ADAPTIVE TRAINING FOR HIDDEN SEMI-MARKOV MODEL

Junichi Yamagishi, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Yokohama, 226-8502 Japan
Email: {junichi.yamagishi,takao.kobayashi}@ip.titech.ac.jp

ABSTRACT

This paper describes an adaptive training technique for hidden semi-Markov model (HSMM). The adaptive training scheme conducts normalization of speaker differences and acoustic variability in both output and state duration distributions of a canonical model by using HSMM-based MLLR adaptation. We incorporate the adaptive training into our HSMM-based speech synthesis system with MLLR adaptation and compare synthesized speech using the adaptive training with that using standard speaker independent training. From the results of subjective tests, we show that the adaptive training outperforms speaker independent training and also show that the speech synthesis system generates speech with better naturalness and intelligibility than the original HSMM-based speech synthesis system.

1. INTRODUCTION

Recently *maximum likelihood linear regression* (MLLR) adaptation [1] has been used widely as an effective speaker adaptation technique that tunes speaker independent models to a new speaker by using a small amount of new speaker's speech data. Furthermore *adaptive training* schemes [2–5] have become the focus of attention as powerful training techniques reducing influence of speaker differences and acoustic variability of a *canonical* model, i.e., an initial seed model of the adaptation.

In an HMM-based speech synthesis system proposed in [6], we showed that the MLLR adaptation can approximate voice characteristics of synthetic speech to those of a target speaker by using a small amount of adaptation data uttered by the target speaker. We also showed that *speaker adaptive training* (SAT) technique [2] can significantly improve the quality of the canonical model[1] and the quality of the synthetic speech after the adaptation. In [6], the adaptive training was conducted for normalizing only state output probability distributions which represent spectrum and F_0 parameters of speech data. However, phone/segmental duration parameters of speech data also have the significant differences among the training speakers and it would occur that the state duration distributions of the canonical model have relatively large dependence on speakers and/or gender included in the training speech database. To obtain higher performance in the speaker adaptation to a wide variety of target speakers, the state duration distributions as well as the output distributions of the canonical model should not have any dependence on speaker and/or gender.

In this paper, we propose an adaptive training technique for normalizing simultaneously spectrum, F_0, and duration parameters in a framework of hidden semi-Markov model (HSMM) [7–9]. The HSMM is a kind of HMM with explicit state duration probability distributions instead of self-transition probabilities, and the HSMM-based adaptive training conducts normalization of speaker differences and acoustic variability in both output and state duration distributions of the canonical model by using HSMM-based MLLR adaptation technique [10]. We incorporate the adaptive training into our HSMM-based speech synthesis system [11] with the MLLR adaptation and show its effectiveness from results of subjective evaluation tests.

[1] In [6], we referred to the canonical model as the *average voice* model.

2. SPEAKER INDEPENDENT TRAINING OF HIDDEN SEMI-MARKOV MODEL

Before deriving HSMM-based adaptive training, we briefly review speaker independent training of hidden semi-Markov model [7–9]. An N-state HSMM λ is specified by initial state probability $\{\pi_i\}_{i=1}^N$, state transition probability $\{a_{ij}\}_{i,j=1,i\neq j}^N$, state output probability distribution $\{b_i(\cdot)\}_{i=1}^N$, and state duration probability distribution $\{p_i(\cdot)\}_{i=1}^N$. In this study we assume that the i-th state output and duration distributions are Gaussian distributions characterized by mean vector $\boldsymbol{\mu}_i$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i$, and mean m_i and variance σ_i^2, respectively,

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

$$p_i(d) = \mathcal{N}(d; m_i, \sigma_i^2). \tag{2}$$

The observation probability of training data $\boldsymbol{O} = \{\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T\}$ of length T given the model λ can be written as

$$P(\boldsymbol{O}|\lambda) = \sum_{i=1}^{N} \sum_{d=1}^{t} \gamma_t^d(i) \qquad {}^\forall\, t \in [1, T] \tag{3}$$

where $\gamma_t^d(i)$ is a probability generating serial observation sequence $\boldsymbol{o}_{t-d+1}, \cdots, \boldsymbol{o}_t$ at i-th state defined by

$$\gamma_t^d(i) = \sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_{t-d}(j) a_{ji} p_i(d) \prod_{s=t-d+1}^{t} b_i(\boldsymbol{o}_s) \beta_t(i). \tag{4}$$

In this equation, $\alpha_t(i)$ and $\beta_t(i)$ are forward and backward probabilities defined by

$$\alpha_t(i) = \sum_{d=1}^{t} \sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_{t-d}(j) a_{ji} p_i(d) \prod_{s=t-d+1}^{t} b_i(\boldsymbol{o}_s) \tag{5}$$

$$\beta_t(i) = \sum_{d=1}^{T-t} \sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij} p_j(d) \prod_{s=t+1}^{t+d} b_j(\boldsymbol{o}_s) \beta_{t+d}(j) \tag{6}$$

where a_{ji} is the transition probability from state j to i, $\alpha_0(i) = \pi_i$, and $\beta_T(i) = 1$. In the following, for simplification, we assume

that HSMM is a simple left-to-right model without skip paths. As a result, the parameter set of HSMM λ can be simplified as $\lambda = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, m, \sigma^2)$.

The conventional speaker independent training based on maximum likelihood (ML) criterion can be formulated as follows:

$$\lambda_{SI} = \underset{\lambda}{\operatorname{argmax}}\, P(\boldsymbol{O}|\lambda). \tag{7}$$

The re-estimation formulas based on Baum-Welch algorithm of the parameter set λ are given by

$$\overline{\boldsymbol{\mu}}_i = \frac{\sum_{t,d}^{T,t} \gamma_t^d(i) \sum_{s=t-d+1}^{t} \boldsymbol{o}_s}{\sum_{t,d}^{T,t} \gamma_t^d(i)\ d} \tag{8}$$

$$\overline{\boldsymbol{\Sigma}}_i = \frac{\sum_{t,d}^{T,t} \gamma_t^d(i) \sum_{s=t-d+1}^{t} (\boldsymbol{o}_s - \overline{\boldsymbol{\mu}}_i)(\boldsymbol{o}_s - \overline{\boldsymbol{\mu}}_i)^\top}{\sum_{t,d}^{T,t} \gamma_t^d(i)\ d} \tag{9}$$

$$\overline{m}_i = \frac{\sum_{t,d}^{T,t} \gamma_t^d(i) \cdot d}{\sum_{t,d}^{T,t} \gamma_t^d(i)} \tag{10}$$

$$\overline{\sigma}_i^2 = \frac{\sum_{t,d}^{T,t} \gamma_t^d(i) \cdot (d - \overline{m}_i)^2}{\sum_{t,d}^{T,t} \gamma_t^d(i)}. \tag{11}$$

3. SPEAKER ADAPTIVE TRAINING FOR HIDDEN SEMI-MARKOV MODEL

Next we derive an HSMM-based speaker adaptive training (SAT) algorithm. The basic idea of SAT algorithm is to use MLLR transformations for representing the acoustic differences among the training speakers and to train a canonical model by using the set of the MLLR transformations. The HSMM-based SAT algorithm makes use of HSMM-based MLLR algorithm [10] in which mean vectors of state output and duration distributions for speaker f are obtained by linearly transforming mean vector of state output and duration distributions of the canonical model,

$$\begin{aligned} b_i(\boldsymbol{o}^{(f)}) &= \mathcal{N}(\boldsymbol{o}; \boldsymbol{W}^{(f)}\boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i) \\ &= \mathcal{N}(\boldsymbol{o}; \boldsymbol{\zeta}^{(f)}\boldsymbol{\mu}_i + \boldsymbol{\epsilon}^{(f)}, \boldsymbol{\Sigma}_i) \end{aligned} \tag{12}$$

$$\begin{aligned} p_i(d) &= \mathcal{N}(d; \boldsymbol{X}^{(f)}\boldsymbol{\phi}_i, \sigma_i^2) \\ &= \mathcal{N}(d; \chi^{(f)} m_i + \nu^{(f)}, \sigma_i^2) \end{aligned} \tag{13}$$

where $\boldsymbol{W}^{(f)} = \left[\boldsymbol{\zeta}^{(f)}, \boldsymbol{\epsilon}^{(f)}\right]$ and $\boldsymbol{X}^{(f)} = \left[\chi^{(f)}, \nu^{(f)}\right]$ are $n \times (n+1)$ and 1×2 transformation matrices of speaker f for state output and duration distributions, respectively, $\boldsymbol{\zeta}^{(f)}$ and $\boldsymbol{\epsilon}^{(f)}$ are $n \times n$ matrix and n-dimensional vector, respectively, and $\boldsymbol{\xi}_i = [\boldsymbol{\mu}_i^\top, 1]^\top$ and $\boldsymbol{\phi}_i = [m_i, 1]^\top$ are $(n+1)$-dimensional and 2-dimensional vectors. Let F be the total number of the training speakers, $\boldsymbol{O} = \{\boldsymbol{O}^{(1)}, \cdots, \boldsymbol{O}^{(F)}\}$ be all training data, and $\boldsymbol{O}^{(f)} = \{\boldsymbol{o}_{1_f}, \cdots, \boldsymbol{o}_{T_f}\}$ be the training data of length T_f for speaker f. Speaker adaptive training based on ML criterion can be formulated as follows[2]:

$$\begin{aligned} (\lambda_{SAT}, \Lambda_{SAT}) &= \underset{\lambda,\Lambda}{\operatorname{argmax}}\, P(\boldsymbol{O}|\lambda, \Lambda) \\ &= \underset{\lambda,\Lambda}{\operatorname{argmax}} \prod_{f=1}^{F} P(\boldsymbol{O}^{(f)}|\lambda, \Lambda^{(f)}) \end{aligned} \tag{14}$$

where $\Lambda = (\Lambda^{(1)}, \cdots, \Lambda^{(F)})$ and $\Lambda^{(f)} = (\boldsymbol{W}^{(f)}, \boldsymbol{X}^{(f)})$. In other words, in the SAT paradigm, the optimum parameter set of λ and the transformation matrices Λ are estimated jointly so as to maximize the likelihood (14). The re-estimation formulas based on Baum-Welch algorithm of the parameter set λ are given by

$$\begin{aligned} \overline{\boldsymbol{\mu}}_i = &\left[\sum_{f,t,d}^{F,T_f,t} \gamma_t^d(i)\, d\, \overline{\boldsymbol{\zeta}}^{(f)\top} \boldsymbol{\Sigma}_i^{-1} \overline{\boldsymbol{\zeta}}^{(f)}\right]^{-1} \cdot \\ &\left[\sum_{f,t,d}^{F,T_f,t} \gamma_t^d(i)\, \overline{\boldsymbol{\zeta}}^{(f)\top} \boldsymbol{\Sigma}_i^{-1} \sum_{s=t-d+1}^{t} (\boldsymbol{o}_{s_f} - \overline{\boldsymbol{\epsilon}}^{(f)})\right] \end{aligned} \tag{15}$$

$$\overline{\boldsymbol{\Sigma}}_i = \frac{\sum_{f,t,d}^{F,T_f,t} \gamma_t^d(i) \sum_{s=t-d+1}^{t} (\boldsymbol{o}_{s_f} - \overline{\boldsymbol{\mu}}_i^{(f)})(\boldsymbol{o}_{s_f} - \overline{\boldsymbol{\mu}}_i^{(f)})^\top}{\sum_{f,t,d}^{F,T_f,t} \gamma_t^d(i)\ d} \tag{16}$$

$$\overline{m}_i = \frac{\sum_{f,t,d}^{F,T_f,t} \gamma_t^d(i)\, \overline{\chi}^{(f)} (d - \overline{\nu}^{(f)})}{\sum_{f,t,d}^{F,T_f,t} \gamma_t^d(i)\, \overline{\chi}^{(f)2}} \tag{17}$$

$$\overline{\sigma}_i^2 = \frac{\sum_{f,t,d}^{F,T_f,t} \gamma_t^d(i)\ (d - \overline{m}_i^{(f)})^2}{\sum_{f,t,d}^{F,T_f,t} \gamma_t^d(i)}, \tag{18}$$

where

$$\overline{\boldsymbol{\mu}}_i^{(f)} = \overline{\boldsymbol{\zeta}}^{(f)} \overline{\boldsymbol{\mu}}_i + \overline{\boldsymbol{\epsilon}}^{(f)} \tag{19}$$

$$\overline{m}_i^{(f)} = \overline{\chi}^{(f)} \overline{m}_i + \overline{\nu}^{(f)}. \tag{20}$$

The re-estimation formulas based on Baum-Welch algorithm of the transformation matrices $\Lambda^{(f)}$ are given by

$$\overline{\boldsymbol{w}}^{(f)\top} = \boldsymbol{G}_l{}^{-1} \boldsymbol{y}_l^\top \tag{21}$$

$$\overline{\boldsymbol{X}}^{(f)} = \left(\sum_{t,r,d}^{T_f,R,t} \frac{\gamma_t^d(r) d}{\sigma_r^2} \boldsymbol{\phi}_r^\top\right) \left(\sum_{t,r,d}^{T_f,R,t} \frac{\gamma_t^d(r)}{\sigma_r^2} \boldsymbol{\phi}_r \boldsymbol{\phi}_r^\top\right)^{-1} \tag{22}$$

where $(n+1) \times (n+1)$ matrix $\boldsymbol{G}_l$ is given by

$$\boldsymbol{G}_l = \sum_{t,r,d}^{T_f,R,t} \gamma_t^d(r)\, d\, \frac{1}{\Sigma_r(l)} \boldsymbol{\xi}_r \boldsymbol{\xi}_r^\top, \tag{23}$$

[2]It is straightforward to estimate multiple transformation matrices for each speaker using tree structures. However, for notational simplicity, we denote the transformation for each speaker as a single transform.

$\boldsymbol{w}^{(f)}$ and $\boldsymbol{y}_l$ are the l-th row vectors of $\boldsymbol{W}^{(f)}$ and $n \times (n+1)$ matrix $\boldsymbol{Y}$ written as

$$\boldsymbol{Y} = \sum_{t,r,d}^{T_f,R,t} \gamma_t^d(r)\ \boldsymbol{\Sigma}_r^{-1} \sum_{s=t-d+1}^{t} \boldsymbol{o}_s\, \boldsymbol{\xi}_r^\top, \tag{24}$$

respectively, $\Sigma_r(l)$ is the l-th diagonal element of $\boldsymbol{\Sigma}_r$, and R is the number of distributions sharing the same transformation matrix.

4. HSMM-BASED SPEECH SYNTHESIS WITH MLLR SPEAKER ADAPTATION

In this study, we use an HSMM-based speech synthesis system with the MLLR adaptation framework. The basic structure is similar to the HMM-based speech synthesis system of [6].

In the training stage, context dependent phoneme HSMMs are trained using multi-speaker speech database. Spectrum, F_0, and duration are modeled by multi-stream HSMMs in which output distributions for spectral and F_0 parts are modeled using continuous probability distribution and multi-space probability distribution [12], respectively. To model variations of spectrum, F_0, and duration, we take several phonetic and linguistic contextual factors such as phoneme identity factors, stress related factors, and locational factors into account. Then, shared decision tree based clustering technique [6] is separately applied to the spectral, F_0, and duration parts of the context dependent phoneme HSMMs. Moreover, we apply re-estimation process using the SAT algorithm described in Sect. 3 to the clustered and tied context dependent HSMMs. The resultant context dependent HSMMs are used as the canonical model of the adaptation.

In the adaptation stage, the canonical model is adapted to a target speaker using a small amount of speech data uttered by the target speaker. We use the HSMM-based MLLR algorithm [10] to adapt spectrum, F_0, and state duration at the same time.

In the synthesis stage, texts are transformed into a context dependent label sequence. In accordance with the label sequence, a sentence HSMM is constructed by concatenating context dependent HSMMs. From the sentence HSMM, spectral and F_0 parameter sequences are obtained based on ML criterion. Finally, by using MLSA filter, speech is synthesized from the generated mel-cepstral and F_0 parameter sequences.

5. EXPERIMENTS

5.1. Experimental Conditions

We used a set of phonetically balanced sentences of ATR Japanese speech database (Set B) for training data of HSMMs. 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. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of fundamental frequency, and their delta and delta-delta coefficients.

We used 5-state left-to-right HSMMs. The canonical model was trained using 1500 sentences, 300 sentences for each of five male speakers. We set a male speaker "mht" as the target speaker, and adapted the canonical model to the target speaker using 50 sentences which were included in the training sentences. In the SAT and MLLR algorithm, multiple transformation matrices were estimated for each speaker using shared-decision-trees constructed in the training stage. For comparison, we also trained a speaker independent HSMMs using the conventional speaker independent training (SIT) described in Sect. 2 and the standard decision tree based clustering [13]. The number of iterations for both SAT and SIT methods is set to 3. Furthermore, we also trained speaker dependent HSMMs using 450 sentences for the target speaker. It is noted that we ignore the duration probability $p_i(d)$ for $d > \mathcal{D}$, where $\mathcal{D}$ is a prescribed maximum duration value, to keep the computational costs in a reasonable range for all training and adaptation paradigms. In this study, we set $\mathcal{D}$ to 100 frames.

Table 1. The number of distributions after clustering.

	SIT	SAT	SD
Spec.	3323	3227	934
F_0	4557	3926	1549
Dur.	1253	1318	199

Table 1 shows the number of distributions included in the models after clustering. The entries for "SIT,""SAT," and "SD" correspond to the models obtained using the conventional SIT method, the proposed SAT method, and the speaker dependent modeling, respectively. In addition, "Spec.,""F_0," and "Dur." represent the spectrum, F_0, and state duration, respectively. We adjusted the number of distributions of the SAT model to a comparable size with that of SIT model.

5.2. Comparison of Speaker Adaptive Training and Speaker Independent Training

We first compared the naturalness and intelligibility of the synthesized speech generated from the models using SAT or SIT method by a paired comparison test. Subjects were seven persons, and presented a pair of synthesized speech samples generated from the models using SAT and SIT methods in random order and then asked which samples had better naturalness and intelligibility. For each subject, 10 test sentences were chosen at random from 53 test sentences which were contained in neither training nor adaptation data sentence set.

Figure 1 shows the preference scores. In the figure, "SAT" represents the results for synthesized speech using the proposed SAT method and "SIT" represents the results for synthesized speech using the conventional SIT method. A confidence interval of 95 % is also shown in the figure. From the figure, we can see that the proposed SAT method significantly outperforms the conventional SIT method at the 95% confidence level. This is because the HSMM-based SAT algorithm can reduce the influence of speaker dependency during re-estimation process, and as a result, more appropriate speaker adaptation of both the output and the state duration distribution was conducted than the normal SIT algorithm and inappropriate transformations were suppressed.

5.3. Comparison of MLLR Adaptation-based Speech Synthesis and Speaker Dependent Speech Synthesis

We then compared the naturalness and intelligibility of the synthesized speech generated from the model using SAT method and the target speaker's dependent model by a paired comparison test. Subjects and other experimental conditions were the same as the evaluation test described in Sect. 5.2.

Figure 2 shows the preference scores. In the figure, "SAT" represents the results for synthesized speech using the proposed SAT method and "SD" represents the result for synthesized speech using the speaker dependent model of the target speaker. It can be seen from the figure that the proposed SAT method significantly outperforms the speaker dependent model of the target speaker at the 95% confidence level. This means that the amount of the training data for the speaker dependent model is not sufficient in order to generate synthetic speech with good naturalness and intelligibility, and the proposed technique would yield a rich canonical model

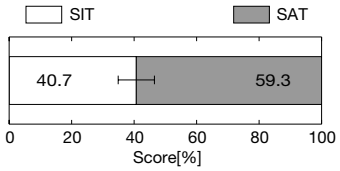

Fig. 1. Evaluation of the HSMM-based speaker adaptive training.

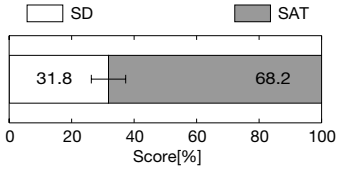

Fig. 2. Evaluation of the MLLR adaptation-based speech synthesis approach.

having a lot of training data with various contextual factors and improving the naturalness and intelligibility of synthesized speech after the speaker adaptation crucially.

5.4. Evaluation on Voice Characteristics and Prosodic Features of Synthesized Speech

We finally conducted a Comparison Category Rating (CCR) test to evaluate voice characteristics and prosodic features of synthesized speech from the model using SAT method and the target speaker's dependent model. Five persons listened to 8 sentences of synthesized speech chosen randomly from 53 test sentences and rated their voice characteristics and prosodic features comparing to those of the reference speech. The reference speech was synthesized by a mel-cepstral vocoder. The rating is a 5-point scale, that is, 5 for very similar, 4 for similar, 3 for slightly similar, 2 for dissimilar and 1 for very dissimilar. For comparison, we also evaluated *average voice* which is synthetic speech generated from the canonical model.

Figure 3 shows the results of the CCR test. "SAT" represents the results for synthesized speech using the proposed SAT method, "SD" represents the result for the speaker dependent model of the target speaker, and "AV" represents the result for the average voice. This result confirms that synthesized speech using the proposed technique has voice characteristics and prosodic features similar to the synthesized speech using the speaker dependent model. However the result also shows the proposed SAT method does not outperform the speaker dependent model of the target speaker in voice characteristics and needs further improvement. This is because all of voice characteristics and prosodic features of the target speaker is not included in a small amount of the adaptation data.

6. CONCLUSIONS

This paper has described an adaptive training technique for hidden semi-Markov model. The adaptive training conducts speaker normalization of both output and state duration distributions of a canonical model by using HSMM-based MLLR adaptation. From the results of subjective tests, we have shown that the adaptive training outperforms speaker independent training. Moreover, we have shown that the speech synthesis system using MLLR adaptation generates synthetic speech with better naturalness and intelligibility than the conventional HSMM-based speech synthesis system. Future work will focus on further evaluations of proposed technique using different speakers and development of proposed technique using other criterion such as MPE criterion [4].

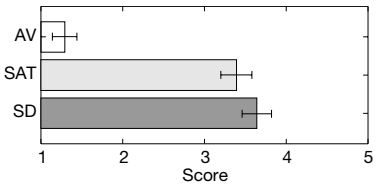

Fig. 3. Evaluation of speaker characteristics of synthesized speech.

7. ACKNOWLEDGMENTS

Authors would like to thank Prof. Keiichi Tokuda, Nagoya Institute of Technology and Dr. Takashi Masuko, Toshiba Corporation for their valuable discussions. A part of this work was supported by JSPS Grant-in-Aid for Scientific Research (B) 15300055 and JSPS Research Fellowships for Young Scientists 164633.

8. REFERENCES

[1] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[2] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP-96*, Oct. 1996, pp. 1137–1140.

[3] M.J.F. Gales, "Multiple-cluster adaptive training schemes," in *Proc. ICASSP 2001*, May 2001, pp. 361–364.

[4] L. Wang and P.C. Woodland, "Discriminative adaptive training using the MPE criterion," in *Proc. ASRU '03*, Nov. 2003, pp. 279–284.

[5] K. Yu and M.J.F. Gales, "Adaptive training using structured transforms," in *Proc. ICASSP 2004*, May 2004, pp. 317–320.

[6] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for hmm-based speech synthesis," *IEICE Trans. Fundamentals*, vol. E86-A, no. 8, pp. 1956–1963, Aug. 2003.

[7] J.D. Ferguson, "Variable duration models for speech," in *Symp. on the Application of Hidden Markov Models to Text and Speech*, 1980, pp. 143–179.

[8] S.E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, no. 1, pp. 29–45, 1986.

[9] M.J. Russell and R.K. Moore, "Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition," in *Proc. ICASSP-85*, Mar. 1985, pp. 5–8.

[10] J. Yamagishi, T. Masuko, and T. Kobayashi, "MLLR adaptation for hidden semi-Markov model based speech synthesis," in *Proc. ICSLP 2004*, Oct. 2004, (to appear).

[11] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. ICSLP 2004*, Oct. 2004, (to appear).

[12] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. ICASSP-99*, Mar. 1999, pp. 229–232.

[13] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Language Technology Workshop*, Mar. 1994, pp. 307–312.