Hidden Semi-Markov Model Based Speech Synthesis

Heiga Zen[†]*, Keiichi Tokuda*[†]*, Takashi Masuko*[‡*]* , Takao Kobayashi*[‡]*, Tadashi Kitamura*[†]

† Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, 466-8555 Japan

‡ Tokyo Institute of Technology, Nagatsuta, Midori-ku, Yokohama, 226-8502 Japan

E-mail: {zen,tokuda,kitamura}@ics.nitech.ac.jp, {masuko,Takao.Kobayashi}@ip.titech.ac.jp

Abstract

In the present paper, a hidden-semi Markov model (HSMM) based speech synthesis system is proposed. In a hidden Markov model (HMM) based speech synthesis system which we have proposed, rhythm and tempo are controlled by state duration probability distributions modeled by single Gaussian distributions. To synthesis speech, it constructs a sentence HMM corresponding to an arbitralily given text and determine state durations maximizing their probabilities, then a speech parameter vector sequence is generated for the given state sequence. However, there is an inconsistency: although the speech is synthesized from HMMs with explicit state duration probability distributions, HMMs are trained without them. In the present paper, we introduce an HSMM, which is an HMM with explicit state duration probability distributions, into the HMM-based speech synthesis system. Experimental results show that the use of HSMM training improves the naturalness of the synthesized speech.

1. Introduction

For any text-to-speech (TTS) synthesis system, controlling timing of the events in the speech signal is one of the difficult problems since there are many contextual factors (e.g., phone identity factors, stress-related factors, locational factors) that affect timing. Furthermore, several factors affecting duration interact with one another. Recently, there have been proposed several approaches to controlling timing using statistical models such as linear regression [1], tree regression [2], and sums-of-products model [3]. By using these techniques, rhythm and tempo of speech were successfully controlled with a small amount of parameters.

On the other hand, we have proposed an HMM-based speech synthesis system in which spectrum, F_0 and duration are modeled simultaneously in a unified framework of HMMs [4]. In this system, rhythm and tempo are controlled by state duration probability distributions. One of major limitation of HMMs is that they do not provide an adequate representation of the temporal structure of speech. This is because the probability of state occupancy decreases exponentially with time. To overcome this limitation, in the HMM-based speech synthesis system each state duration probability distribution is explicitly modeled by a single Gaussian distribution. They are estimated from statistical variables obtained in the last iteration of the forward-backward algorithm, and then clustered by a decision tree-based context clustering [5]: they are not reestimated in the Baum-Welch iteration. In the synthesis stage, we construct a sentence HMM corresponding to an arbitrarily given text and determine state durations maximizing their probabilities. Then a speech parameter vector sequence is generated for the given state sequence by speech parameter generation algorithm (case 1 in [6]).

However, there is an inconsistency: although speech is synthesized from HMMs with explicit state duration probability distributions, HMMs are trained without them. In the present paper, we introduce an HSMM, which is an HMM with explicit state duration probability distributions, into not only for synthesis but also training in the HMM-based speech synthesis system.

The rest of the present paper organized as follows. Section 2 describes the likelihood computation of the HMM and overview of the HMM-based speech synthesis system. Section 3 describes the likelihood computation of the HSMM and derives its reestimation formulas to construct an HSMM-based speech synthesis system. Results of subjective listening tests are shown in Section 4. Concluding remarks and future plans are presented in the final section.

2. The hidden Markov model

2.1. Likelihood computation of the HMM

The model likelihood of an HMM λ illustrated in Fig. 1(a) for an observation vector sequence $\boldsymbol{o} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ can be computed efficiently by the forward-backward algorithm. First, we define partial forward likelihood $\alpha_t(\cdot)$ as follows:

$$\alpha_t(j) = P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_t, q_t = j \mid \lambda) \tag{1}$$

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

where a_{ij} is a state transition probability from i-th state to j-th state, $b_j(\boldsymbol{o}_t)$ is an output probability of observation vector $\boldsymbol{o}_t$ from j-th state, N is a total number of HMM states. To begin the recursion Eq. (2), we set $\alpha_1(j) = \pi_j b_j(\boldsymbol{o}_t), 1 \le j \le N$, where π_j is an initial state probability of j-th state. Secondly, partial backward likelihood $\beta_t(\cdot)$ is defined as follows:

$$\beta_t(i) = P(\boldsymbol{o}_{t+1}, \ldots, \boldsymbol{o}_T \mid q_t = i, \lambda) \tag{3}$$

$$= \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_t) \beta_{t+1}(j). \quad 1 \le t \le T,\ 1 \le i \le N \tag{4}$$

To begin the recursion Eq. (18), we set $\beta_T(i) = 1, 1 \le i \le N$. From Eqs. (2) and (4), $P(\boldsymbol{o} \mid \lambda)$ is computed as

$$P(\boldsymbol{o} \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i) \cdot \beta_t(i). \quad 1 \le t \le T \tag{5}$$

2.2. HMM-based speech synthesis

Figure 2 shows the overview of the current HMM-based speech synthesis system. In the training stage, output vector of HMM

[*] Presently, with the Corporate Research & Development Center, Toshiba Corporation.

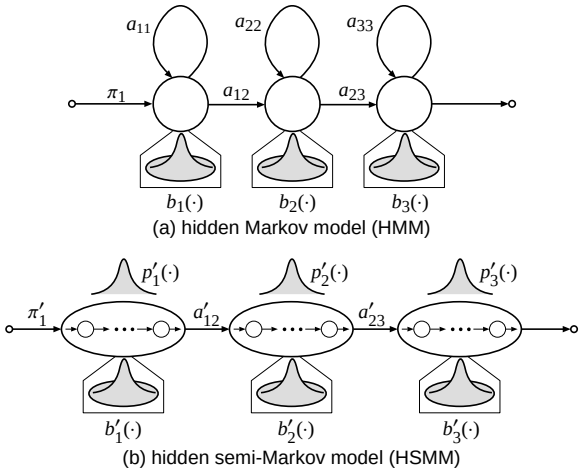

Figure 1: Examples of an HMM and an HSMM with 3-state left-to-right with no skip structures

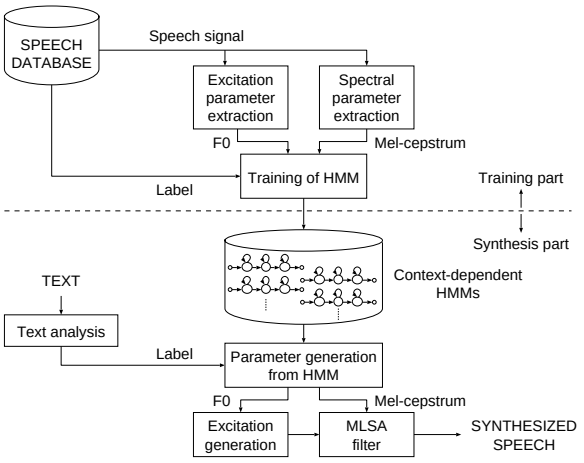

Figure 2: An overview of the HMM-based speech synthesis system

consists of spectrum part and F_0 part. In this system, the spectrum part consists of mel-cepstral coefficients, their delta and delta-delta coefficients and F_0 part consists of $\log F_0$, its delta and delta-delta. Although the spectrum part can be modeled by continuous HMM, F_0 part cannot be modeled by continuous or discrete HMM since the observation sequence of F_0 is composed of one-dimensional continuous value and discrete symbol which represents "unvoiced". To model such observation sequence, we have proposed a new kind of HMM based on multi-space probability distribution (MSD-HMM) [7]. The MSD-HMM includes both discrete HMM and continuous HMM as special cases. As a result, MSD-HMM can model F_0 patterns without heuristic assumption.

In the synthesis stage, first an arbitrarily given text to be synthesized is converted to a context-dependent label sequence and a sentence HMM is constructed by concatenating context-dependent HMMs according to the label sequence. Secondly, we have to determine state durations $\boldsymbol{D}$ of the sentence HMM λ which maximize their probability

$$\log P(\boldsymbol{D} \mid \lambda) = \sum_{k=1}^{K} \log p_k(d_k), \tag{6}$$

where K is a total number of HMM states in sentence HMM λ, d_k is the state duration of k-th state, and $p_k(\cdot)$ is the state duration probability distribution of k-th state. However, one of major limitation of HMMs is that they do not provide an adequate representation of the temporal structure of speech. This is because the probability of state occupancy decreases exponentially with time. The probability of d consecutive observation in state k is the probability of taking the self-loop at state k for d times, which can be written as

$$p_k(d) = a_{kk}^{d} \cdot (1 - a_{kk}). \tag{7}$$

Hence, state durations $\bar{\boldsymbol{D}}$ maximizing Eq. (6) with state duration probability distributions given by Eq. (7) are determined as

$$\bar{\boldsymbol{D}} = \arg\max_{\boldsymbol{D}} \log P(\boldsymbol{D} \mid \lambda) \tag{8}$$

$$= \{1, \ldots, 1\}. \tag{9}$$

To avoid this, state duration probability distributions are explicitly modeled by single Gaussian distributions in the HMM-based speech synthesis system. They are estimated from statistical variables obtained in the last iteration of the forward-backward algorithm. The mean ξ_j and the variance σ_j of the state duration probability distribution of j-th state are estimated as

$$\xi_j = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(j) \cdot (t_1 - t_0 + 1)}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(j)}, \tag{10}$$

$$\sigma_j = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(j) \cdot (t_1 - t_0 + 1)^2}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(j)} - \xi_j^2, \tag{11}$$

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

$$\chi_{t_0,t_1}(j) = \Bigg[\left\{ \sum_{i \neq j} \alpha_{t_0-1}(i) a_{ij} \right\} \cdot \prod_{s=t_0}^{t_1} b_j(\boldsymbol{o}_s) \cdot a_{jj}^{t_1-t_0} \cdot \left\{ \sum_{k \neq j} a_{jk} b_k\left(\boldsymbol{o}_{t_1+1}\right) \beta_{t_1+1}(k) \right\} \Bigg] \Big/ P(\boldsymbol{o} \mid \lambda). \tag{12}$$

Since each state duration probability distribution is modeled by a single Gaussian distribution, state durations $\bar{\boldsymbol{D}}$ maximizing Eq. (6) with Gaussian state duration probability distributions are determined as

$$\bar{\boldsymbol{D}} = \arg\max_{\boldsymbol{D}} \log P(\boldsymbol{D} \mid \lambda) \tag{13}$$

$$= \{\xi_1, \ldots, \xi_K\}. \tag{14}$$

Thirdly, a speech parameter vector sequence is generated for a given state sequence by the parameter generation algorithm (case 1 in [6]). Finally, speech waveform is synthesized directly from the generated speech parameter vector sequence.

However, there is an inconsistency: although HMMs are trained without explicit state duration models, speech parameter vector sequence is generated from HMMs with explicit state duration models.

3. The hidden semi-Markov model

In this section, we describe an HSMM [8, 9] which can be considered as an HMM with explicit state duration probability distributions, and introduce it into not only for synthesis but also training in the HMM-based speech synthesis system.

3.1. Likelihood computation of the HSMM

We can compute the model likelihood of an HSMM λ', which is illustrated in Fig. 1(b), for an observation vector sequence $\boldsymbol{o}$ by a generalized forward-backward algorithm [9]. We can compute partial forward likelihood $\alpha'_t(\cdot)$ and partial backward likelihood $\beta'_t(\cdot)$ recursively as follows:

$$\alpha'_0(j) = \pi_j, \tag{15}$$

$$\alpha'_t(j) = \sum_{d=1}^{t} \sum_{\substack{i=1,\\ i\neq j}}^{N'} \alpha'_{t-d}(i) a'_{ij} p'_j(d) \prod_{s=t-d+1}^{t} b'_j(\boldsymbol{o}_s), \quad 1 \leq t \leq T \tag{16}$$

$$\beta'_T(i) = 1, \tag{17}$$

$$\beta'_t(i) = \sum_{d=1}^{T-t} \sum_{\substack{j=1,\\ j\neq i}}^{N'} a'_{ij} p'_j(d) \prod_{s=t+1}^{t+d} b'_j(\boldsymbol{o}_s) \beta'_{t+d}(j), \quad 1 \leq t < T \tag{18}$$

where a'_{ij}, $b'_j(\boldsymbol{o}_t)$, N', $p'_j(d)$, and π'_j are a state transition probability from i-th state to j-th state, an output probability of observation vector $\boldsymbol{o}_t$ from j-th state, a total number of HSMM states, a state duration probability of j-th state, and an initial state probability of j-th state, respectively. From above equations, $P(\boldsymbol{o} \mid \lambda')$ is given by

$$P(\boldsymbol{o} \mid \lambda') = \sum_{i=1}^{N'} \sum_{\substack{j=1,\\ j\neq i}}^{N'} \sum_{d=1}^{t} \alpha'_{t-d}(i) a'_{ij} p'_j(d) \prod_{s=t-d+1}^{t} b'_j(\boldsymbol{o}_s) \beta'_t(j). \tag{19}$$

3.2. HSMM-based speech synthesis system

In this section, we derive the parameter reestimation formulas of HSMMs, especially for constructing an HSMM-based speech synthesis system.

3.2.1. State duration probability distribution

When the the state duration probability distribution of j-th state of HSMM λ' is modeled by a single Gaussian distribution[1] with mean ξ'_j and variance σ'_j, the parameter reestimation formulas of them are derived as follows:

$$\overline{\xi}'_j = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi'_{t_0,t_1}(j) \cdot (t_1 - t_0 + 1)}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi'_{t_0,t_1}(j)}, \tag{20}$$

$$\overline{\sigma}'_j = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi'_{t_0,t_1}(j) \cdot (t_1 - t_0 + 1)^2}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi'_{t_0,t_1}(j)} - \left(\overline{\xi}'_j\right)^2, \tag{21}$$

where

$$\chi'_{t_0,t_1}(j) = \frac{\sum_{i\neq j} \alpha'_{t_0-1}(i) a'_{ij} \cdot \prod_{s=t_0}^{t_1} b'_j(\boldsymbol{o}_s) \cdot p'_j(t_1 - t_0 + 1) \cdot \beta'_{t_1}(j)}{P(\boldsymbol{o} \mid \lambda')}. \tag{22}$$

3.2.2. State output probability distribution

In the HMM-based speech synthesis system, the MSD-HMM was used for modeling F_0 patterns. Hence, we derive the parameter reestimation formulas for the HSMM based on MSD (MSD-HSMM).

We consider a sample space composed of G spaces. Each space is an n_g-dimensional real space $\mathbb{R}^{n_g}$, specified by space index g. Each space has its probability w_g, where $\sum_{g=1}^{G} w_g = 1$. If $n_g > 0$, each space has a pdf function $f_g(\boldsymbol{x}_t), \boldsymbol{x}_t \in \mathbb{R}^{n_g}$, where $\int f_g(\boldsymbol{x}_t)\, d\boldsymbol{x}_t = 1$. Each event E, which will be considered in the present paper, is represented by a random vector $\boldsymbol{o}_t$ composed of a set of space indices X_t and continuous random variable $\boldsymbol{x}_t \in \mathbb{R}^{n_g}$, that is,

$$\boldsymbol{o}_t = (X_t, \boldsymbol{x}_t), \tag{23}$$

where all spaces specified by X_t are n-dimensional. On the other hand, X_t does not necessarily include all indices specifying n-dimensional spaces. It is noted that not only the observation vector $\boldsymbol{x}_t$ but also the space index set X_t is a random variable, which is determined by an observation device (or feature extractor) at each observation. The output probability of observation $\boldsymbol{o}_t$ from j-th state is defined by

$$b'_j(\boldsymbol{o}_t) = \sum_{g\in S(\boldsymbol{o}_t)} w_{jg} f_{jg}(V(\boldsymbol{o}_t)), \tag{24}$$

where

$$S(\boldsymbol{o}_t) = X_t, \quad V(\boldsymbol{o}_t) = \boldsymbol{x}_t. \tag{25}$$

When $f_{jg}(\cdot)$, $n_g > 0$ is the n_g-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}_{jg}$ and covariance matrix $\boldsymbol{\Sigma}_{jg}$, the reestimation formulas of w_{jg}, $\boldsymbol{\mu}_{jg}$ and $\boldsymbol{\Sigma}_{jg}$ can be derived as follows:

$$\overline{w}_{jg} = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j,g)}{\sum_{h=1}^{G} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j,h)}, \tag{26}$$

$$\overline{\boldsymbol{\mu}}_{jg} = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \zeta_t^d(j,g)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j,g)}, \quad n_g > 0 \tag{27}$$

$$\overline{\boldsymbol{\Sigma}}_{jg} = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \eta_t^d(j,g)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j,g)}, \quad n_g > 0 \tag{28}$$

where

$$\gamma_t^d(j,g) = \sum_{\substack{i=1,\\ i\neq j}}^{N'} \alpha'_{t-d}(i) a'_{ij} p'_j(d) \beta'_t(j) \cdot \sum_{\substack{s=t-d+1,\\ g\in S(\boldsymbol{o}_s)}}^{t} w_{jg} \mathcal{N}\left(V(\boldsymbol{o}_s) \mid \boldsymbol{\mu}_{jg}, \boldsymbol{\Sigma}_{jg}\right) \prod_{\substack{k=t-d+1,\\ k\neq s}}^{t} b'_j(\boldsymbol{o}_k), \tag{29}$$

$$\zeta_t^d(j,g) = \sum_{\substack{i=1,\\ i\neq j}}^{N'} \alpha'_{t-d}(i) a'_{ij} p'_j(d) \beta'_t(j) \cdot \sum_{\substack{s=t-d+1,\\ g\in S(\boldsymbol{o}_s)}}^{t} w_{jg} \mathcal{N}\left(V(\boldsymbol{o}_s) \mid \boldsymbol{\mu}_{jg}, \boldsymbol{\Sigma}_{jg}\right) \prod_{\substack{k=t-d+1,\\ k\neq s}}^{t} b'_j(\boldsymbol{o}_k) \cdot V(\boldsymbol{o}_s), \tag{30}$$

$$\eta_t^d(j,g) = \sum_{\substack{i=1,\\ i\neq j}}^{N'} \alpha'_{t-d}(i) a'_{ij} p'_j(d) \beta'_t(j) \cdot \sum_{\substack{s=t-d+1,\\ g\in S(\boldsymbol{o}_s)}}^{t} w_{jg} \mathcal{N}\left(V(\boldsymbol{o}_s) \mid \boldsymbol{\mu}_{jg}, \boldsymbol{\Sigma}_{jg}\right) \prod_{\substack{k=t-d+1,\\ k\neq s}}^{t} b'_j(\boldsymbol{o}_k) \cdot \left[V(\boldsymbol{o}_s) - \boldsymbol{\mu}_{jg}\right]\left[V(\boldsymbol{o}_s) - \boldsymbol{\mu}_{jg}\right]^{\top}. \tag{31}$$

3.2.3. Decision tree-based context clustering

To construct the HSMM-based speech synthesis system, we have to re-derive the decision tree-based context clustering technique for the HSMM. Because of limitation of space, we cannot provide a complete derivation of it in the HSMM. However, it can be used in the HSMM because it was derived by ignoring the state sequence probability [5].

[1] The reestimation formulas when the state duration probability distribution is modeled by the gamma distribution was derived in [9].

Table 1: The number of distributions after context clustering

speaker	model	Spec.	F_0	Dur.
FTK	HMM	956	1392	404
	HSMM	963	1427	343
FYM	HMM	870	1365	368
	HSMM	874	1360	343
MHT	HMM	969	1133	338
	HSMM	969	1150	313
MYI	HMM	728	1234	377
	HSMM	737	1217	361

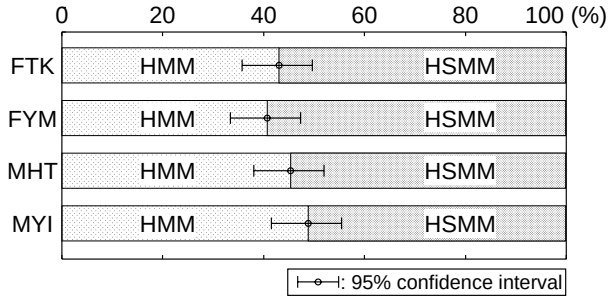

Figure 3: Preference scores

4. Experiments

4.1. Experimental conditions

We used phonetically balanced 450 sentences from ATR Japanese speech database B-set uttered by 2 female speakers (FTK and FYM) and 2 male speakers (MHT and MYI) for training speaker-dependent HMMs and HSMMs. Based on phoneme labels and linguistic information included in the database, we made context-dependent labels. We used 42 phonemes including silence and pause.

Speech signals were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift, and mel-cepstral coefficients were obtained by a mel-cepstral analysis technique. Fundamental frequency (F_0) values included in the database were used. Feature vector consisted of spectrum and F_0 parameter vectors. Spectrum parameter vector consisted of 25 mel-cepstral coefficients including the zeroth coefficient, their delta and delta-delta coefficients. F_0 parameter vector consisted of log F_0, its delta and delta-delta. We used 5-state left-to-right with no skip HMM/HSMM structure. The state output probability distributions were consisted of spectrum and F_0 part. The spectrum part was modeled by a single Gaussian distribution with diagonal covariance matrix. The F_0 part was modeled by an MSD consisted of a single Gaussian distribution with diagonal covariance matrix (voiced space) and a single discrete distribution which outputs only one symbol (unvoiced space).

4.2. Experimental results

Table 1 shows the total number of distributions after context clustering based on an MDL criterion [4, 10]. It can be seen from the table that the total numbers of model parameters are almost same.

To evaluate the effectiveness of the use of HSMM training algorithm, subjective listening tests were conducted. We compared the naturalness of the synthesized speech generated from HMMs and HSMMs by paired comparison tests. Subjects were 8 persons, and presented a pair of synthesized speech from HMMs and HSMMs in random order and then asked which speech sounded more natural. For each subject, 20 test sentences were chosen at random from 53 test sentences not contained in the training data sentence set. Experiments were carried out in a sound proof room.

Figure 3 shows the preference scores. It can be seen from the figure that the training algorithm of the HSMM provides the higher performance than that of the HMM. Interestingly, we have observed that the training algorithm of the HSMM improves the naturalness in not only duration but also spectrum and F_0.

5. Conclusion

In the present paper, we introduced a hidden semi-Markov model (HSMM), which is a hiden Markov model (HMM) with explicit state duration probability distributions, into an HMM-based speech synthesis system which we have proposed. Use of HSMM training enable us to solve an inconsistency between training and synthesis stage. From the results of subjective listening tests, we have shown that the use of HSMM training improved the naturalness of synthesized speech.

Future work will focus on the introduction of other distribution such as gamma distribution or logarithmic Gaussian distribution into state duration probability distribution.

6. Acknowledgments

Authors would like to thank Dr. Frank K. Soong for helpful discussions.

7. References

[1] 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*, G. Bailly and C. Benoit, Eds. Elsevier Science Publishers, 1992, pp. 255–263.

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

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

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. of Eurospeech*, vol. 5, 1999, pp. 2347–2350.

[5] J. Odell, "The use of context in large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, 1995.

[6] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. of ICASSP*, 1995, pp. 660–663.

[7] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. of ICASSP*, 1999, pp. 229–232.

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

[9] S. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer, Speech and Language*, vol. 1, pp. 29–45, 1986.

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