SPEECH PARAMETER GENERATION ALGORITHMS FOR HMM-BASED SPEECH SYNTHESIS

Keiichi Tokuda [1], *Takayoshi Yoshimura* [1], *Takashi Masuko* [2], *Takao Kobayashi* [2], *Tadashi Kitamura* [1],

[1]Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555 Japan

[2]Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan

Email: {tokuda,yossie,kitamura}@ics.nitech.ac.jp, {masuko,tkobayas}@ip.titech.ac.jp

ABSTRACT

This paper derives a speech parameter generation algorithm for HMM-based speech synthesis, in which speech parameter sequence is generated from HMMs whose observation vector consists of spectral parameter vector and its dynamic feature vectors. In the algorithm, we assume that the state sequence (state and mixture sequence for the multi-mixture case) or a part of the state sequence is unobservable (i.e., hidden or latent). As a result, the algorithm iterates the forward-backward algorithm and the parameter generation algorithm for the case where state sequence is given. Experimental results show that by using the algorithm, we can reproduce clear formant structure from multi-mixture HMMs as compared with that produced from single-mixture HMMs.

1. INTRODUCTION

The increasing availability of large speech databases makes it possible to construct speech synthesis systems, which are referred to as data-driven or corpus-based approach, by applying statistical learning algorithms. These systems, which can be automatically trained, not only generate natural and high quality synthetic speech but also can reproduce voice characteristics of the original speaker.

For constructing such a system, the use of hidden Markov models (HMMs) has arisen largely. HMMs have successfully been applied to modeling the sequence of speech spectra in speech recognition systems, and the performance of HMM-based speech recognition systems have been improved by techniques which utilize the flexibility of HMMs: context-dependent modeling, dynamic feature parameters, mixtures of Gaussian densities, tying mechanism, speaker and environment adaptation techniques. HMM-based approaches to speech synthesis can be categorized as follows:

1. Transcription and segmentation of speech database [1].
2. Construction of inventory of speech segments [2], [3].
3. Run-time selection of multiple instances of speech segments [4], [5].
4. Speech synthesis from HMMs themselves [6]–[9].

In approaches 1–3, by using a waveform concatenation algorithm, e.g., PSOLA algorithm, a high quality synthetic speech could be produced. However, to obtain various voice characteristics, large amounts of speech data are necessary, and it is difficult to collect, segment, and store these data. On the other hand, in approach 4, voice characteristics of synthetic speech can be changed by transforming HMM parameters appropriately. From this point of view, we have proposed parameter generation algorithms [10], [11] for HMM-based speech synthesis, and constructed a speech synthesis system [8], [9]. Actually, we have shown that we can change voice characteristics of synthetic speech by applying a speaker adaptation technique [12], [13] or a speaker interpolation technique [14]. The main feature of the system is the use of dynamic feature: by inclusion of dynamic coefficients in the feature vector, the dynamic coefficients of the speech parameter sequence generated in synthesis are constrained to be realistic, as defined by the parameters of the HMMs.

This paper describes algorithms for speech parameter generation from HMMs. In the previously proposed algorithms [10], [11] for speech parameter generation, state sequence is assumed to be given, or determined based on a maximum likelihood criterion. In this paper, we derive a new algorithm in which we assume that the state sequence (state and mixture sequence for the multi-mixture case) or a part of the state sequence is unobservable (i.e., hidden or latent), and show examples of speech spectra generated by the proposed algorithm.

The rest of this paper is organized as follows. Section 2 summarizes the previously proposed algorithms for speech parameter generation, and derives a new algorithm. Experimental results are shown in Section 3. Concluding remarks are presented in the final section.

2. SPEECH PARAMETER GENERATION BASED ON MAXIMUM LIKELIHOOD CRITERION

For a given continuous mixture HMM λ, we derive an algorithm for determining speech parameter vector sequence

$$\boldsymbol{O} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \ldots, \boldsymbol{o}_T^\top\right]^\top \tag{1}$$

in such a way that

$$P(\boldsymbol{O}|\lambda) = \sum_{\text{all } \boldsymbol{Q}} P(\boldsymbol{O}, \boldsymbol{Q}|\lambda) \tag{2}$$

is maximized with respect to $\boldsymbol{O}$, where

$$\boldsymbol{Q} = \{(q_1, i_1), (q_2, i_2), \ldots, (q_T, i_T)\} \tag{3}$$

is the state and mixture sequence, i.e., (q, i) indicates the i-th mixture of state q. We assume that the speech parameter vector $\boldsymbol{o}_t$ consists of the static feature vector $\boldsymbol{c}_t = [c_t(1), c_t(2), \ldots, c_t(M)]^\top$

This work was partially supported by the Ministry of Education, Science, Sports and Culture Japan, Grant-in-Aid for Scientific Research (B) 2, 1055125, 1998, Encouragement of Young Scientists (A), 10780226, 1998.

(e.g., cepstral coefficients) and dynamic feature vectors $\Delta \boldsymbol{c}_t$, $\Delta^2 \boldsymbol{c}_t$ (e.g., delta and delta-delta cepstral coefficients, respectively), that is, $\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta \boldsymbol{c}_t^\top, \Delta^2 \boldsymbol{c}_t^\top]^\top$, where the dynamic feature vectors are calculated by

$$\Delta \boldsymbol{c}_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau) \boldsymbol{c}_{t+\tau} \tag{4}$$

$$\Delta^2 \boldsymbol{c}_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} w^{(2)}(\tau) \boldsymbol{c}_{t+\tau}. \tag{5}$$

We have derived algorithms [10], [11] for solving the following problems:

Case 1. For given λ and $\boldsymbol{Q}$, maximize $P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$ with respect to $\boldsymbol{O}$ under the conditions (4), (5).

Case 2. For a given λ, maximize $P(\boldsymbol{O}, \boldsymbol{Q}|\lambda)$ with respect to $\boldsymbol{Q}$ and $\boldsymbol{O}$ under the conditions (4), (5).

In this section, we will review the above algorithms and derive an algorithm for the problem:

Case 3. For a given λ, maximize $P(\boldsymbol{O}|\lambda)$ with respect to $\boldsymbol{O}$ under the conditions (4), (5).

2.1. Case 1: Maximizing $P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$ with respect to $\boldsymbol{O}$

First, consider maximizing $P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$ with respect to $\boldsymbol{O}$ for a fixed state and mixture sequence $\boldsymbol{Q}$. The logarithm of $P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$ can be written as

$$\log P(\boldsymbol{O}|\boldsymbol{Q}, \lambda) = -\frac{1}{2} \boldsymbol{O}^\top \boldsymbol{U}^{-1} \boldsymbol{O} + \boldsymbol{O}^\top \boldsymbol{U}^{-1} \boldsymbol{M} + K \tag{6}$$

where

$$\boldsymbol{U}^{-1} = \text{diag}\left[\boldsymbol{U}_{q_1,i_1}^{-1}, \boldsymbol{U}_{q_2,i_2}^{-1}, ..., \boldsymbol{U}_{q_T,i_T}^{-1}\right] \tag{7}$$

$$\boldsymbol{M} = \left[\boldsymbol{\mu}_{q_1,i_1}^\top, \boldsymbol{\mu}_{q_2,i_2}^\top, ..., \boldsymbol{\mu}_{q_T,i_T}^\top\right]^\top \tag{8}$$

$\boldsymbol{\mu}_{q_t,\, i_t}$ and $\boldsymbol{U}_{q_t,\, i_t}$ are the $3M \times 1$ mean vector and the $3M \times 3M$ covariance matrix, respectively, associated with i_t-th mixture of state q_t, and the constant K is independent of $\boldsymbol{O}$.

It is obvious that $P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$ is maximized when $\boldsymbol{O} = \boldsymbol{M}$ without the conditions (4), (5), that is, the speech parameter vector sequence becomes a sequence of the mean vectors. Conditions (4), (5) can be arranged in a matrix form:

$$\boldsymbol{O} = \boldsymbol{W}\boldsymbol{C} \tag{9}$$

where

$$\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_T]^\top \tag{10}$$

$$\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_T]^\top \tag{11}$$

$$\boldsymbol{w}_t = \left[\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}\right] \tag{12}$$

$$\begin{aligned} \boldsymbol{w}_t^{(n)} = \big[& \underbrace{\boldsymbol{0}_{M\times M}, \ldots, \boldsymbol{0}_{M\times M}}_{\text{1st}}, w^{(n)}(-L_-^{(n)}) \underbrace{\boldsymbol{I}_{M\times M}}_{(t-L_-^{(n)})\text{-th}}, \\ & \ldots, w^{(n)}(0) \underbrace{\boldsymbol{I}_{M\times M}}_{t\text{-th}}, \ldots, w^{(n)}(L_+^{(n)}) \underbrace{\boldsymbol{I}_{M\times M}}_{(t+L_+^{(n)})\text{-th}}, \\ & \underbrace{\boldsymbol{0}_{M\times M}, \ldots, \boldsymbol{0}_{M\times M}}_{T\text{-th}} \big]^\top, \quad n = 0, 1, 2 \end{aligned} \tag{13}$$

$L_-^{(0)} = L_+^{(0)} = 0$, and $w^{(0)}(0) = 1$. Under the condition (9), maximizing $P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$ with respect to $\boldsymbol{O}$ is equivalent to that with respect to $\boldsymbol{C}$. By setting

$$\frac{\partial \log P(\boldsymbol{W}\boldsymbol{C}|\boldsymbol{Q}, \lambda)}{\partial \boldsymbol{C}} = \boldsymbol{0}, \tag{14}$$

we obtain a set of equations

$$\boldsymbol{W}^\top \boldsymbol{U}^{-1} \boldsymbol{W} \boldsymbol{C} = \boldsymbol{W}^\top \boldsymbol{U}^{-1} \boldsymbol{M}^\top. \tag{15}$$

For direct solution of (15), we need $O(T^3 M^3)$ operations[1] because $\boldsymbol{W}^\top \boldsymbol{U}^{-1} \boldsymbol{W}$ is a $TM \times TM$ matrix. By utilizing the special structure of $\boldsymbol{W}^\top \boldsymbol{U}^{-1} \boldsymbol{W}$, (15) can be solved by the Cholesky decomposition or the QR decomposition with $O(TM^3L^2)$ operations[2], where $L = \max_{n\in\{1,2\}, s\in\{-,+\}} L_s^{(n)}$. Equation (15) can also be solved by an algorithm derived in [10], [11], which can operate in a time-recursive manner [16].

2.2. Case 2: Maximizing $P(\boldsymbol{O}, \boldsymbol{Q}|\lambda)$ with respect to $\boldsymbol{O}$ and $\boldsymbol{Q}$

This problem can be solved by evaluating $\max_{\boldsymbol{C}} P(\boldsymbol{O}, \boldsymbol{Q}|\lambda) = \max_{\boldsymbol{C}} P(\boldsymbol{O}|\boldsymbol{Q}, \lambda) P(\boldsymbol{Q}|\lambda)$ for all $\boldsymbol{Q}$. However, it is impractical because there are too many combinations of $\boldsymbol{Q}$. We have developed a fast algorithm for searching for the optimal or sub-optimal state sequence keeping $\boldsymbol{C}$ optimal in the sense that $P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$ is maximized with respect to $\boldsymbol{C}$ [10], [11].

To control temporal structure of speech parameter sequence appropriately, HMMs should incorporate state duration densities. The probability $P(\boldsymbol{O}, \boldsymbol{Q}|\lambda)$ can be written as $P(\boldsymbol{O}, \boldsymbol{Q}|\lambda) = P(\boldsymbol{O}, \boldsymbol{i}|\boldsymbol{q}, \lambda) P(\boldsymbol{q}|\lambda)$, where $\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$, $\boldsymbol{i} = \{i_1, i_2, \ldots, i_T\}$, and the state duration probability $P(\boldsymbol{q}|\lambda)$ is given by

$$\log P(\boldsymbol{q}|\lambda) = \sum_{n=1}^{N} \log p_{q_n}(d_{q_n}) \tag{16}$$

where the total number of states which have been visited during T frames is N, and $p_{q_n}(d_{q_n})$ is the probability of d_{q_n} consecutive observations in state q_n. If we determine the state sequence $\boldsymbol{q}$ only by $P(\boldsymbol{q}|\lambda)$ independently of $\boldsymbol{O}$, maximizing $P(\boldsymbol{O}, \boldsymbol{Q}|\lambda) = P(\boldsymbol{O}, \boldsymbol{i}|\boldsymbol{q}, \lambda) P(\boldsymbol{q}|\lambda)$ with respect to $\boldsymbol{O}$ and $\boldsymbol{Q}$ is equivalent to maximizing $P(\boldsymbol{O}, \boldsymbol{i}|\boldsymbol{q}, \lambda)$ with respect to $\boldsymbol{O}$ and $\boldsymbol{i}$. Furthermore, if we assume that state output probabilities are single-Gaussian, $\boldsymbol{i}$ is unique. Therefore, the solution is obtained by solving (15) in the same way as the Case 1.

2.3. Case 3: Maximizing $P(\boldsymbol{O}|\lambda)$ with respect to $\boldsymbol{O}$

We derive an algorithm based on an EM algorithm, which find a critical point of the likelihood function $P(\boldsymbol{O}|\lambda)$. An auxiliary function of current parameter vector sequence $\boldsymbol{O}$ and new parameter vector sequence $\boldsymbol{O}'$ is defined by

$$Q(\boldsymbol{O}, \boldsymbol{O}') = \sum_{\text{all } \boldsymbol{Q}} P(\boldsymbol{O}, \boldsymbol{Q}|\lambda) \log P(\boldsymbol{O}', \boldsymbol{Q}|\lambda). \tag{17}$$

[1] When $\boldsymbol{U}_{q,i}$ is diagonal, it is reduced to $O(T^3 M)$ since each of the M-dimensions can be calculated independently.

[2] When $\boldsymbol{U}_{q,i}$ is diagonal, it is reduced to $O(TML^2)$. Furthermore, when $L_-^{(1)} = -1$, $L_+^{(1)} = 0$, and $w^{(2)}(i) \equiv 0$, it is reduced to $O(TM)$ as described in [15].

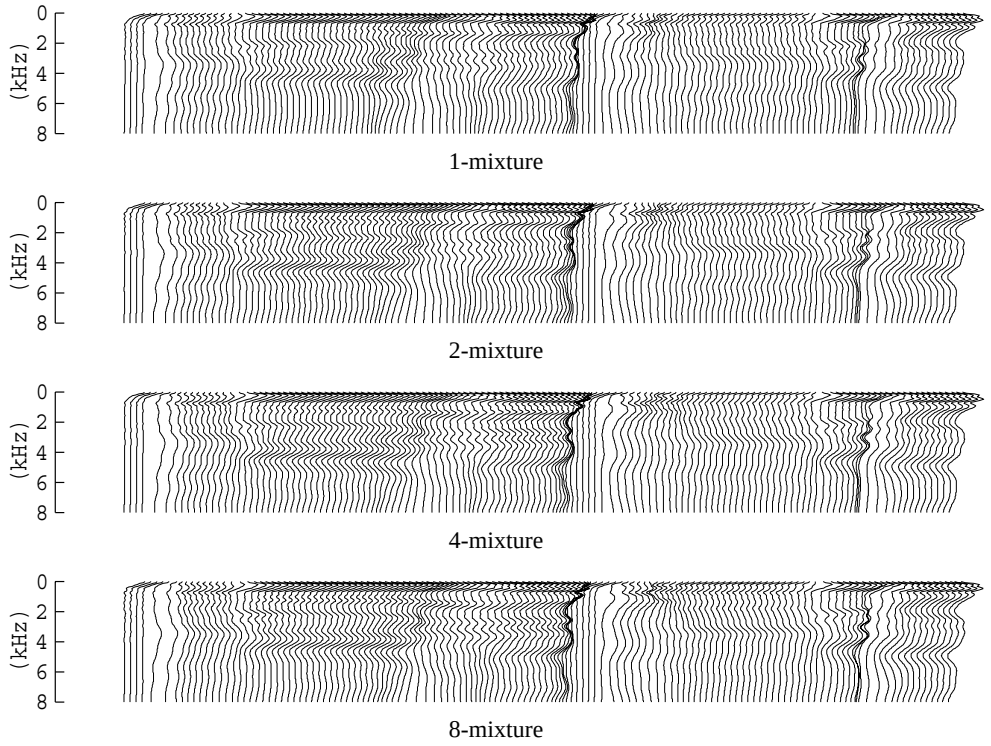

Figure 1: Generated spectra for a sentence fragment "kiNzokuhiroo."

It can be shown that by substituting $\boldsymbol{O}'$ which maximizes $Q(\boldsymbol{O}, \boldsymbol{O}')$ for $\boldsymbol{O}$, the likelihood increases unless $\boldsymbol{O}$ is a critical point of the likelihood. Equation (17) can be written as

$$Q(\boldsymbol{O}, \boldsymbol{O}') = P(\boldsymbol{O}|\lambda) \left\{ -\frac{1}{2} \boldsymbol{O}'^{\top} \overline{\boldsymbol{U}^{-1}} \boldsymbol{O}' + \boldsymbol{O}'^{\top} \overline{\boldsymbol{U}^{-1} \boldsymbol{M}} + \overline{K} \right\} \tag{18}$$

where

$$\overline{\boldsymbol{U}^{-1}} = \text{diag}\left[\overline{\boldsymbol{U}_1^{-1}}, \overline{\boldsymbol{U}_2^{-1}}, \ldots, \overline{\boldsymbol{U}_T^{-1}} \right] \tag{19}$$

$$\overline{\boldsymbol{U}_t^{-1}} = \sum_{q,i} \gamma_t(q,i) \boldsymbol{U}_{q,i}^{-1} \tag{20}$$

$$\overline{\boldsymbol{U}^{-1} \boldsymbol{M}} = \left[\overline{\boldsymbol{U}_1^{-1} \boldsymbol{\mu}_1}^{\top}, \overline{\boldsymbol{U}_2^{-1} \boldsymbol{\mu}_2}^{\top}, \ldots, \overline{\boldsymbol{U}_T^{-1} \boldsymbol{\mu}_T}^{\top} \right]^{\top} \tag{21}$$

$$\overline{\boldsymbol{U}_t^{-1} \boldsymbol{\mu}_t} = \sum_{q,i} \gamma_t(q,i) \, \boldsymbol{U}_{q,i}^{-1} \boldsymbol{\mu}_{q,i} \tag{22}$$

and the constant $\overline{K}$ is independent of $\boldsymbol{O}'$. The occupancy probability $\gamma_t(q,i)$ defined by

$$\gamma_t(q,i) = P(q_t = (q,i)|\boldsymbol{O}, \lambda) \tag{23}$$

can be calculated with the forward-backward inductive procedure. Under the condition $\boldsymbol{O}' = \boldsymbol{W}\boldsymbol{C}'$, $\boldsymbol{C}'$ which maximizes $Q(\boldsymbol{O}, \boldsymbol{O}')$ is given by the following set of equations:

$$\boldsymbol{W}^{\top} \overline{\boldsymbol{U}^{-1}} \boldsymbol{W} \boldsymbol{C}' = \boldsymbol{W}^{\top} \overline{\boldsymbol{U}^{-1} \boldsymbol{M}}. \tag{24}$$

The above set of equations has the same form as (15). Accordingly, it can be solved by the algorithm for solving (15).

The whole procedure is summarized as follows:

Step 0. Choose an initial parameter vector sequence $\boldsymbol{C}$.

Step 1. Calculate $\gamma_t(q,i)$ with the forward-backward algorithm.

Step 2. Calculate $\overline{\boldsymbol{U}^{-1}}$ and $\overline{\boldsymbol{U}^{-1} \boldsymbol{M}}$ by (19)–(22), and solve (24).

Step 3. Set $\boldsymbol{C} = \boldsymbol{C}'$. If a certain convergence condition is satisfied, stop; otherwise, goto Step 1.

From the same reason as Case 2, HMMs should incorporate state duration densities. If we determine the state sequence $\boldsymbol{q}$ only by $P(\boldsymbol{q}|\lambda)$ independently of $\boldsymbol{O}$ in a manner similar to the previous section, only the mixture sequence $\boldsymbol{i}$ is assumed to be unobservable[3]. Further, we can also assume that $\boldsymbol{Q}$ is unobservable but phoneme or syllable durations are given.

3. EXAMPLE

We used phonetically balanced 450 sentences from ATR Japanese speech database for training. Speech signal were sampled at 16 kHz and windowed by a 25.6-ms Blackman window with a 5-ms shift, and then mel-cepstral coefficients were obtained by a mel-cepstral analysis technique [18]. Feature vector consists of 25 mel-cepstral coefficients including the zeroth coefficient, their delta and delta-delta coefficients. We used 5-state left-to-right HMMs. Decision-tree based state clustering [19] was applied to the context-dependent phoneme model set, and the resultant HMM set has approximately 900 states. We assume that the state and mixture sequence is unobservable but phoneme durations are given from phoneme duration densities in a manner similar to [20]. It has found that a few iterations are sufficient for convergence of the

[3]For this problem, an algorithm based on a direct search has also been proposed in [17].

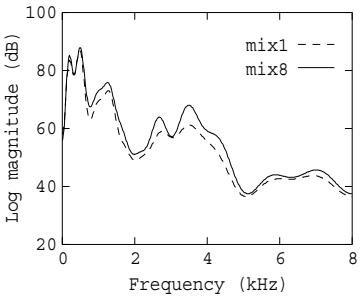

Figure 2: Spectra obtained from 1-mixture HMMs and 8-mixture HMMs.

proposed algorithm. Fig. 1 shows generated spectra for a Japanese sentence fragment "kiNzokuhiroo" taken from a sentence which is not included in the training data. Fig. 2 compares two spectra obtained from single-mixure HMMs and 8-mixture HMMs, respectively, for the same temporal position of the sentence fragment.

It is seen from Fig. 1 and Fig. 2 that with increasing mixtures, the formant structure of the generated spectra get clearer. From informal listening of the synthetic speech, it has been observed that the quality of the synthetic speech is considerably improved by increasing mixtures.

When we use single-mixture HMMs, the formant structure of spectrum corresponding to each mean vector $\boldsymbol{\mu}_{q,i}$ might be vague since $\boldsymbol{\mu}_{q,i}$ is the average of different speech spectra. One can increase the number of decision tree leaf clusters. However, it might result in perceivable discontinuities in synthetic speech since overly large tree will be overspecialized to training data and generalized poorly. We expect that the proposed algorithm can avoid this situation in a simple manner.

4. CONCLUSION

This paper has derived an algorithm for speech parameter generation from HMM whose observation vector consists of spectral parameter vector and its dynamic feature vectors. In the algorithm, we assume that the state sequence (state and mixture sequence for the multi-mixture case) or a part of the state sequence is unobservable (i.e., hidden or latent). The algorithm is based on an EM algorithm. As a result, it iterates the forward-backward algorithm and the parameter generation algorithm for the case where state sequence is given. Experimental results have shown that by using the algorithm, we can reproduce clear formant structure from multi-mixture HMMs as compared with that produced from single-mixture HMMs. Results of subjective evaluation tests of the synthetic speech will be presented in the near future.

5. REFERENCES

[1] A. Ljolje, J. Hirschberg and J. P. H. van Santen, "Automatic speech segmentation for concatenative inventory selection," in Progress in Speech Synthesis, ed. J. P. H. van Santen, R. W. Sproat, J. P. Olive and J. Hirschberg, Springer-Verlag, New York, 1997.

[2] R. E. Donovan and P. C. Woodland, "Automatic speech synthesiser parameter estimation using HMMs," in *Proc. ICASSP*, 1995, pp.640–643.

[3] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith and M. Plumpe, "Recent improvements on Microsoft's trainable text-to-speech system -Whistler," in *Proc. ICASSP*, 1997, pp.959–962.

[4] H. Hon, A. Acero, X. Huang, J. Liu and M. Plumpe, "Automatic generation of synthesis units for trainable text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp.293–306.

[5] R. E. Donovan and E. M. Eide, "The IBM Trainable Speech Synthesis System," in *Proc. ICSLP*, 1998, pp.1703–1706.

[6] A. Falaschi, M. Giustiniani and M. Verola, "A hidden Markov model approach to speech synthesis," in *Proc. EUROSPEECH*, 1989, pp.187–190.

[7] M. Giustiniani and P. Pierucci, "Phonetic ergodic HMM for speech synthesis," in *Proc. EUROSPEECH*, 1991, pp.349–352.

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

[9] 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. EUROSPEECH*, 1999, pp.2347–2350.

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

[11] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," in *Proc. EUROSPEECH*, 1995, pp.757–760.

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

[13] M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *Proc. ESCA/COCOSDA Third International Workshop on Speech Synthesis*, 1998, pp.273–276.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Speaker Interpolation in HMM-Based Speech Synthesis System," in *Proc. EUROSPEECH*, 1997, pp.2523–2526.

[15] A. Acero, "Formant analysis and synthesis using hidden Markov models," in *Proc. EUROSPEECH*, 1999, pp.1047–1050.

[16] K. Koishida, K. Tokuda, T. Masuko and T. Kobayashi, "Vector quantization of speech spectral parameters using statistics of dynamic features," in *Proc. ICSP*, 1997, pp.247–252.

[17] W. Tachiwa and S. Furui, "A study of speech synthesis using HMMs," in *Proc. Spring Meeting of Acoustical Society of Japan*, Mar. 1999, pp.239–240 (in Japanese).

[18] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, pp.137–140.

[19] J. J. Odell, "The use of context in large vocabulary speech recognition," PhD thesis, Cambridge University, 1995.

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