SPECTRAL CONVERSION BASED ON MAXIMUM LIKELIHOOD ESTIMATION CONSIDERING GLOBAL VARIANCE OF CONVERTED PARAMETER

Tomoki Toda[†‡], *Alan W Black*[†], *Keiichi Tokuda*[‡]

[†]Language Technologies Institute, Carnegie Mellon University, USA
[‡]Graduate School of Engineering, Nagoya Institute of Technology, Japan
{tomoki,awb}@cs.cmu.edu, {tomoki,tokuda}@ics.nitech.ac.jp

ABSTRACT

This paper describes a novel spectral conversion method for the voice transformation. We perform spectral conversion between speakers using a Gaussian Mixture Model (GMM) on joint probability density of source and target features. A smooth spectral sequence can be estimated by applying maximum likelihood (ML) estimation using dynamic features to the GMM-based mapping. However, the degradation of the converted speech quality is still caused due to an over-smoothing of the converted spectra, which is inevitable in the conventional ML-based parameter estimation. In order to alleviate the over-smoothing, we propose an ML-based conversion taking account of the global variance of the converted parameter in each utterance. Experimental results show that the performance of the voice conversion can be improved by using the global variance information. Moreover, it is demonstrated that the proposed algorithm is more effective than spectral enhancement by postfiltering.

1. INTRODUCTION

Voice conversion is a potential technique for flexibly synthesizing various types of speech. This technique can modify speech characteristics using conversion rules statistically extracted from a small amount of training data. A typical application of voice conversion is speaker conversion [1]. This conversion can be extended to the cross-language speaker conversion [2][3]. Although we focus on the spectral conversion in this paper, prosodic conversion as well as spectral conversion is important to more properly realize speaker personality [4].

As a typical spectral conversion method, a mapping algorithm using a Gaussian Mixture Model (GMM) has been proposed by Stylianou [5]. In this method, the mapping between source and target features is determined using the GMM on joint probability density of those features [6]. In each mixture, the conditional target mean vector for the given source vector is calculated by a simple linear conversion using the correlation between the source and target features. The converted vector is defined as the weighted sum of the conditional mean vectors, where the conditional probabilities that the source vector belongs to each one of the mixtures are used as weights. Although this mapping method can improve the conversion accuracy compared with the VQ-based mapping [1], the performance of the conversion is insufficient. The converted speech quality is deteriorated by some factors, e.g., an excessive smoothing of converted spectra [7] and spectral discontinuities [8].

In this paper, we perform the spectral conversion based on the maximum likelihood (ML) criterion. In order to alleviate the spectral discontinuities, we consider the correlation between frames by applying a parameter generation algorithm with dynamic features [9] to the GMM-based mapping. This conversion algorithm makes it possible to estimate a more appropriate spectral sequence compared with the conventional GMM-based algorithm. However, the over-smoothing problem of the converted spectra still remains to be solved. In order to address this problem, we propose an ML-based conversion algorithm taking account of the global variance of the converted spectra in each utterance. The effectiveness of using the global variance information is demonstrated by results of objective and subjective evaluations.

The paper is organized as follows. In **Section 2**, the ML-based spectral conversion is described. In **Section 3**, the ML-based spectral conversion considering the global variance is described. In **Section 4**, experimental evaluations are described. Finally, we summarize this paper in **Section 5**.

2. ML-BASED SPECTRAL CONVERSION

We use $2D$-dimensional acoustic features $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ (source speaker's) and $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ (target speaker's) consisting of D-dimensional static and dynamic features, where $\top$ denotes transposition of the vector. As described in [6], a GMM on joint probability $p(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\Theta})$ is trained in advance with training data consisting of time-aligned features determined by Dynamic Time Warping (DTW), where $\boldsymbol{\Theta}$ denotes model parameters. The i-th mixture has a weight w_i, mean vectors $\boldsymbol{\mu}_i^{(X)}$ and $\boldsymbol{\mu}_i^{(Y)}$, and covariance matrices $\boldsymbol{\Sigma}_i^{(XX)}$, $\boldsymbol{\Sigma}_i^{(XY)}$, $\boldsymbol{\Sigma}_i^{(YX)}$, and $\boldsymbol{\Sigma}_i^{(YY)}$. In this paper, we use the diagonal covariance matrices.

2.1. Likelihood function for spectral conversion

Let $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ be a time sequence of the source feature vectors and $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ be that of the target feature vectors. We perform the spectral conversion based on maximizing the following likelihood function,

$$p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\Theta}) = \sum_{\text{all } \boldsymbol{m}} p(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{\Theta}) p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{m},\boldsymbol{\Theta}), \tag{1}$$

where $\boldsymbol{m} = \{m_{i1}, m_{i2}, \cdots, m_{iT}\}$ is a mixture sequence. At frame t, $p(m_i|\boldsymbol{X}_t,\boldsymbol{\Theta})$ and $p(\boldsymbol{Y}_t|\boldsymbol{X}_t, m_i, \boldsymbol{\Theta})$ are given by

$$p(m_i|\boldsymbol{X}_t,\boldsymbol{\Theta}) = \frac{w_i N(\boldsymbol{X}_t; \boldsymbol{\mu}_i^{(X)}, \boldsymbol{\Sigma}_i^{(XX)})}{\sum_{j=1}^{M} w_j N(\boldsymbol{X}_t; \boldsymbol{\mu}_j^{(X)}, \boldsymbol{\Sigma}_j^{(XX)})}, \tag{2}$$

$$p(\boldsymbol{Y}_t|\boldsymbol{X}_t, m_i, \boldsymbol{\Theta}) = N(\boldsymbol{Y}_t; \boldsymbol{E}_t(m_i), \boldsymbol{D}(m_i)), \tag{3}$$

where

$$\boldsymbol{E}_t(m_i) = \boldsymbol{\mu}_i^{(Y)} + \boldsymbol{\Sigma}_i^{(YX)} \boldsymbol{\Sigma}_i^{(XX)^{-1}} (\boldsymbol{X}_t - \boldsymbol{\mu}_i^{(X)}), \quad (4)$$

$$\boldsymbol{D}(m_i) = \boldsymbol{\Sigma}_i^{(YY)} - \boldsymbol{\Sigma}_i^{(YX)} \boldsymbol{\Sigma}_i^{(XX)^{-1}} \boldsymbol{\Sigma}_i^{(XY)}. \quad (5)$$

The total number of mixtures is M. The normal distribution with $\boldsymbol{\mu}_i^{(X)}$ and $\boldsymbol{\Sigma}_i^{(XX)}$ is represented as $N(\boldsymbol{X}_t; \boldsymbol{\mu}_i^{(X)}, \boldsymbol{\Sigma}_i^{(XX)})$.

2.2. Spectral determination by maximizing likelihood

We consider the optimum mixture sequence $\boldsymbol{m}$ for maximizing the likelihood function. First, $\boldsymbol{m}$ is determined so that the output probability $p(\boldsymbol{X}|\boldsymbol{m}, \boldsymbol{\Theta})$ is maximized. Then, the logarithm of the likelihood function is written as

$$\log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\Theta}) = -\frac{1}{2} \boldsymbol{Y}^\top \boldsymbol{D}_{\boldsymbol{m}}^{-1} \boldsymbol{Y} + \boldsymbol{Y}^\top \boldsymbol{D}_{\boldsymbol{m}}^{-1} \boldsymbol{E}_{\boldsymbol{m}} + K, \quad (6)$$

where

$$\boldsymbol{E}_{\boldsymbol{m}} = \left[\boldsymbol{E}_1(m_{i1}), \boldsymbol{E}_2(m_{i2}), \cdots, \boldsymbol{E}_T(m_{iT})\right], \quad (7)$$

$$\boldsymbol{D}_{\boldsymbol{m}}^{-1} = \text{diag}\left[\boldsymbol{D}(m_{i1})^{-1}, \boldsymbol{D}(m_{i2})^{-1}, \cdots, \boldsymbol{D}(m_{iT})^{-1}\right]. \quad (8)$$

The constant K is independent of $\boldsymbol{Y}$. The relationship between a sequence of the static feature vectors $\boldsymbol{y} = \left[\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \cdots, \boldsymbol{y}_T^\top\right]^\top$ and a sequence of the static and dynamic feature vectors $\boldsymbol{Y}$ can be represented as a linear conversion,

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \quad (9)$$

where $\boldsymbol{W}$ is a transformation matrix described in [9]. We set coefficients of a delta window to (-0.5, 0, 0.5) in this paper. Under the condition (9), $\boldsymbol{y}$ that maximizes the logarithmic likelihood function is given by

$$\boldsymbol{y} = \left(\boldsymbol{W}^\top \boldsymbol{D}_{\boldsymbol{m}}^{-1} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{D}_{\boldsymbol{m}}^{-1} \boldsymbol{E}_{\boldsymbol{m}}. \quad (10)$$

We can also maximize the logarithm of the likelihood function $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})$ by employing the EM algorithm [9]. There was little difference between the conversion accuracy when using the optimum mixture sequence and that when using the EM algorithm in our preliminary experiments.

Results of our informal evaluations showed that the ML-based conversion causes the performance improvement of voice conversion compared with the conventional GMM-based conversion [5]. These results were similar to those described in [10].

3. ML-BASED SPECTRAL CONVERSION CONSIDERING GLOBAL VARIANCE

3.1. Global variance

The global variance of the static feature vectors in each utterance is written as

$$\boldsymbol{v}(\boldsymbol{y}) = \left[v^{(1)}, v^{(2)}, \cdots, v^{(D)}\right]^\top, \quad (11)$$

$$v^{(d)} = \frac{1}{T} \sum_{t=1}^{T} \left(y_t^{(d)} - \frac{1}{T} \sum_{\tau=1}^{T} y_\tau^{(d)} \right)^2, \quad (12)$$

where $y_t^{(d)}$ is the d-th component of the target static feature vector at frame t.

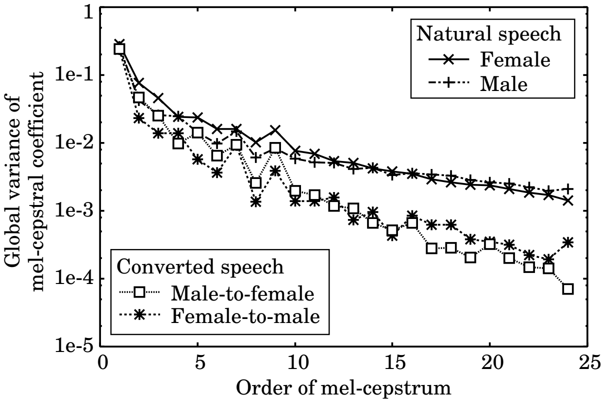

Fig. 1. Global variance of mel-cepstra in an utterance. The average of global variances in 50 utterances is shown in each case.

Figure 1 shows global variances of the converted mel-cepstra and of the natural mel-cepstra of target speech. The experimental conditions are the same as afterward described in **Section 4.1**. It can be observed that the global variances of the converted mel-cepstra are smaller than those of the natural mel-cepstra. Removed variance features are regarded as a noise in modeling acoustic probability density. Surely, this smoothing causes error reduction of the spectral conversion. However, it also causes the degradation of the converted speech quality because those removed features are still necessary for synthesizing high-quality speech.

3.2. Spectral conversion considering global variance

We define a new likelihood function consisting of two probabilities for a sequence of the target feature vectors and for the global variance of the target static feature vectors as follows:

$$L = \log\left\{p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\Theta})^\omega \cdot p(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\Theta}_v)\right\}, \quad (13)$$

where $p(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\Theta}_v)$ is modeled by the normal distribution. A set of model parameters $\boldsymbol{\Theta}_v$ consists of the mean vector $\boldsymbol{\mu}^{(v)}$ and the covariance matrix $\boldsymbol{\Sigma}^{(vv)}$ for the global variance vector $\boldsymbol{v}(\boldsymbol{y})$. The constant ω denotes the weight for the likelihood of the target feature vector sequence. In this paper, ω is set to the ratio of the number of dimensions between $\boldsymbol{v}(\boldsymbol{y})$ and $\boldsymbol{Y}$, i.e., $1/(2T)$.

In order to maximize the likelihood L with respect to $\boldsymbol{y}$, we employ a steepest descent algorithm using the first derivative,

$$\frac{\partial L}{\partial \boldsymbol{y}} = \left(-\boldsymbol{W}^\top \boldsymbol{D}_{\boldsymbol{m}}^{-1} \boldsymbol{W} \boldsymbol{y} + \boldsymbol{W}^\top \boldsymbol{D}_{\boldsymbol{m}}^{-1} \boldsymbol{E}_{\boldsymbol{m}}\right)\omega + \left[v_1^{(1)\prime}, v_1^{(2)\prime}, \cdots, v_2^{(1)\prime}, v_2^{(2)\prime}, \cdots, v_T^{(D)\prime}\right]^\top, \quad (14)$$

$$v_t^{(d)\prime} = -\frac{2}{T} \sum_{i=1}^{D} s_v^{(d,i)} \left(v^{(i)} - \mu_v^{(i)}\right) \left(y_t^{(d)} - \frac{1}{T} \sum_{\tau=1}^{T} y_\tau^{(d)}\right), \quad (15)$$

where $\mu_v^{(i)}$ and $s^{(d,i)}$ denote the i-th component of $\boldsymbol{\mu}^{(v)}$ and the element in the d-th row and the i-th column of $\boldsymbol{\Sigma}^{(vv)^{-1}}$, respectively. The target static sequence estimated from Eq. (10) is used as the initial parameter.

In this paper, we calculate $\boldsymbol{\mu}^{(v)}$ and the diagonal covariance matrix $\boldsymbol{\Sigma}^{(vv)}$ using the target static feature vectors in the training data. We can also estimate these parameters in each utterance from other features, e.g., the global variance of the given source feature vectors or that of the converted vectors by Eq. (10). In our preliminary experiment, there was not the large difference between results when using the parameters directory extracted from the training data and when using the estimated parameters.

4. EXPERIMENTAL EVALUATIONS

In order to investigate the effectiveness of using the global variance information, we compared the proposed algorithm ("MLGV") with the ML-based conversion not considering the global variance ("ML") and the ML-based conversion with spectral enhancement by postfiltering ("ML+PF") [11] in the speaker conversion.

4.1. Experimental conditions

We performed male-to-female and female-to-male voice conversions using MOCHA database [12] consisting of 460 sentences in each speaker. We selected 50 sentences at random as an evaluation set. Then, we selected 50 sentences as a training set from remaining 410 sentences so that the diphone coverage for all sentences was maximized. The resulting diphone coverage of the training set was 91.4%. The total duration of the training data of which silence parts were removed was 2.7 minutes for the female speaker and 2.3 minutes for the male speaker.

We used mel-cepstrum as a spectral feature. The first through 24-th mel-cepstral coefficients were extracted from 16 kHz sampling speech data. The STRAIGHT analysis method [13] was employed for the spectral extraction.

We determined the optimum number of mixtures so that the mel-cepstral distortion between the converted and target mel-cepstra was minimized in the evaluation set. As a result, the number of mixtures was set to 128.

We performed objective evaluations based on the likelihood and subjective evaluations on the converted speech quality and the conversion accuracy for speaker individuality. We varied the coefficient β of the postfilter for mel-cepstrum [11] from 0.1 to 0.8. In order to measure only the performance of the spectral conversion, we synthesized the converted speech using the natural prosodic features automatically extracted from target speech as follows. A time-alignment for modifying duration was performed with DTW, and then at each frame, F_0 and total power of the converted linear spectrum were set to each target value. The STRAIGHT synthesis method [13] was employed as a speech synthesizer.

4.2. Objective evaluation

Table 1 shows the logarithmic likelihood on the mel-cepstral sequence $\log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})$, which is normalized by the number of frames. Results for the natural mel-cepstral sequence of the target are also shown in this table. It is reasonable that the likelihood decreases by applying the criterion on the global variance or the postfilter to the ML-based conversion. However, the likelihoods in those cases don't fall below those of the target mel-cepstrum.

Table 1 also shows the logarithmic likelihood on the global variance $\log p(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\Theta}_v)$. It can be seen that the likelihoods in the ML-based conversion are very low. Although the likelihood can be improved by using the postfilter, the improved likelihood is much lower than that of the target. On the other hand, the likelihood when considering the global variance ("MLGV") is larger than that of the target.

Table 1. Logarithmic likelihood on mel-cepstral sequence and logarithmic likelihood on global variance. Left numbers in each column show results for the female-to-male conversion, and right numbers show results for the male-to-female conversion.

	Mel-cepstra		Global variance	
ML	113.7	109.6	−36.1	−112.0
MLGV	107.7	103.1	136.6	136.5
ML+PF ($\beta = 0.1$)	113.2	109.2	−16.8	−91.3
ML+PF ($\beta = 0.2$)	112.0	108.4	1.1	−72.1
ML+PF ($\beta = 0.3$)	110.0	107.0	16.8	−55.2
ML+PF ($\beta = 0.4$)	107.3	105.2	29.2	−41.7
ML+PF ($\beta = 0.5$)	103.9	102.8	37.3	−32.7
ML+PF ($\beta = 0.6$)	99.8	100.1	40.1	−29.3
ML+PF ($\beta = 0.7$)	95.1	96.9	36.2	−32.8
ML+PF ($\beta = 0.8$)	90.0	93.4	24.5	−44.6
Target	86.5	82.2	120.7	118.6

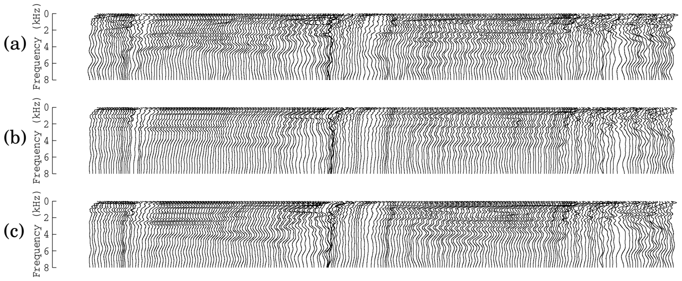

Fig. 2. An example of spectra of (a) target speech, (b) converted speech by the ML, and (c) converted speech by the ML using global variance, for a sentence fragment "farmers grow oats."

These results demonstrate that the converted mel-cepstral sequence having similar characteristics to the target is estimated by applying the criterion on the global variance to the ML-based conversion. An example of spectra of the target and converted voices is shown in **Fig. 2**.

4.3. Subjective evaluation

We performed a MOS test on speech quality and XAB test on speaker individuality. In the MOS test, an opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). In the XAB test, an analysis-synthesized target speech was presented as X, and the converted voices were presented as A and B. Listeners were asked to choose either A or B as being more similar to X. We used 25 sentences in the evaluation set[1]. The number of listeners was five.

Figure 3 shows results of the MOS test. The proposed algorithm can obviously improve the converted speech quality. Although the spectral enhancement by postfiltering is also effective for improving the quality, the improved quality is inferior to that by the proposed algorithm. In postfiltering, the mel-cepstral coefficients except for the first coefficient are basically emphasized at a constant rate [11]. Whereas, in the ML using the global variance, the emphasis rate is varied according to the conditional probability distribution at each of frames and dimensions. Therefore, more reasonable enhancement is performed compared with postfiltering.

[1] Some samples are available in
http://kt-lab.ics.nitech.ac.jp/˜tomoki/ICASSP2005/index.html

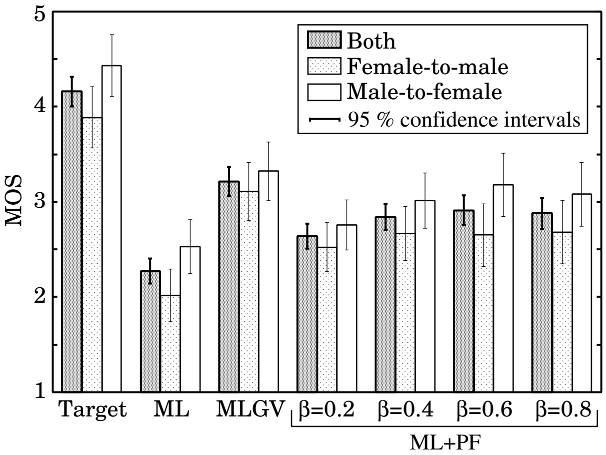

Fig. 3. Results of MOS test on speech quality. “Target” shows the result for analysis-synthesized target speech using the 0-th thorough 24-th mel-cepstral coefficients.

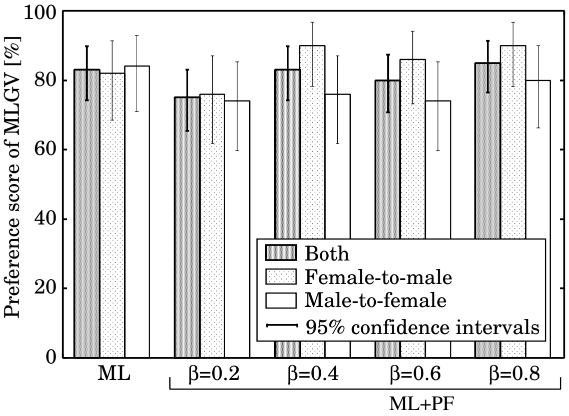

Fig. 4. Results of XAB test on speaker individuality. Preference score shows the rate of judging that the converted speech by the ML using global variance is more similar to the target compared with that by the other methods.

Figure 4 shows results of the XAB test. It is observed that the proposed algorithm can synthesize a more similar speech to the target than the other methods. It seems that these results are caused by the improvement of speech quality while keeping the characteristics of target spectra. Since it might be that the global variance feature itself contributes to the speaker individuality, we will further investigate the effect of using the global variance information on speaker conversion.

5. CONCLUSIONS

We proposed a spectral conversion method based on maximum likelihood considering the global variance of the converted parameter in each utterance. Experimental results demonstrated that the performance of voice conversion can be significantly improved by using the global variance information. Moreover, it was shown that the proposed algorithm is more effective than the postfilter-based spectral enhancement. We can apply the proposed method to both spectral and F_0 generation algorithms for the HMM-based Text-to-Speech synthesis [11].

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 Prof. Hideki Kawahara of Wakayama University in Japan for permission to use the STRAIGHT analysis-synthesis method.

6. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. *J. Acoust. Soc. Jpn. (E)*, Vol. 11, No. 2, pp. 71–76, 1990.

[2] M. Abe, K. Shikano, and H. Kuwabara. Statistical analysis of bilingual speaker’s speech for cross-language voice conversion. *J. Acoust. Soc. Am.*, Vol. 90, No. 1, pp. 76–82, 1991.

[3] M. Mashimo, T. Toda, H. Kawanami. K. Shikano, and N. Campbell. Cross-language voice conversion evaluation using bilingual databases. *IPSJ Journal*, Vol. 43, No. 7, pp. 2177–2185, 2002.

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

[5] Y. Stylianou. *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*. Ph.D. Thesis, Ecole Nationale Supérieure des Télécommunications, 1996.

[6] A. Kain. *High Resolution Voice Transformation*. Ph.D. Thesis, OGI School of Science and Engineering, Oregon Health and Science University, 2001.

[7] T. Toda. *High-quality and flexible speech synthesis with segment selection and voice conversion*. Ph.D. Thesis, Graduate School of Information Science, Nara Institute of Science and Technology, 2003.

[8] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu. Voice conversion with smoothed GMM and MAP adaptation. *Proc. EUROSPEECH*, pp. 2413–2416, Geneva, Switzerland, Sep. 2003.

[9] 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.

[10] T. Toda, A.W. Black, and K. Tokuda. Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. *Proc. 5th ISCA Speech Synthesis Workshop*, pp. 31–36, Pittsburgh, USA, June 2004.

[11] 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.

[12] A. Wrench. The MOCHA-TIMIT articulatory database. *http://www.cstr.ed.ac.uk/artic/mocha.html*, Queen Margaret University College, 1999.

[13] 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.