Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis

Tomoki Toda† *and Keiichi Tokuda*‡

† Graduate School of Information Science, Nara Institute of Science and Techonology, Japan
‡ Graduate School of Engineering, Nagoya Institute of Technology, Japan
tomoki@is.naist.jp, tokuda@ics.nitech.ac.jp

Abstract

This paper describes a novel parameter generation algorithm for the HMM-based speech synthesis. The conventional algorithm generates a trajectory of static features that maximizes an output probability of a parameter sequence consisting of the static and dynamic features from HMMs under an actual constraint between the two features. The generated trajectory is often excessively smoothed due to the statistical processing. Using the over-smoothed trajectory causes the muffled sound. In order to alleviate the over-smoothing effect, we propose the generation algorithm considering not only the output probability used for the conventional method but also that of a global variance (GV) of the generated trajectory. The latter probability works as a penalty for a reduction of the variance of the generated trajectory. A result of a perceptual evaluation demonstrates that the proposed method causes large improvements of the naturalness of synthetic speech.

1. Introduction

It is no doubtful that the corpus-based approach [1] has caused the dramatic improvements of Text-to-Speech (TTS) [2]. It has enabled us to construct a TTS system without professional expertise. So far, many generic synthesis methods have been established.

There are two main techniques of corpus-based speech synthesis, i.e., sample-based synthesis and statistical synthesis. The sample-based synthesis such as unit selection [3] directly uses acoustic inventories selected from a speech corpus for synthesizing a speech waveform. Main advantage of this method is that high-quality speech keeping original voice characteristics is synthesized by concatenating natural acoustic units. However, since the desired units with target attributes are not always in the corpus, other units with similar attributes to the target are used instead. The concatenation of such units often causes audible discontinuities. Signal processing alleviates those discontinuities but it causes other artificial sounds. Consequently, the large-sized speech corpus with consistent voice quality is inevitable to achieve high-quality synthetic speech, which is indeed hard to be prepared. One of the biggest problems of the sample-based synthesis is difficult to flexibly synthesize speech with rich voice characteristics.

On the other hand, the statistical synthesis such as Context Oriented Clustering (COC) [4] uses averaged acoustic inventories statistically extracted from the speech corpus. Synthetic speech based on those inventories has smooth and consistent quality. Moreover, it is more robust to the corpus size compared with the sample-based method because unseen acoustics are generated with an interpolation of the inventories having similar attributes to the target. However, voice quality of synthetic speech is muffled compared with that of natural speech because complex characteristics of speech are removed in the statistical processing. In general, this method is inferior to the sample-based method in terms of naturalness of synthetic speech.

A hidden Markov model (HMM) has widely been used in speech recognition. As one of the statistical synthesis methods, we focus on the HMM-based speech synthesis method [5][6]. This method has many advantages as follows: 1) it is well known that the HMM is suitable for modeling a time sequence of speech acoustics, 2) we can apply many techniques for HMM-based speech recognition to speech synthesis, 3) because the HMM is mathematically tractable, voice characteristics of synthetic speech are easily controlled by modifying acoustic statistics in the manner mathematically supported. The HMM-based synthesis method directly generates speech parameters from HMMs so that an output probability of the parameter is maximized under a constraint on an explicit relationship between static and dynamic features [7]. Consequently, a smoothed parameter trajectory is generated but it is excessively smoothed due to the statistical processing. Using multiple mixtures alleviates the over-smoothing effect [7] but it also causes another problem of over-training.

Recently, we proposed a voice conversion method considering a global variance (GV) of the converted trajectory for alleviating the over-smoothing effect [8]. It is shown that this method is superior to a spectral enhancement technique with the postfilter that is widely used for improving the speech quality [9]. In this paper, we apply the idea of considering GV to not only spectral parameter generation but also F_0 parameter generation in the HMM-based speech synthesis.

The paper is organized as follows. In **Section 2**, we describe the conventional parameter generation algorithm. In **Section 3**, we describe the proposed algorithm considering GV. In **Section 4**, experimental evaluations are described. Finally, we summarize this paper in **Section 5**.

2. Conventional Parameter Generation Algorithm

We assume a D-dimensional static feature vector $\boldsymbol{c}_t = [c_t(1), c_t(2), \cdots, c_t(D)]^\top$ at frame t. We use a speech parameter vector $\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta^2\boldsymbol{c}_t^\top]^\top$ consisting of not only the static feature vector but also dynamic feature vectors $\Delta\boldsymbol{c}_t$, $\Delta^2\boldsymbol{c}_t$, which are calculated by

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

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

In this paper, we treat feature vectors at all frames over an utterance as a time sequence vector. The sequence vectors of $\boldsymbol{o}_t$ and $\boldsymbol{c}_t$ are written as

$$\boldsymbol{O} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \cdots, \boldsymbol{o}_T^\top\right]^\top, \tag{3}$$

$$\boldsymbol{C} = \left[\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top, \tag{4}$$

respectively, and the relationship between those is represented as

$$\boldsymbol{O} = \boldsymbol{W}\boldsymbol{C}, \tag{5}$$

where

$$\begin{aligned}
\boldsymbol{W} &= \left[\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_T\right]^\top, && (6)\\
\boldsymbol{w}_t &= \left[\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}\right], && (7)\\
\boldsymbol{w}_t^{(n)} &= \Big[\underset{\text{1st}}{\boldsymbol{0}_{D\times D}}, \cdots, \boldsymbol{0}_{D\times D}, \underset{(t-L_-^{(n)})\text{-th}}{w^{(n)}(-L_-^{(n)})\boldsymbol{I}_{D\times D}}, \\
&\qquad \cdots, \underset{t\text{-th}}{w^{(n)}(0)\boldsymbol{I}_{D\times D}}, \cdots, \underset{(t+L_+^{(n)})\text{-th}}{w^{(n)}(-L_+^{(n)})\boldsymbol{I}_{D\times D}}, \\
&\qquad \boldsymbol{0}_{D\times D}, \cdots, \underset{T\text{-th}}{\boldsymbol{0}_{D\times D}}\Big]^\top, \quad n = 0, 1, 2 && (8)
\end{aligned}$$

$L_-^{(0)} = L_+^{(0)} = 0$, and $w^{(0)}(0) = 1$.

For a given continuous mixture HMM $\boldsymbol{\lambda}$, an output probability of the parameter vectors $\boldsymbol{O}$ is written as

$$P(\boldsymbol{O}|\boldsymbol{\lambda}) = \sum_{\text{all}\boldsymbol{Q}} P(\boldsymbol{O}, \boldsymbol{Q}|\boldsymbol{\lambda}), \tag{9}$$

where

$$\boldsymbol{Q} = \{(q_1, i_1), (q_2, i_2), \cdots, (q_T, i_T)\} \tag{10}$$

is the state and mixture sequence, i.e., (q, i) indicates the i-th mixture of state q. We determine the static feature sequence $\boldsymbol{C}$ that maximizes the output probability. In order to reduce computational cost, the current HMM-based speech synthesis system [5][6] determines the sub-optimum state sequence independently of $\boldsymbol{O}$ so that the state duration probability $P(\boldsymbol{q}|\boldsymbol{\lambda})$ is maximized, where $\boldsymbol{q} = \{q_1, q_2, \cdots, q_T\}$. Moreover, the sub-optimum mixture sequence $\boldsymbol{i} = \{i_1, i_2, \cdots, i_T\}$ is also determined when a single Gaussian is used as each state output probability. Under such conditions, we maximize the following log-scaled output probability with respect to $\boldsymbol{C}$,

$$\log P(\boldsymbol{O}|\boldsymbol{Q}, \boldsymbol{\lambda}) = -\frac{1}{2}\boldsymbol{O}^\top\boldsymbol{U}^{-1}\boldsymbol{O} + \boldsymbol{O}^\top\boldsymbol{U}^{-1}\boldsymbol{M} + K, \tag{11}$$

where

$$\begin{aligned}
\boldsymbol{U}^{-1} &= \text{diag}\left[\boldsymbol{U}_{q_1,i_1}^{-1}, \boldsymbol{U}_{q_2,i_2}^{-1}, \cdots, \boldsymbol{U}_{q_T,i_T}^{-1}\right], && (12)\\
\boldsymbol{M} &= \left[\boldsymbol{\mu}_{q_1,i_1}^\top, \boldsymbol{\mu}_{q_2,i_2}^\top, \cdots, \boldsymbol{\mu}_{q_T,i_T}^\top\right]^\top, && (13)
\end{aligned}$$

$\boldsymbol{\mu}_{q_t,i_t}^\top$ and $\boldsymbol{U}_{q_t,i_t}^{-1}$ 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. The constant K is independent of $\boldsymbol{O}$. Under the condition (5), we determine $\boldsymbol{C}$ that maximizes the output probability by setting

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

Consequently, we obtain a set of equations

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

Although we assume that the sub-optimum state and mixture sequence is given in this paper, we can also determine $\boldsymbol{C}$ by directly maximizing $P(\boldsymbol{O}|\boldsymbol{\lambda})$ with EM algorithm [7].

3. Proposed Parameter Generation Algorithm Considering GV

The GV [8] of the static feature vectors is defined as

$$\begin{aligned}
\boldsymbol{v}(\boldsymbol{C}) &= \left[v(1), v(2), \cdots, v(D)\right]^\top, && (16)\\
v(d) &= \frac{1}{T}\sum_{t=1}^{T}\left(c_t(d) - \overline{c}(d)\right)^2, && (17)\\
\overline{c}(d) &= \frac{1}{T}\sum_{\tau=1}^{T} c_\tau(d). && (18)
\end{aligned}$$

The proposed method determines the static feature sequence considering not only the output probability of the static and dynamic feature vectors but also that of the GV. Specifically, instead of maximizing the probability (11), we maximize the following criterion, which is based on a product of the two output probabilities, with respect to the static feature sequence $\boldsymbol{C}$,

$$L = \log\left\{p(\boldsymbol{O}|\boldsymbol{Q}, \boldsymbol{\lambda})^\omega \cdot p(\boldsymbol{v}(\boldsymbol{C})|\boldsymbol{\lambda}_v)\right\}, \tag{19}$$

where $p(\boldsymbol{v}(\boldsymbol{C})|\boldsymbol{\lambda}_v)$ is modeled by a single Gaussian distribution. A set of model parameters $\boldsymbol{\lambda}_v$ consists of the mean vector $\boldsymbol{\mu}_v$ and the covariance matrix $\boldsymbol{\Sigma}_v = \boldsymbol{P}_v^{-1}$ for the GV. This Gaussian model $\boldsymbol{\lambda}_v$ and the HMMs $\boldsymbol{\lambda}$ are independently trained from the speech corpus. The constant ω denotes the weight controlling a balance between the two probabilities. In this paper, ω is set to the ratio of the number of dimensions between vectors $\boldsymbol{v}(\boldsymbol{C})$ and $\boldsymbol{O}$, i.e., $1/(3T)$.

In order to determine $\boldsymbol{C}$ that maximizes L, we iteratively update $\boldsymbol{C}$ with the gradient method,

$$\boldsymbol{C}^{(i+1)\text{-th}} = \boldsymbol{C}^{(i)\text{-th}} + \alpha \cdot \boldsymbol{\Delta}\boldsymbol{C}^{(i)\text{-th}}, \tag{20}$$

where α is a step size parameter. If we use the steepest decent algorithm using only the first derivative written as

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{C}} &= \omega\left(-\boldsymbol{W}^\top\boldsymbol{U}^{-1}\boldsymbol{W}\boldsymbol{C} + \boldsymbol{W}^\top\boldsymbol{U}^{-1}\boldsymbol{M}\right)\\
&\qquad + \left[\boldsymbol{v}_1'^\top, \boldsymbol{v}_2'^\top, \cdots, \boldsymbol{v}_T'^\top\right]^\top, && (21)\\
\boldsymbol{v}_t' &= \left[v_t'(1), v_t'(2), \cdots, v_t'(D)\right]^\top, && (22)\\
v_t'(d) &= -\frac{2}{T}\boldsymbol{p}_v^{(d)\top}\left(\boldsymbol{v}(\boldsymbol{C}) - \boldsymbol{\mu}_v\right)\left(c_t(d) - \overline{c}(d)\right), && (23)
\end{aligned}$$

where $\boldsymbol{p}_v^{(d)}$ is the d-th vector of $\boldsymbol{P}_v$, the vector $\boldsymbol{\Delta}\boldsymbol{C}^{(i)\text{-th}}$ is written as

$$\boldsymbol{\Delta}\boldsymbol{C}^{(i)\text{-th}} = \left.\frac{\partial L}{\partial \boldsymbol{C}}\right|_{\boldsymbol{C}=\boldsymbol{C}^{(i)\text{-th}}}. \tag{24}$$

Moreover, we may also use the Newton-Raphson method using not only the first derivative but also the second derivative written as

$$\frac{\partial^2 L}{\partial \boldsymbol{C} \partial \boldsymbol{C}^\top} = -\omega \boldsymbol{W}^\top \boldsymbol{U}^{-1} \boldsymbol{W} + \begin{bmatrix} \boldsymbol{v}''_{1,1} & \cdots & \boldsymbol{v}''_{1,T} \\ \vdots & \ddots & \vdots \\ \boldsymbol{v}''_{T,1} & \cdots & \boldsymbol{v}''_{T,T} \end{bmatrix}, \quad (25)$$

$$\boldsymbol{v}''_{t_1,t_2} = \begin{bmatrix} {v''}^{(1,1)}_{t_1,t_1} & \cdots & {v''}^{(1,D)}_{t_1,t_2} \\ \vdots & \ddots & \vdots \\ {v''}^{(D,1)}_{t_1,t_2} & \cdots & {v''}^{(D,D)}_{t_1,t_2} \end{bmatrix}, \quad (26)$$

$${v''}^{(d_1,d_2)}_{t_1,t_2} = -\frac{2}{T^2} \left\{ \beta \cdot \boldsymbol{p}_v^{(d_1)\top} (\boldsymbol{v}(\boldsymbol{C}) - \boldsymbol{\mu}_v) + 2 z^{(d_1,d_2)}_{t_1,t_2} \right\}, (27)$$

$$z^{(d_1,d_2)}_{t_1,t_2} = p_v^{(d_1,d_2)} \{c_{t_1}(d_1) - \overline{c}(d_1)\}\{c_{t_2}(d_2) - \overline{c}(d_2)\}, \quad (28)$$

where

$$\beta = \begin{cases} \text{if } (d_1 \neq d_2), & 0, \\ \text{else if } (t_1 \neq t_2), & -1, \\ \text{else}, & T-1. \end{cases} \quad (29)$$

The vector $\boldsymbol{\Delta C}^{(i)\text{-th}}$ is written as

$$\boldsymbol{\Delta C}^{(i)\text{-th}} = -\left(\frac{\partial^2 L}{\partial \boldsymbol{C} \partial \boldsymbol{C}^\top} \right)^{-1} \left. \frac{\partial L}{\partial \boldsymbol{C}} \right|_{\boldsymbol{C} = \boldsymbol{C}^{(i)\text{-th}}}. \quad (30)$$

Large computational cost is necessary to calculate the second derivative, i.e., the Hessian matrix. Furthermore, it is not always a positive definite matrix. In this paper, we approximate the second derivative using only diagonal elements.

A concept of the proposed method is that the parameter generation is performed under a constraint on the variance of the generated parameter trajectory, i.e., GV. In the criterion (19), we can consider the probability $p(\boldsymbol{v}(\boldsymbol{C}))$ as a penalty term for a reduction of the GV.

4. Experimental Evaluations

4.1. Experimental conditions

We trained HMMs for each of 4 Japanese speakers (2 males, MHT and MYI, and 2 females, FTK and FYM). We used 450 sentences of phonetically balanced 503 sentences from ATR Japanese speech database B-set as training data for each speaker. Context-dependent labels were prepared from phoneme and linguistic labels included in the ATR database.

As a spectral parameter, we used 0th through 24th mel-cepstral coefficients obtained from the smoothed spectrum analyzed by STRAIGHT [10]. As a source parameter, we used a log-scaled F_0 automatically extracted from a waveform. Each of the spectral and F_0 parameter vectors included the static feature and their delta and delta-delta features. Frame shift was set to 5 ms.

A spectral part was modeled by continuous HMMs of which each state output probability was modeled by a single Gaussian distribution with a diagonal covariance matrix. As for an F_0 part, we used the HMMs based on multi-space probability distribution (MSD-HMMs) [11] to model a time sequence consisting of continuous values, i.e., log-scaled F_0s, and discrete symbols that represent unvoiced frames. Static, delta, and delta-delta F_0s were treated in different streams. We constructed context-dependent HMMs for each part with a decision-tree based context clustering technique based on an MDL criterion [12]. We also trained context-dependent HMMs for modeling the state duration probabilities.

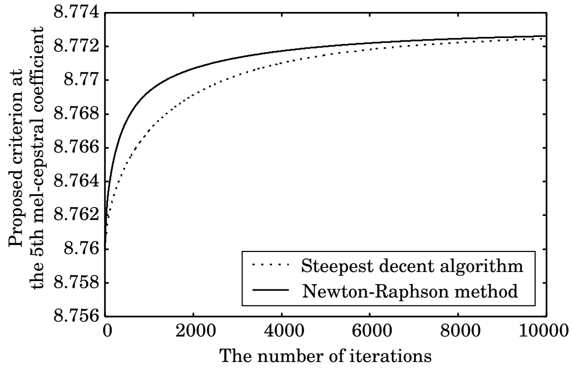

Figure 1: *Comparison of convergence performance between two gradient methods.*

In the synthesis, we concatenated HMMs for given input contexts and then we determined a sequence of probability density functions (PDFs) in the manner as described in **Section 2**. A mel-cepstrum sequence was directly generated from PDFs. In the F_0 parameter generation, we firstly determined unvoiced frames based on the output probability of the unvoiced symbol from the MSD-HMMs. Then, we generated an F_0 parameter sequence from a PDF sequence that doesn't include the unvoiced frames. Inverse variances for the dynamic features were set to 0 at the boundaries between voiced and unvoiced frames. A simple excitation was constructed with a pulse train and noise based on the generated F_0 parameters. Then, a speech waveform was synthesized with the MLSA filter [13] based on the generated mel-cepstra.

4.2. Investigation of iterative generation process

We firstly generate a parameter trajectory with the conventional algorithm, and then we estimate the trajectory that maximizes the criterion (19) using the gradient method. We can perform these processes at each dimension because we use diagonal covariance matrices in this paper.

We investigated which trajectory is better as an initial value used for the gradient method, the generated trajectory by the conventional algorithm or the trajectory to which the generated one is converted so that its GV is equal to the mean of the GV model $\boldsymbol{\lambda}_v$. Results showed that the latter has a larger value of the criterion than the former. Therefore, we use the converted trajectory as the initial value. Note that this result depends on the weight ω in the criterion.

We also investigated which gradient method has a better performance of the convergence, the steepest decent algorithm or the Newton-Raphson method. The step size parameter was optimized for each method so that the criterion converged as fast as possible. One example of the convergence of the criterion is shown in **Figure 1**. The Newton-Raphson method has the better convergence compared with the steepest decent algorithm when using the initial value as mentioned above. We found this tendency at the most of cases. Therefore, we use the Newton-Raphson method in this paper.

4.3. Perceptual evaluation

We performed an opinion test on the naturalness of synthetic speech to demonstrate the effectiveness of the proposed method. We evaluated the following five voices: 1) synthetic speech with the spectral and F_0 parameters generated by the conventional

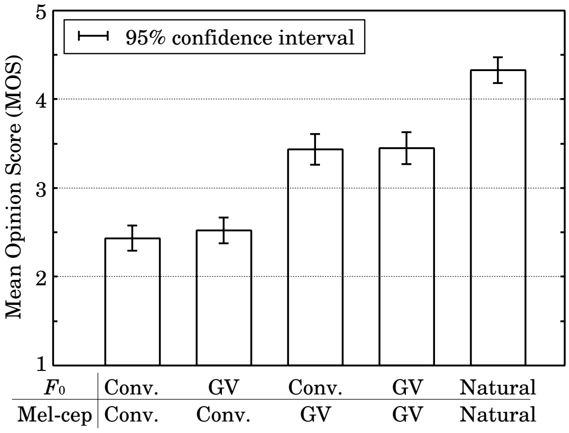

Figure 2: *Mean opinion score (MOS) for each synthetic voice. MOS is calculated over results for all of four speakers. "Conv." and "GV" denote that the conventional method and the proposed method are employed respectively for each of "F_0" and "Mel-cep" parameter generations. "Natural" denotes that parameters extracted from natural speech are used, i.e., analysis-synthesized speech.*

method, 2) synthetic speech with the spectral parameter generated by the conventional method and the F_0 parameter generated by the proposed method, 3) synthetic speech with the spectral parameter generated by the proposed method and the F_0 parameter generated by the conventional method, 4) synthetic speech with the spectral and F_0 parameters generated by the proposed method, and 5) analysis-synthesized speech. Seven Japanese listeners participated in the test. Each listener evaluated 25 samples consisting of five sentences for each speaker. Those sentences were randomly selected for each listener from 53 sentences, which were not included in the training data.

Figure 2 shows a result of the test. It is observed that the proposed method works very well in the spectral parameter generation. Considering the GV in the F_0 parameter generation slightly causes an improvement of the naturalness of synthetic speech. One of reasons why the F_0 improvements are small is possibly that the GV vectors used in the training were affected by errors of the automatic F_0 extraction, especially halving and doubling, which were often observed on the extracted F_0s. Those errors make the GV inappropriately large.

The improved quality is still worse than that of the analysis-synthesized speech. This quality difference is caused by not only the insufficient accuracy of the generated spectral and F_0 parameters but also that of duration modeling. Further improvements of the acoustic modeling are indispensable for achieving higher-quality synthetic speech.

5. Conclusions

We proposed a parameter generation algorithm considering global variance (GV) of the generated parameters for the HMM-based speech synthesis. The proposed method generated a time sequence of static features that maximized a criterion based on not only an output probability of a time sequence of the static and dynamic features but also that of the GV under a constraint that the dynamic features and the GV were calculated from the static features. We applied this algorithm to both spectral and F_0 parameter generations. As a result of the perceptual evaluation, it was shown that the proposed algorithm causes the large improvements of the naturalness of synthetic speech.

Acknowledgment: This research was supported in part by JSPS (Japan Society for the Promotion of Science) Research Fellowships for Young Scientists. The authors are grateful to Dr. Hideki Kawahara of Wakayama University in Japan for permission to use the STRAIGHT analysis-synthesis method.

6. References

[1] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. *Proc. ICASSP*, pp. 679–682, New York, USA, Apr. 1988.

[2] A.K. Syrdal, C.W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K-S. Lee, and M.J. Makashay. Corpus-based techniques in the AT&T NextGen synthesis system. *Proc. ICSLP*, Vol. 3, pp. 410–415, Beijing, China, Oct. 2000.

[3] A.J. Hunt and A.W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. *Proc. ICASSP*, pp. 373–376, Atlanta, USA, May 1996.

[4] S. Nakajima and H. Hamada. Automatic generation of synthesis units based on context oriented clustering. *Proc. ICASSP*, pp. 659–662, New York, USA, Apr. 1988.

[5] K. Tokuda, H. Zen, and A.W. Black. An HMM-based speech synthesis system applied to English. *Proc. IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, USA, Sep. 2002.

[6] 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, Budapest, Hungary, Sep. 1999.

[7] 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, Istanbul, Turkey, June 2000.

[8] T. Toda, A.W. Black, and K. Tokuda. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. *Proc. ICASSP*, Vol. 1, pp. 9–12, Philadelphia, USA, Mar. 2005.

[9] T. Yoshimura. *Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based Text-to-Speech systems*. Ph.D. Thesis, Department of Electrical and Computer Engineering, Nagoya Institute of Technology, 2001.

[10] H. Kawahara, I. Masuda-Katsuse, and A.de 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 Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[11] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. *Proc. ICASSP*, pp. 229–232, Phoenix, U.S.A., May 1999.

[12] K. Shinoda and T. Watanabe. Acoustic modeling based on the MDL criterion for speech recognition. *Proc. EUROSPEECH*, pp. 99–102, Rhodes, Greece, Sep. 1997.

[13] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, An adaptive algorithm for mel-cepstral analysis of speech. *Proc. ICASSP*, Vol. 1, pp. 137–140, San Francisco, USA, Mar. 1992.