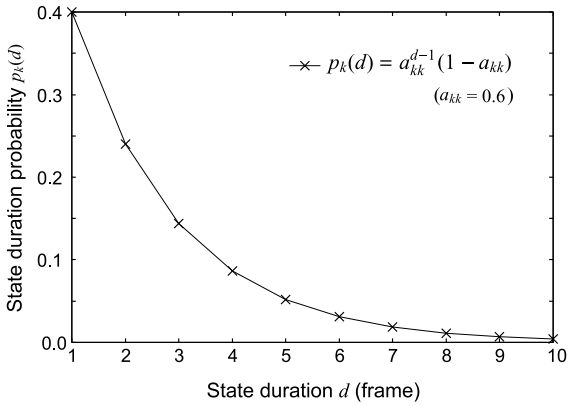


Fig. 3 Example of state duration probability of an HMM ($a_{kk} = 0.6$).

PDF of the j-th state. However, one of the major limitations of the HMM is that it does not adequately represent the temporal structure of speech. This is because state duration PDFs of HMMs are implicitly modeled by their state self-transition probabilities. This means that the probability of d consecutive observations in the j-th state is given by the probability of taking the self-loop at the j-th state for d times as

$$p_j(d) = a_{jj}^{d-1} \cdot \left(1 - a_{jj}\right). \tag{11}$$

The above equation shows that state duration probabilities follow a geometric distribution. Figure 3 plots an example of a state duration probability of an HMM. It can be seen from the figure that state duration probability decreases exponentially with time. Accordingly, the state durations that maximize Eq. (10) are determined as

$$\bar{\boldsymbol{d}} = \arg\max_{\boldsymbol{d}} \log P(\boldsymbol{d} \mid \lambda) \tag{12}$$

$$= \arg\max_{d_1,\ldots,d_N} \sum_{j=1}^{N} \left\{(d_j - 1)\log a_{jj} + \log\left(1 - a_{jj}\right)\right\} \tag{13}$$

$$= \{1, \ldots, 1\}. \tag{14}$$

The above equations show that all expected state durations become 1. This is not useful for controlling the temporal structure of speech. To avoid this problem, the HMM-based speech synthesis system represents state durations PDFs explicitly by Gaussian distributions [11].† They are estimated from statistical variables obtained in the last iteration of the EM algorithm. The mean ξ_j and the variance σ_j^2 of the state duration at the j-th state are estimated as

$$p_j(d) = \mathcal{N}\left(d \mid \xi_j, \sigma_j^2\right), \tag{15}$$

$$\xi_j = \frac{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(j) \cdot (t_1 - t_0 + 1)}{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(j)}, \tag{16}$$

$$\sigma_j^2 = \frac{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(j) \cdot (t_1 - t_0 + 1)^2}{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(j)} - \xi_j^2, \tag{17}$$

where $\chi_{t_0,t_1}(j)$ is the probability of occupying the j-th state from time t_0 to t_1, which can be written as

$$\chi_{t_0,t_1}(j) = P\left(q_{t_0-1} \neq j, q_{t_0} = \ldots = q_{t_1} = j, q_{t_1+1} \neq j \mid \boldsymbol{o}, \lambda\right)$$

$$= \frac{1}{P(\boldsymbol{o} \mid \lambda)} \left\{\sum_{\substack{i=1 \\ i \neq j}}^{N} \alpha_{t_0-1}(i) a_{ij}\right\} \cdot a_{jj}^{t_1 - t_0}$$

$$\cdot \prod_{s=t_0}^{t_1} b_j(\boldsymbol{o}_s) \left\{\sum_{\substack{k=1 \\ k \neq j}}^{N} a_{jk} b_k(\boldsymbol{o}_{t_1+1}) \beta_{t_1+1}(k)\right\}. \tag{18}$$

Since each state duration PDF is represented by a Gaussian distribution, the state durations that maximize Eq. (10) are determined as

$$\bar{\boldsymbol{d}} = \arg\max_{\boldsymbol{d}} \log P(\boldsymbol{d} \mid \lambda_l) \tag{19}$$

$$= \arg\max_{d_1,\ldots,d_K} \sum_{k=1}^{K} \log \mathcal{N}\left(d_k \mid \xi_k, \sigma_k^2\right) \tag{20}$$

$$= \{\xi_1, \ldots, \xi_K\}. \tag{21}$$

Third, the speech parameter generation algorithm [16] generates the sequences of mel-cepstral coefficients and $\log F_0$ values that maximize their output probabilities. The state duration PDFs and the expected state durations $\bar{\boldsymbol{d}}$ are used in the speech parameter generation procedure. Finally, a speech waveform is synthesized directly from the generated speech parameter vectors by a speech synthesis filter.

This system defines a speech synthesis problem in a generative model framework and solves it based on the ML criterion. However, there is an inconsistency: although state duration PDFs are explicitly used for generating speech parameter vector sequences from the HMMs, they have not been incorporated into the expectation step of the EM algorithm: model parameters are estimated without considering the state duration PDFs. This inconsistency can make the synthesized speech sound less natural.

3. HSMM-Based Speech Synthesis System

To resolve the inconsistency of the HMM-based speech synthesis system, we introduce a hidden semi-Markov model

†Although the gamma and log Gaussian distributions have been applied to state duration modeling in HMM-based speech synthesis [12], [13], the Gaussian distribution is widely used because it is mathematically easy to use (e.g., easy to derive speaker adaptation). It is obvious that modeling state durations by continuous distributions is inappropriate in the sense of statistical modeling because the state durations of HMMs are inherently discrete. However, the use of continuous distributions provides better flexibility to control the temporal structure of synthesized speech.

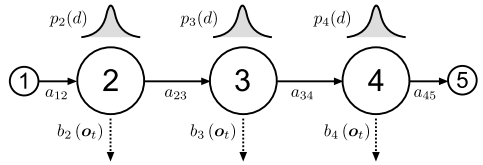


Fig. 4 Example of a five-state, left-to-right, no skip HSMM.

(HSMM) [19]–[21]. This model can be viewed as an HMM with explicit state duration PDFs. The use of HSMMs can solve the above inconsistency because we can incorporate the state duration PDFs explicitly into both the synthesis and the training parts of the system. In this section, the generalized forward-backward algorithm (expectation step), parameter reestimation formulae (maximization step), decision tree-based context clustering technique, and speech parameter generation algorithms, which are required to build an HSMM-based speech synthesis system, are described.

3.1 Generalized Forward-Backward Algorithm

An N-state continuous HSMM λ' is specified by sets of initial state probabilities $\{\pi_i\}_{i=1}^N$, state transition probabilities $\{a_{ij}\}_{i,j=1}^N$, state output PDFs $\{b_i(\cdot)\}_{i=1}^N$, and state duration PDFs $\{p_i(\cdot)\}_{i=1}^N$. Here, we assume that the first and N-th states are beginning and ending null states, respectively, as illustrated in Fig. 4. Thus, the initial state probabilities $\{\pi_i\}_{i=1}^N$ become $\pi_1 = 1$ and $\pi_2 =, \ldots, \pi_N = 0$.

For the given HSMM λ', the output probability of an observation vector sequence $\boldsymbol{o} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$ of length T can be computed efficiently using the generalized forward-backward algorithm [21]–[23]. Partial forward probability variables $\alpha'_t(\cdot)$ and partial backward probability variables $\beta'_t(\cdot)$ are defined as follows:

$$\alpha'_0(j) = \begin{cases} 1 & j = 1 \\ 0 & \text{otherwise} \end{cases}, \tag{22}$$

$$\begin{aligned}\alpha'_t(j) &= P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_t, q_t = j \mid q_{t+1} \neq j, \lambda') \\ &= \sum_{d=1}^{t} \sum_{\substack{i=1, \\ i\neq j}}^{N} \alpha'_{t-d}(i) a_{ij} p_j(d) \prod_{s=t-d+1}^{t} b_j(\boldsymbol{o}_s) \\ &\quad \left(\begin{array}{l} t = 1, 2, \ldots, T \\ 1 \leq j \leq N \end{array} \right),\end{aligned} \tag{23}$$

$$\beta_{T+1}(i) = \begin{cases} 1 & i = N \\ 0 & \text{otherwise} \end{cases}, \tag{24}$$

$$\beta'_T(i) = a_{iN} \beta_{T+1}(N) \quad (1 \leq i \leq N), \tag{25}$$

$$\begin{aligned}\beta'_t(i) &= P(\boldsymbol{o}_{t+1}, \ldots, \boldsymbol{o}_T, q_t = i \mid q_{t+1} \neq i, \lambda') \\ &= \sum_{d=1}^{T-t} \sum_{\substack{j=1, \\ j\neq i}}^{N} a_{ij} p_j(d) \prod_{s=t+1}^{t+d} b_j(\boldsymbol{o}_s) \beta'_{t+d}(j) \\ &\quad \left(\begin{array}{l} t = T-1, \ldots, 1 \\ 1 \leq i \leq N \end{array} \right),\end{aligned} \tag{26}$$

where $a_{11} =, \ldots, = a_{N1} = 0$, $a_{N2} =, \ldots, = a_{NN} = 0$, and we assume that $p_1(\cdot) = p_N(\cdot) = 0$, and $b_1(\cdot) = b_N(\cdot) = 1$. From the above equations, $P(\boldsymbol{o} \mid \lambda')$ is given by

$$\begin{aligned}P(\boldsymbol{o} \mid \lambda') &= \sum_{i=1}^{N} \sum_{\substack{j=1, \\ j\neq i}}^{N} \sum_{d=1}^{t} \alpha'_{t-d}(i) a_{ij} p_j(d) \\ &\quad \cdot \prod_{s=t-d+1}^{t} b_j(\boldsymbol{o}_s) \beta'_t(j) \quad (1 \leq t \leq T).\end{aligned} \tag{27}$$

The drawback of the HSMMs is that the above recursions require on the order of $O(N^2T^2)$ calculations, as compared with $O(N^2T)$ of the HMM. If a simple left-to-right structure illustrated in Fig. 4 is assumed, it reduces to $O(NT^2)$. Furthermore, by limiting the maximum duration to D, it further reduces to $O(NDT)$ [22]. Although the use of HSMMs increases computational cost, it is still possible to perform the above recursions using the currently available computational resources.

3.2 Parameter Reestimation Formulae

The ML criterion is used to estimate parameters of HSMMs. In common with the HMM training, the EM algorithm may be used. Let us assume that the state duration probability of the j-th state of an HSMM λ' is modeled by a Gaussian distribution[†] with mean ξ_j and variance σ_j^2. The reestimation formulae of ξ_j and σ_j^2 are derived as follows:

$$p_j\left(d_j\right) = \mathcal{N}\left(d_j \mid \xi_j, \sigma_j^2\right), \tag{28}$$

$$\bar{\xi}_j = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi'_{t_0,t_1}(j) \cdot (t_1 - t_0 + 1)}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi'_{t_0,t_1}(j)}, \tag{29}$$

$$\bar{\sigma}_j^2 = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi'_{t_0,t_1}(j) \cdot (t_1 - t_0 + 1)^2}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi'_{t_0,t_1}(j)} - \left(\bar{\xi}_j\right)^2, \tag{30}$$

where $\chi'_{t_0,t_1}(j)$ is the probability of occupying the j-th state of the HSMM λ' from time t_0 to t_1, which can be written as

$$\begin{aligned}\chi'_{t_0,t_1}(j) &= \frac{1}{P(\boldsymbol{o} \mid \lambda')} \sum_{\substack{i=1 \\ i\neq j}}^{N} \alpha'_{t_0-1}(i) a_{ij} \cdot \prod_{s=t_0}^{t_1} b_j(\boldsymbol{o}_s) \\ &\quad \cdot p_j(t_1 - t_0 + 1) \cdot \beta'_{t_1}(j),\end{aligned} \tag{31}$$

In the HMM-based speech synthesis system, the MSD-HMMs have been used to model $\log F_0$ sequences. Thus, we derive reestimation formulae for HSMM based on multi-space probability distributions (MSD-HSMMs) in order to

[†]The HSMM with continuous state duration PDFs is also known as continuously variable duration HMM (CVD-HMM) [21].

construct the HSMM-based speech synthesis system.

A sample space composed of G spaces is considered. Each space is an n_g-dimensional real space $\mathbb{R}^{n_g}$, specified by a space index g. We consider that each observation $\boldsymbol{o}_t$ consists of a set of space indexes X_t (e.g., $X_t = \{1\}$, $X_t = \{1, 3, 5\}$, or $X_t = \{1, 2, \ldots, G\}$) and a continuous random variable $\boldsymbol{x}_t \in \mathbb{R}^{n_g}$, that is,

$$\boldsymbol{o}_t = (X_t, \boldsymbol{x}_t), \tag{32}$$

where all spaces specified by each X_t should have the same dimensionality. On the other hand, X_t does not necessarily include all indices specifying the same dimensional spaces. It is noted that both the observation vector $\boldsymbol{x}_t$ and the space index set X_t are random variables that are determined by an observation device (or feature extractor) at each observation.

Each space has its probability w_g, where $\sum_{g=1}^{G} w_g = 1$. If $n_g > 0$, each space has a PDF $f_g(\boldsymbol{x}_t), \boldsymbol{x}_t \in \mathbb{R}^{n_g}$, where $\int_{\mathbb{R}^{n_g}} f_g(\boldsymbol{x}_t)\, d\boldsymbol{x}_t = 1$. If $n_g = 0$, we assume that $\boldsymbol{x}_t$ takes only one sample point. Therefore, we have $f_g(\boldsymbol{x}_t) = 1$ if $n_g = 0$. Accordingly, the output probability of the observation $\boldsymbol{o}_t$ for the j-th state is defined by

$$b_j(\boldsymbol{o}_t) = \sum_{g \in S(\boldsymbol{o}_t)} w_{jg} f_{jg}(V(\boldsymbol{o}_t)), \tag{33}$$

where

$$S(\boldsymbol{o}_t) = X_t, \quad V(\boldsymbol{o}_t) = \boldsymbol{x}_t. \tag{34}$$

Let us assume that $f_{jg}(\cdot)$, $n_g > 0$ is the n_g-dimensional multi-variate Gaussian distribution with the mean vector $\boldsymbol{\mu}_{jg}$ and covariance matrix $\boldsymbol{\Sigma}_{jg}$. The reestimation formulae of w_{jg}, $\boldsymbol{\mu}_{jg}$, and $\boldsymbol{\Sigma}_{jg}$ are derived as follows:

$$\bar{w}_{jg} = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j,g)}{\sum_{h=1}^{G} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j,h)}, \tag{35}$$

$$\bar{\boldsymbol{\mu}}_{jg} = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \zeta_t^d(j,g)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j,g)}, \quad n_g > 0 \tag{36}$$

$$\bar{\boldsymbol{\Sigma}}_{jg} = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \eta_t^d(j,g)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j,g)}, \quad n_g > 0 \tag{37}$$

where $\gamma_t^d(j,g)$, $\zeta_t^d(j,g)$, and $\eta_t^d(j,g)$ are the occupancy probability and first and second order statistics, respectively, given by

$$\gamma_t^d(j,g) = \frac{1}{P(\boldsymbol{o} \mid \lambda')} \sum_{\substack{i=1, \\ i \neq j}}^{N} \alpha'_{t-d}(i) a_{ij} p_j(d) \beta'_t(j) \cdot \sum_{\substack{s=t-d+1, \\ g \in S(\boldsymbol{o}_s)}}^{t} w_{jg} \mathcal{N}\left(V(\boldsymbol{o}_s) \mid \boldsymbol{\mu}_{jg}, \boldsymbol{\Sigma}_{jg}\right) \prod_{\substack{k=t-d+1, \\ k \neq s}}^{t} b_j(\boldsymbol{o}_k), \tag{38}$$

$$\zeta_t^d(j,g) = \frac{1}{P(\boldsymbol{o} \mid \lambda')} \sum_{\substack{i=1, \\ i \neq j}}^{N} \alpha'_{t-d}(i) a_{ij} p_j(d) \beta'_t(j) \cdot \sum_{\substack{s=t-d+1, \\ g \in S(\boldsymbol{o}_s)}}^{t} w_{jg} \mathcal{N}\left(V(\boldsymbol{o}_s) \mid \boldsymbol{\mu}_{jg}, \boldsymbol{\Sigma}_{jg}\right) \prod_{\substack{k=t-d+1, \\ k \neq s}}^{t} b_j(\boldsymbol{o}_k) \cdot V(\boldsymbol{o}_s), \tag{39}$$

$$\eta_t^d(j,g) = \frac{1}{P(\boldsymbol{o} \mid \lambda')} \sum_{\substack{i=1, \\ i \neq j}}^{N} \alpha'_{t-d}(i) a_{ij} p_j(d) \beta'_t(j) \cdot \sum_{\substack{s=t-d+1, \\ g \in S(\boldsymbol{o}_s)}}^{t} w_{jg} \mathcal{N}\left(V(\boldsymbol{o}_s) \mid \boldsymbol{\mu}_{jg}, \boldsymbol{\Sigma}_{jg}\right) \prod_{\substack{k=t-d+1, \\ k \neq s}}^{t} b_j(\boldsymbol{o}_k) \cdot \left[V(\boldsymbol{o}_s) - \boldsymbol{\mu}_{jg}\right]\left[V(\boldsymbol{o}_s) - \boldsymbol{\mu}_{jg}\right]^{\top}. \tag{40}$$

3.3 Decision Tree-Based Context Clustering

There are a number of contextual factors (e.g., phone identity, accent, stress, location, part-of-speech) that affect spectrum, excitation, and duration of speech. In the HMM-based speech synthesis system, context-dependent models are used to capture these factors. If context-dependent models that take account of more combinations of the above contextual factors are constructed, we should be able to obtain more accurate models. However, as the number of contextual factors increases, the number of possible combinations also increases exponentially. As a result, we cannot estimate model parameters robustly. Furthermore, it is impossible to prepare a speech database that includes every possible combination of contextual factors.

To avoid this problem, a variety of parameter sharing techniques have been developed [24]–[27]. The use of phonetic decision trees [15] is one good solution to this problem. This technique has been extended for MSD-HMMs [28] and state duration PDFs [11]. In the HMM-based speech synthesis system, distributions of spectrum, excitation, and duration are clustered separately because they have their own influential contextual factors.

Although the decision tree-based context clustering technique was originally been derived for the HMM, it can be applied to the HSMM with no modifications. The following assumptions are made in the same manner as described by Odell [15]:

- The occupancy probabilities $\gamma_t^d(j,g)$ are not altered during the clustering procedure. In practice, careful selection of the initial state assignments ensures that there are no significant changes.

- The contribution of state transition and duration probabilities to the total probability is negligible. This is related to the previous point. Although the transition and duration probabilities will have a significant effect on total likelihood, their contribution would only change if changes occurred in the state assignments. These are assumed to be fixed throughout the clustering procedure, so the contribution of the transition and duration probabilities is constant and unaffected by the clustering.
- The total likelihood of the training data can be approximated by an average of the log likelihoods weighted by the probability of state occupancy. This is an approximation unless the state occupancy probabilities are zero or one, as is the case for deterministic state assignments, but is often nearly true for probabilistic assignments.

These assumptions would also hold in the HSMM. By providing the state occupancy probabilities using the generalized forward-backward algorithm instead of the normal forward-backward algorithm, the state output PDFs of the HSMM can be clustered by decision trees in the same manner as the HMMs. Furthermore, the state duration PDFs can also be clustered by decision trees using the same procedure as described by Yoshimura et al. [11].

3.4 Speech Parameter Generation Algorithm

For the HMM-based speech synthesis system, three algorithms for generating a speech parameter vector sequence for a given HMM have been derived [16]. These algorithms aim to solve the following three problems:

Case 1. For given λ, $\boldsymbol{q}$ and $\boldsymbol{g}$, maximize $P(\boldsymbol{o} \mid \boldsymbol{q}, \boldsymbol{g}, \lambda)$ with respect to $\boldsymbol{o}$,

Case 2. For given λ, maximize $P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{g} \mid \lambda)$ with respect to $\boldsymbol{q}$, $\boldsymbol{g}$, and $\boldsymbol{o}$,

Case 3. For given λ, maximize $P(\boldsymbol{o} \mid \lambda)$ with respect to $\boldsymbol{o}$,

where λ is an MSD-HMM, $\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$ is a state sequence, $\boldsymbol{g} = \{g_1, g_2, \ldots, g_T\}$ is a space index sequence, and g_t is a space index at time t.

In the Case 1 algorithm, the expected state durations $\bar{\boldsymbol{d}}$ (see Eq. (21)) are used to give $\boldsymbol{q}$. Then a speech parameter vector sequence is generated from λ according to given $\boldsymbol{q}$ and $\boldsymbol{g}$. The objective function of the Case 2 algorithm can be factorized as

$$P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{g} \mid \lambda) = P(\boldsymbol{o} \mid \boldsymbol{q}, \boldsymbol{g}, \lambda) P(\boldsymbol{q}, \boldsymbol{g} \mid \lambda) \tag{41}$$

$$= P(\boldsymbol{o} \mid \boldsymbol{q}, \boldsymbol{g}, \lambda) P(\boldsymbol{g} \mid \boldsymbol{q}, \lambda) P(\boldsymbol{q} \mid \lambda). \tag{42}$$

In this algorithm, the expected state durations $\bar{\boldsymbol{d}}$ are used as the initial $\boldsymbol{q}$, and the state duration PDFs $p_j(d)$ are used for giving $P(\boldsymbol{q} \mid \lambda)$. On the other hand, in the Case 3 algorithm, for giving initial $\boldsymbol{q}$, the state duration PDFs are not incorporated explicitly. This is because the Case 3 algorithm is derived based on the EM algorithm for the HMM, not the HSMM. To b the state duration PDFs explicitly, it should be re-derived based on the EM algorithm for the HSMMs.

This algorithm aims to find a critical point of output probability $P(\boldsymbol{o} \mid \lambda')$ with respect to $\boldsymbol{o}$. An auxiliary function of the current speech parameter vector sequence $\boldsymbol{o}$ and the new one $\bar{\boldsymbol{o}}$ is defined by

$$Q(\boldsymbol{o}, \bar{\boldsymbol{o}}) = \sum_{\text{all } \boldsymbol{g}, \boldsymbol{d}, \boldsymbol{q}} P(\boldsymbol{o}, \boldsymbol{g}, \boldsymbol{d}, \boldsymbol{q} \mid \lambda') \log P(\bar{\boldsymbol{o}}, \boldsymbol{g}, \boldsymbol{d}, \boldsymbol{q} \mid \lambda'), \tag{43}$$

where $\boldsymbol{d} = \{d_1, d_2, \ldots, d_N\}$ is a set of state durations and N is the number of HSMM states. It can be shown that substituting $\boldsymbol{o}$, which maximizes Eq. (43) for $\boldsymbol{o}$, increases the output probability unless $\boldsymbol{o}$ is a critical point. Equation (43) can be rewritten as

$$Q(\boldsymbol{o}, \bar{\boldsymbol{o}}) = P(\boldsymbol{o} \mid \lambda) \left\{ -\frac{1}{2} \bar{\boldsymbol{o}}^\top \overline{\boldsymbol{\Sigma}^{-1}} \bar{\boldsymbol{o}} + \bar{\boldsymbol{o}}^\top \overline{\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}} + C \right\} \tag{44}$$

where C is a constant independent of $\bar{\boldsymbol{o}}$, and $\overline{\boldsymbol{\Sigma}^{-1}}$ and $\overline{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}$ are an expected inverse covariance matrix and an expected inverse covariance matrix times mean vector, respectively, given as

$$\overline{\boldsymbol{\Sigma}^{-1}} = \operatorname{diag}\left[\overline{\boldsymbol{\Sigma}_1^{-1}}, \overline{\boldsymbol{\Sigma}_2^{-1}}, \ldots, \overline{\boldsymbol{\Sigma}_T^{-1}} \right], \tag{45}$$

$$\overline{\boldsymbol{\Sigma}_t^{-1}} = \sum_{\tau=1}^{T} \sum_{d=1}^{\tau} \sum_{j=1}^{N} \sum_{h=1}^{G} \delta(t, \tau, d) \cdot \gamma_\tau^d(j, h) \boldsymbol{\Sigma}_{jh}^{-1}, \tag{46}$$

$$\overline{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} = \left[\overline{\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1}^\top, \overline{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2}^\top, \ldots, \overline{\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\mu}_T}^\top \right]^\top, \tag{47}$$

$$\overline{\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t} = \sum_{\tau=1}^{T} \sum_{d=1}^{\tau} \sum_{j=1}^{N} \sum_{h=1}^{G} \delta(t, \tau, d) \cdot \gamma_\tau^d(j, h) \boldsymbol{\Sigma}_{jh}^{-1} \boldsymbol{\mu}_{jh}, \tag{48}$$

$$\delta(t, \tau, d) = \begin{cases} 1 & \tau - d + 1 \le t \le \tau \\ 0 & \text{otherwise} \end{cases}. \tag{49}$$

We assume that a speech parameter vector $\boldsymbol{o}_t$ consists of an M-dimensional static feature vector

$$\boldsymbol{c}_t = [c_t(1), c_t(2), \ldots, c_t(M)]^\top \tag{50}$$

and its first and second order dynamic feature vectors, that is

$$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta \boldsymbol{c}_t^\top, \Delta^2 \boldsymbol{c}_t^\top \right]^\top, \tag{51}$$

where $\Delta \boldsymbol{c}_t$ and $\Delta^2 \boldsymbol{c}_t$ are given by

$$\Delta \boldsymbol{c}_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau) \boldsymbol{c}_{t+\tau}, \quad \Delta \boldsymbol{c}_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} w^{(2)}(\tau) \boldsymbol{c}_{t+\tau}, \tag{52}$$

and $w^{(i)}(\cdot)$ are window coefficients for calculating the i-th order dynamic features. The conditions in Eq. (52) can be arranged in a matrix form:

$$\boldsymbol{o} = \boldsymbol{W} \boldsymbol{c}, \tag{53}$$

where

$$\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \ldots, \boldsymbol{c}_T^\top\right]^\top, \tag{54}$$

$$\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_T]^\top \otimes \boldsymbol{I}_{M\times M}, \tag{55}$$

$$\boldsymbol{W}_t = \left[\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}\right], \tag{56}$$

$$\boldsymbol{w}_t^{(0)} = \left[\underbrace{0, \ldots, 0}_{t-1}, 1, \underbrace{0, \ldots, 0}_{T-t}\right]^\top, \tag{57}$$

$$\boldsymbol{w}_t^{(d)} = \left[\underbrace{0, \ldots, 0}_{t-L_-^{(d)}-1}, w^{(d)}(-L_-^{(d)}), \ldots, w^{(d)}(L_+^{(d)}), \underbrace{0, \ldots, 0}_{T-\left(t+L_+^{(d)}\right)}\right]^\top. \tag{58}$$

Under the conditions in Eq. (53), the static feature vector sequence that maximizes Eq. (43) can be determined by solving the following set of linear equations:

$$\boldsymbol{W}^\top \overline{\boldsymbol{\Sigma}^{-1}} \boldsymbol{W} \bar{\boldsymbol{c}} = \boldsymbol{W} \overline{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}. \tag{59}$$

This can be done efficiently using the Cholesky or QR decomposition [16]. The whole procedure is summarized as follows:

Step 0. Set the initial speech parameter vector sequence $\boldsymbol{c}$;
Step 1. Calculate $\gamma_t^d(j, g)$ using the generalized forward-backward algorithm;
Step 2. Calculate $\overline{\boldsymbol{\Sigma}^{-1}}$ and $\overline{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}$ and solve Eq. (59);
Step 3. Set $\boldsymbol{c} = \bar{\boldsymbol{c}}$. If a certain convergence condition is satisfied, stop; otherwise, go back to Step 1;

Results of an informal experiment showed that the use of the Case 3 algorithm slightly improved the quality of synthesized speech but drastically increased the computational cost compared with the Case 1 algorithm. Therefore, no further experiments evaluating the performance of the Case 3 algorithm were conducted for this study.

4. Experiments

4.1 Experimental Conditions

The first 450 of the phonetically balanced 503 sentences from the ATR Japanese speech database B-set [29], uttered by two female speakers (FTK and FYM) and two male speakers (MHT and MYI), were used for training. The remaining 53 sentences were used for evaluation. Speech signals were sampled at a rate of 16 kHz and windowed with a 5 ms shift, and mel-cepstral coefficients were obtained from STRAIGHT-spectrum [30]. Fundamental frequency values included in the database were used. Aperiodicity measures in the frequency domain based on a ratio between the lower and upper smoothed spectral envelopes to represent the relative energy distribution of aperiodic components [31] were also extracted. Feature vectors consisted of spectrum, F_0, and aperiodicity parameter vectors. The spectrum parameter vectors consisted of 39 STRAIGHT mel-cepstral coefficients including the zeroth coefficient, their delta and delta-delta coefficients. The F_0 parameter vectors consisted of $\log F_0$, its delta and delta-delta. The aperiodicity parameter vectors consisted of average values of the aperiodicity measures in five frequency bands, i.e., 0-1 kHz, 1-2 kHz, 2-4 kHz, 4-6 kHz, and 6-8 kHz [32], and their delta and delta-delta. A seven-state (including the beginning and ending null states), left-to-right, no skip structure was used both for HMM and HSMM. Each state output PDF was composed of spectrum, F_0, and aperiodicity streams. The spectrum and aperiodicity streams were modeled by single multi-variate Gaussian distributions with diagonal covariance matrices. The F_0 stream was modeled by a multi-space probability distribution consisting of a Gaussian distribution for voiced frames and a discrete distribution for unvoiced frames. Each state duration PDF was modeled by a five-dimensional (equal to the number of emitting states in each phoneme model) multivariate Gaussian distribution.

Forty-two Japanese phonemes including a silence and a pause were used, and context-dependent labels were formulated based on phoneme labels and linguistic information included in the database. In this paper, the following contextual factors were taken into account:

- phoneme:
  - the current phoneme
  - the preceding and succeeding two phonemes
- mora:[†]
  - distance between the accent nucleus and position of the current mora in the current accentual phrase
  - the position of the current mora in the current accentual phrase
- morpheme:
  - the part of speech, conjugate type, and conjugate form of the preceding, current, and succeeding morphemes
- accentual phrase:
  - the number of morae in the preceding, current, and succeeding accentual phrases
  - the type of accents in the preceding, current, and succeeding accentual phrases
  - the position of the current accentual phrase in the current breath phrase
- breath phrase:
  - the number of morae, accentual phrases of the preceding, current, and succeeding breath phrases
  - the position of the current breath phrase in the utterance
- utterance:
  - the number of morae, accentual phrases, and breath phrases in the utterance.

The decision tree-based context clustering technique was

[†]A mora is a syllable-sized unit in Japanese.

Table 1 Number of leaf nodes of constructed decision trees for spectrum, F_0, aperiodicity, and duration.

Speakers	Models	Spect.	F_0	Ap.	Dur.
FTK	HMM	610	1,246	621	320
	HSMM	648	1,320	655	321
FYM	HMM	582	1,389	784	387
	HSMM	615	1,479	791	312
MHT	HMM	752	1,139	707	336
	HSMM	807	1,138	747	265
MYI	HMM	469	1,316	819	360
	HSMM	513	1,307	835	273

Table 2 Number of evaluated pairs of speech samples assigned to each combination of speaker and speaking rate.

Speakers	Speaking rates					
	2.00	1.50	1.00	0.75	0.50	Total
FTK	54	54	51	64	63	286
FYM	58	62	55	39	53	267
MHT	54	48	51	68	58	279
MYI	49	50	62	50	57	268
Total	215	214	219	221	231	1,100

separately applied to distributions for spectrum, F_0, aperiodicity, and state duration. We used training procedure described by Zen et al. [33] in this experiment.

4.2 Experimental Results

Table 1 shows the total number of leaf-nodes after decision tree-based context clustering. In this paper, the minimum description length (MDL) criterion [34] was used to stop tree growth [2], [35]. It can be seen from the table that the total number of model parameters for both HMMs and HSMMs are about the same.

The effectiveness of HSMMs was evaluated in a subjective listening test. To compare the duration controlling ability of these two systems, synthesized speech samples in different speaking rates (0.5, 0.75, 1.0, 1.5, or 2.0 × average duration estimated by the state duration PDFs[†]) were used. The ML-based method described by Yoshimura et al. [11] was used to determine state durations.

To evaluate these models in practical conditions, we used the speech parameter generation algorithm considering global variance [32]. This is an extension of the basic speech parameter generation algorithm [16] and a significant improvement over the basic algorithm has been reported.[††]

Twenty-two subjects were presented a pair of synthesized utterances from the HMM and HSMM-based systems in random order and then asked which speech sounded more natural. For each subject, 50 pairs were chosen at random from 1060 pairs of synthesized speech (four speakers × five speaking rates × 53 test sentences). Table 2 shows the number of evaluated pairs assigned to each combination of speaker and speaking rate. This experiment was carried out in a sound proof room using headphones.

Figures 5 and 6 show preference scores averaged over speakers and speaking rates, respectively. It can be seen from the figures that the use of HSMMs improved the reported naturalness of synthesized speech, especially when the speaking rates were slower or faster than the average speaking rates. Interestingly, most of the subjects observed that the use of HSMMs improved the reported naturalness both in duration and in spectrum and excitation.

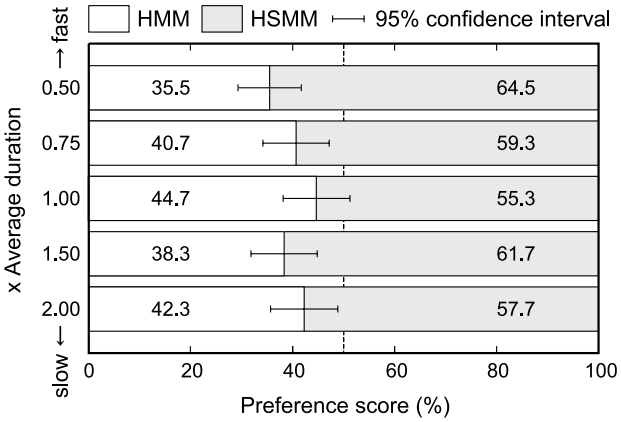


Fig. 5 Preference scores between HMM and HSMM-based systems for different speaking rates.

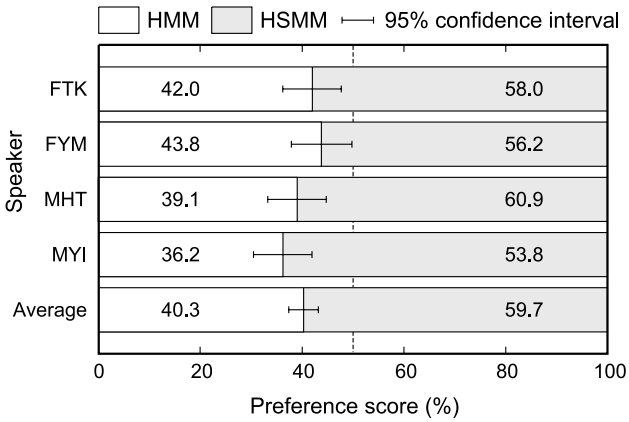


Fig. 6 Preference scores between HMM and HSMM-based systems for different speakers.

5. Conclusion

In this paper, a statistical speech synthesis system based on a hidden semi-Markov model (HSMM), which can be viewed as a hidden Markov model (HMM) with explicit state duration models, was developed and evaluated. The use of HSMMs enables us to explicitly incorporate state duration PDFs into both the synthesis and the training parts of the system. Subjective listening test results showed that the use of HSMMs improved the reported naturalness of synthesized speech.

Future work will focus on the use of other distributions

[†]The speaking rates of average duration for speakers mht, myi, ftk, and fym were 5.33, 8.82, 7.37, and 7.89 (mora/sec), respectively.

[††]Experimental results using the basic speech parameter generation algorithm has been described by Zen et al. [36].

such as gamma or log Gaussian distributions for state duration PDFs.

Acknowledgments

The authors thank Drs. Frank K. Soong, Yoshihiko Nankaku, and Junichi Yamagishi for helpful discussions. This work was partly supported by the MEXT e-Society project.

References

[1] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," Proc. ICASSP, pp.389–392, 1996.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. Eurospeech, pp.2347–2350, 1999.

[3] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," Proc. ICASSP, pp.1611–1614, 1997.

[4] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," Proc. ICASSP, pp.805–808, 2001.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," Proc. Eurospeech, pp.2523–2526, 1997.

[6] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," Proc. ICSLP, pp.1269–1272, 2002.

[7] N. Kaiki, K. Takeda, and Y. Sagisaka, "Linguistic properties in the control of segmental duration for speech synthesis," in Talking Machines: Theories, Models, and Designs, ed. G. Bailly and C. Benoit, pp.255–263, Elsevier Science Publishers, 1992.

[8] M. Riley, "Tree-based modelling of segmental duration," in Talking Machines: Theories, Models, and Designs, ed. G. Bailly and C. Benoit, pp.265–273, Elsevier Science Publishers, 1992.

[9] N. Iwahashi and Y. Sagisaka, "Statistical modelling of speech segment duration by constrained tree regression," IEICE Trans. Inf. & Syst., vol.E83-D, no.7, pp.1550–1559, July 2000.

[10] J. van Santen, C. Shih, B. Möbius, E. Tzoukermann, and M. Tanenblatt, "Multi-lingual duration modelling," Proc. Eurospeech, pp.2651–2654, 1997.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," Proc. ICSLP, pp.29–32, 1998.

[12] Y. Ishimatsu, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Investigation of state duration model based on gamma distribution for HMM-based speech synthesis," IEICE Technical Report, SP2001-81, 2001.

[13] J. Yamagishi, T. Masuko, and Kobayashi, "A study on state duration modeling using lognormal distribution for HMM-based speech synthesis," Proc. ASJ, pp.225–226, March 2004.

[14] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of Royal Statistics Society, vol.39, pp.1–38, 1977.

[15] J. Odell, The Use of Context in Large Vocabulary Speech Recognition, Ph.D. thesis, Cambridge University, 1995.

[16] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. ICASSP, pp.1315–1318, 2000.

[17] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol.77, no.2, pp.257–285, 1989.

[18] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," IEICE Trans. Inf. & Syst., vol.E85-D, no.3, pp.455–464, March 2002.

[19] J. Ferguson, "Variable duration models for speech," Proc. Symposium on the Application Hidden Markov Models to Text and Speech, pp.143–179, 1980.

[20] M. Russell and R. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," Proc. ICASSP, pp.5–8, 1985.

[21] S. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," Comput. Speech Lang., vol.1, pp.29–45, 1986.

[22] C. Mitchell, M. Harper, and L. Jamieson, "On the complexity of explicit duration HMMs," IEEE Trans. Speech Audio Process., vol.3, no.3, pp.213–217, 1995.

[23] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMMs to segment models," IEEE Trans. Speech Audio Process., vol.4, no.5, pp.360–378, 1996.

[24] K.F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," IEEE Trans. Acoust. Speech Signal Process., vol.38, no.4, pp.599–609, 1990.

[25] M. Ostendorf and H. Singer, "HMM topology design using maximum likelihood successive state splitting," Comput. Speech Lang., vol.11, no.1, pp.17–41, 1997.

[26] M.Y. Hwang, X. Huang, and F. Alleva, "Predicting unseen triphones with senones," Proc. ICASSP, pp.311–314, 1993.

[27] P. Woodland and S. Young, "Benchmark DARPA RM results with the HTK portable HMM toolkit," Proc. DARPA Continuous Speech Recognition Workshop, pp.71–76, 1992.

[28] T. Masuko, K. Tokuda, N. Miyazaki, and T. Kobayashi, "Pitch pattern generation using multi-space probability distribution HMM," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J85-D-II, no.7, pp.1600–1609, July 2000.

[29] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Commun., vol.9, pp.357–363, 1990.

[30] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds," Speech Commun., vol.27, pp.187–207, 1999.

[31] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," Proc. MAVEBA, pp.13–15, 2001.

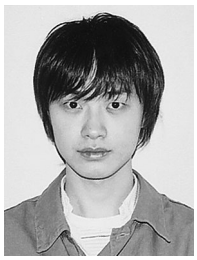
[32] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," Proc. Interspeech (Eurospeech), pp.2801–2804, 2005.

[33] H. Zen and T. Toda, "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005," Proc. Interspeech, pp.93–96, 2005.

[34] J. Rissanen, Stochastic Complexity in Stochastic Inquiry, World Scientific Publishing Company, 1980.

[35] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," Proc. Eurospeech, pp.99–102, 1997.

[36] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," Proc. ICSLP, pp.1185–1180, 2004.



Heiga Zen received the A.E. degree in electronic and information engineering from Suzuka National College of Technology, Suzuka, Japan, in 1999, and the B.E., M.E., and Dr.Eng. degrees in computer science, electrical and computer engineering, and computer science and engineering from Nagoya Institute of Technology, Nagoya, Japan, in 2001, 2003, and 2006, respectively. During 2003, he was an intern researcher at ATR Spoken Language Translation Research Laboratories (ATR-SLT), Kyoto, Japan. From June 2004 to May 2005, he was an intern/co-op researcher in the Human Language Technology group at IBM T.J. Watson Research Center, Yorktown Heights, NY. He is currently a postdoctoral fellow of the MEXT e-Society project at Nagoya Institute of Technology. His research interests include statistical speech recognition and synthesis. He received the Awaya Award in 2006 from the Acoustical Society of Japan (ASJ). He is a member of ASJ.

Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr.Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 1984, 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. From 1996 to 2004 he was an Associate Professor at the Department of Computer Science, Nagoya Institute of Technology as an Associate Professor, and now he is a Professor at the same institute. He is also an Invited Researcher at ATR Spoken Language Translation Research Laboratories (ATR-SLT), Kyoto, Japan, and was a Visiting Researcher at the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2002. He is a co-recipient of the Paper Award and the Inose Award both from the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) in 2001, and the TELECOM System Technology Award from the Telecommunications Advancement Foundation, Japan, in 2001. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society. His research interests include speech coding, speech synthesis and recognition, and statistical machine learning.

Takashi Masuko received the B.E. degree in computer science, M.E. degree in intelligence science, and Dr.Eng degree in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 1993, 1995, and 2003, respectively. In 1995, he joined the Precision and Intelligence Laboratory, Tokyo Institute of Technology as a Research Associate. He is currently with Corporate Research & Development Center, Toshiba Corporation, Kawasaki, 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.

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 of 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.

Tadashi Kitamura received the B.E. degree in electronics engineering from Nagoya Institute of Technology, Nagoya, in 1973, and M.E. and Dr.Eng. degrees from Tokyo Institute of Technology, Tokyo, in 1975 and 1978, respectively. In 1978 He joined the Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology, as a Research Associate. In 1983 he joined Nagoya Institute of Technology, as a Assistant Professor. He is currently a Professor of Graduate School of Engineering of Nagoya Institute of Technology. His current interests include speech processing, image processing and multi-modal biometrics. He is a member of IEEE, ISCA, ASJ, IPSJ and ITE.