Modulation Spectrum-Based Post-Filter for GMM-Based Voice Conversion

Shinnosuke Takamichi*†, Tomoki Toda*, Alan W Black† and Satoshi Nakamura*
* Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan
E-mail: shinnosuke-t@is.naist.jp
† Language Technologies Institute, Carnegie Mellon University (CMU), U. S. A

Abstract—**This paper addresses an over-smoothing effect in Gaussian Mixture Model (GMM)-based Voice Conversion (VC). The flexible use of the statistical approach is one of the major reason why this approach is widely applied to the speech-based systems. However, quality degradation by over-smoothed speech parameter converted is unavoidable problem of statistical modeling. One of common approaches to this over-smoothness in conversion step is to compensate generated features, such as Global Variance (GV), that explicitly express the over-smoothing effect. In statistical Text-To-Speech (TTS) synthesis, we have recently introduced a Modulation Spectrum (MS) which is an extended form of GV, and have proposed MS-based Post-Filter (MSPF) in Hidden Markov Model (HMM)-based TTS synthesis. In this paper, we apply the MSPF to GMM-based VC. Because the MS of speech parameters is degraded through GMM-based conversion process, we perform the post-filter due to MS modification of converted parameters. The experimental evaluation yields the quality benefits by the proposed post-filter.**

I. Introduction

Statistical Voice Conversion (VC) is an effective technique for modifying speech parameters to convert non-linguistic information while keeping linguistic information unchanged. It have gained popularity due to its flexible application to speech-based systems such as disability-aid [1], singing-voice synthesis [2], speech-to-speech translation [3], and non-native speech modification [4]. Gaussian Mixture Model (GMM)-based VC is the state-of-the-art statistical approach. Trajectory-level conversion from the source speaker to the target speaker is performed through the GMMs that the relationship between these speakers is jointly trained. One of the biggest issues in the GMM-based VC is quality degradation of converted speech. Because the accuracy to model fluctuating speech parameters is insufficient, the over-smoothed speech parameters are generated in the conversion stage based on the Maximum Likelihood (ML)-based criterion [5], and this over-smoothness causes the degradation of speech quality and speaker individuality in the converted speech.

Many attempts to address quality improvements are reported. Post-filtering to the converted speech parameters is common approach in conversion stage. We classify these methods into two types: speech emphasis and parameter conversion. The former, such as formant emphasis [6] and peak-to-valley emphasis [7], emphasizes the converted parameters based on the knowledge of speech perception. The latter, such as Global Variance (GV)-based conversion [8] and event-based conversion [9], converts the specific features of the converted parameter into natural one. What latter methods are automatically trainable is a big advantages. T. Toda et al. integrated GV into parameter generation process [5], which generates speech parameters considering GV statistics. Although not only the generation methods but also GV itself are widely applied [10], the quality degradation are still large because the over-smoothness are still remained. For quality improvements in HMM-based TTS synthesis, we have introduced a new feature called Modulation Spectrum (MS) [11], [12] which is regarded as extended form of GV, and have proposed the MS-based Post-Filter (MSPF) [13]. MSPF can achieve the high-quality speech generation by modifying the MS of the generated speech in HMM-based speech synthesis.

In this paper, we apply the MSPF to GMM-based VC. Because the training and conversion stage in GMM-based VC degrade the MS of speech parameters, we filter the generated parameters to close its MS to natural one in the conversion stage. The MSPF can generate the naturally-fluctuated temporal parameter sequence. The result of perceptual assessment demonstrate the quality gain by the proposed method.

II. Conversion Algorithm in Gaussian Mixture Model-based Voice Conversion [5]

In training stage, a joint probability density of speech parameters of the source and target speaker's voice is modeled with a GMM using a parallel data set as follows:

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t | \boldsymbol{\lambda}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t\end{bmatrix}; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right), \quad (1)$$

where $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ and $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ are joint static and dynamic feature vectors of the source and target speakers, respectively. $\boldsymbol{x}_t = \left[x_t(1), \cdots, x_t(D)\right]^\top$ and $\boldsymbol{y}_t = \left[y_t(1), \cdots, y_t(D)\right]^\top$ are D-dimensional static feature vectors of source and target speakers at frame t, respectively. $\mathcal{N}\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denotes the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The total number of mixture components is M. $\boldsymbol{\lambda}$ is a GMM parameter sets consisting of the mixture-component weight α_m, the mean vector $\boldsymbol{\mu}_m^{(X,Y)}$ and the covariance matrix $\boldsymbol{\Sigma}_m^{(X,Y)}$ of the m-th mixture component. Also, we estimate statistics of target speaker's GV $\boldsymbol{v}\left(\boldsymbol{y}\right)$ that is the 2nd moment of parameter

978-616-361-823-8 © 2014 APSIPA

trajectory $\boldsymbol{y} = \left[\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_T^\top\right]^\top$, which is defined as:

$$\begin{aligned} \boldsymbol{v}(\boldsymbol{y}) &= \left[v(1), \cdots, v(d), \cdots, v(D)\right]^\top, & (2) \\ v(d) &= \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\right)^2, & (3) \end{aligned}$$

where T is the number of frames.

In conversion stage, the parameter sequence of source speaker's voice $\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_T^\top\right]^\top$ is converted to maximize both GMM likelihood and GV likelihood:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{y} | \boldsymbol{W}\boldsymbol{x}, \boldsymbol{\lambda}\right) P\left(\boldsymbol{v}(\boldsymbol{y}) | \boldsymbol{\lambda}_v\right)^{\omega}, \tag{4}$$

where $\boldsymbol{\lambda}_v$ is GV parameter sets, $\boldsymbol{W}$ is a weight matrix to calculate dynamic feature vector sequence [14], ω is a weight of GV likelihood. In this paper, we approximate GMM with a single mixture component [5]. Scaling in temporal domain is compensated by considering GV statistics. However, speech quality and speaker individuality of the converted speech excessively degrades because the converted parameter trajectory is still temporally over-smoothed.

III. Modulation Spectrum (MS) and MS-based post-filter (MSPF) for GMM-based VC

A. Modulation Spectrum (MS)

The MS $\boldsymbol{s}(\boldsymbol{y})$ is defined as log-scaled power spectrum of the temporal sequence $\boldsymbol{y}$, which is calculated as

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{y}) &= \left[\boldsymbol{s}(1)^\top, \cdots, \boldsymbol{s}(d)^\top, \cdots, \boldsymbol{s}(D)^\top\right]^\top, & (5) \\ \boldsymbol{s}(d) &= \left[s_d(0), \cdots, s_d(f), \cdots, s_d(D_s)\right]^\top, & (6) \end{aligned}$$

where $s_d(f)$ is the f-th MS of the d-th dimension of the parameter sequence $\left[y_1(d), \cdots, y_T(d)\right]^\top$, f is a modulation frequency index, D_s is one half number of the DFT length. In this paper, the MS is calculated from an utterance that is zero-padded to set its length to $2D_s$. As illustrated in Figure 1 consisting of MSs of the natural and converted mel-ceptral coefficient parameter sequences, the training and conversion processes deteriorate the MS of speech parameters[1]. Moreover, we have found more excessive degradation in the higher quefrency and higher modulation frequency component.

B. MS-based Post-Filtering (MSPF)

The post-filter is trained using the training data including natural and converted speech of the target speaker's voice.

1) Training: The following probability distribution function is estimated from natural speech parameter trajectory:

$$P\left(\boldsymbol{s}(\boldsymbol{y}) | \boldsymbol{\lambda}_s\right) = \mathcal{N}\left(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{\mu}^{(\mathrm{N})}, \boldsymbol{\Sigma}^{(\mathrm{N})}\right), \tag{7}$$

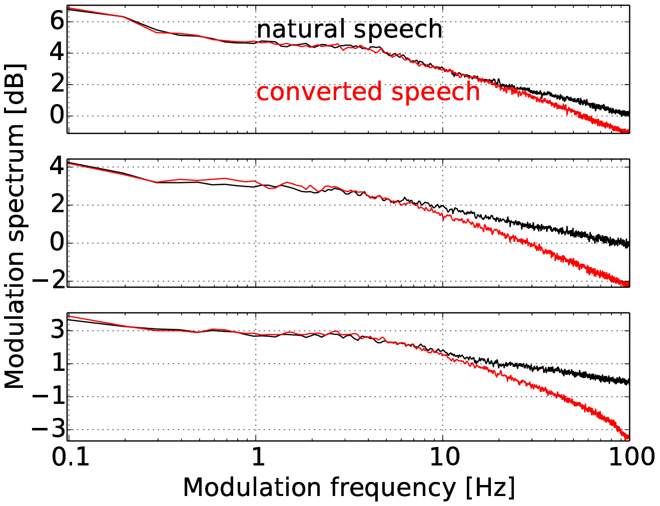

Fig. 1. Averaged modulation spectra of 1st, 5th, and 9th mel-cepstral coefficient sequences from above.

where $\mathcal{N}\left(\cdot; \boldsymbol{\mu}^{(\mathrm{N})}, \boldsymbol{\Sigma}^{(\mathrm{N})}\right)$ is a Gaussian distribution of a mean vector $\boldsymbol{\mu}^{(\mathrm{N})} = \left[\mu_{1,0}^{(\mathrm{N})}, \cdots, \mu_{D,D_s}^{(\mathrm{N})}\right]^\top$ and a diagonal covariance matrix $\boldsymbol{\Sigma}^{(\mathrm{N})} = \operatorname{diag}\left[\left(\sigma_{1,0}^{(\mathrm{N})}\right)^2, \cdots, \left(\sigma_{D,D_s}^{(\mathrm{N})}\right)^2\right]$, $\mu_{d,f}^{(\mathrm{N})}$ and $\left(\sigma_{d,f}^{(\mathrm{N})}\right)^2$ is a mean and a variance of $s_d(f)$ and $\boldsymbol{\lambda}_s$ is a parameter set of MS. Probability distribution function $\mathcal{N}\left(\cdot; \boldsymbol{\mu}^{(\mathrm{G})}, \boldsymbol{\Sigma}^{(\mathrm{G})}\right)$ is estimated in the same manner using the speech parameter trajectory generated with the generation method described in Section II.

2) Conversion: The following filter is applied to the converted speech parameter sequence $\boldsymbol{y}$:

$$\begin{aligned} s_d'(f) &= (1-k)s_d(f) \\ &+ k\left[\frac{\sigma_{d,f}^{(\mathrm{N})}}{\sigma_{d,f}^{(\mathrm{G})}}\left(s_d(f) - \mu_{d,f}^{(\mathrm{G})}\right) + \mu_{d,f}^{(\mathrm{N})}\right], \end{aligned} \tag{8}$$

where k is a post-filter emphasis coefficient valued between 0 and 1. The MS of the converted speech become nearly in natural MS as increasing k. The finally filtered parameter trajectory is calculated from the filtered MS and frequency phase characteristics of the parameter trajectory, which are calculated before filtering.

As we have illustrated in Figure 1, higher modulation frequency components in the converted MS are degraded. Therefore, the MSPF runs like adaptive high-pass filter applied to temporal parameter sequence, and the filtered temporal parameters include natural fluctuation.

IV. Experimental Evaluations

A. Experimental Conditions

In our experiments, we prepared two Japanese speakers of male and female. We selected 50 parallel sentences of subset A from phonetically balanced 503 sentences included in the ATR Japanese speech database [15] for training, and 50 sentences of subset B for evaluation. We trained male-to-female and female-to-male GMMs. Speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 1th-through-24th mel-cepstral coefficients were used as spectral parameters and log-scaled F_0 and 5 band-aperiodicity [16], [17] were used as

[1] As we have reported in [13], MSs of the converted speech are recovered by compensating GV

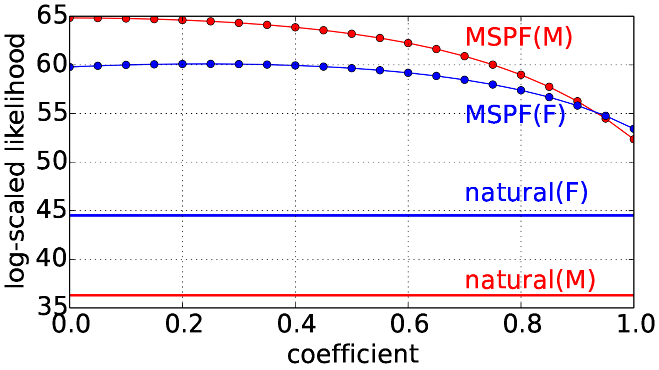

Fig. 2. GMM likelihood for the filtered spectral parameters.

excitation parameters. The STRAIGHT analysis-synthesis system [18] is employed for parameter extraction and waveform generation. The spectral parameters and aperiodic components is converted through a 64-mixture GMM and a 16-mixture GMM, respectively. The log-scaled F_0 is linearly converted. The DFT length to calculate MS is set to 2048, which is over the maximum frame length in training and evaluation data. The MSPF is applied to only spectral parameters. We didn't apply both parameter generation considering GV and the MSPF to the aperiodic components because these methods do not cause any large improvement in the aperiodic component.

We conducted objective and subjective evaluations with two systems: 1) **"Conv"**: parameter trajectory converted considering GV described in Section II, and 2) **"MSPF"**: filtered parameter trajectory by MSPF. In all evaluations, "M" and "F" indicate target male speaker and target female speaker, respectively.

B. Objective Evaluation for Tuning Emphasis Coefficient

In order to determine the filter emphasis coefficient, we calculate the GMM likelihood, GV likelihood, and MS likelihood for filtered parameter trajectory of the evaluation data under settings the emphasis coefficient from 0 to 1. For comparison, the likelihood for natural speech parameter sequence is calculated, which is labeled as **"natural."**

The GMM likelihood, GV likelihood, and MS likelihood are shown in Figure 2, Figure 3, and Figure 4, respectively. Note that "MSPF" at setting coefficient to 0.0 is equal to "Conv." It is observed that the GMM likelihood of "MSPF" decreases as the coefficient increases, and the MS likelihood of "MSPF" increase as the coefficient increases, and it is the closest to "natural" when coefficient is 1.0. We can also see that the magnitude correlation between "MSPF" and "natural" never change in all settings in GMM likelihood and MS likelihood. On the other hand, the transition of GV likelihood shows different tendency between two speakers. We find that the GV likelihood of "MSPF" is the closest to that of "natural" when the coefficients are 0.90 for male speaker and 0.70 for female speaker. From these results, we determined the filter emphasis coefficients to 0.90 and 0.70, respectively.

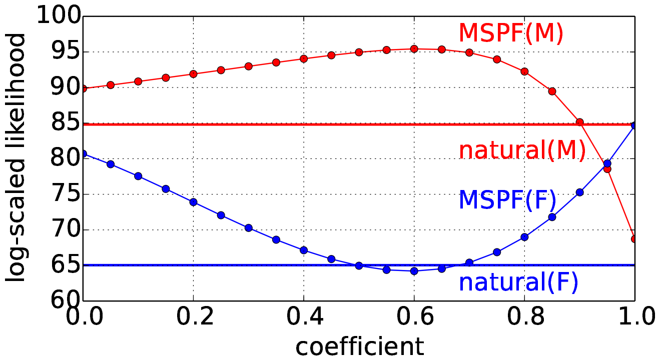

Fig. 3. GV likelihood for the filtered spectral parameters.

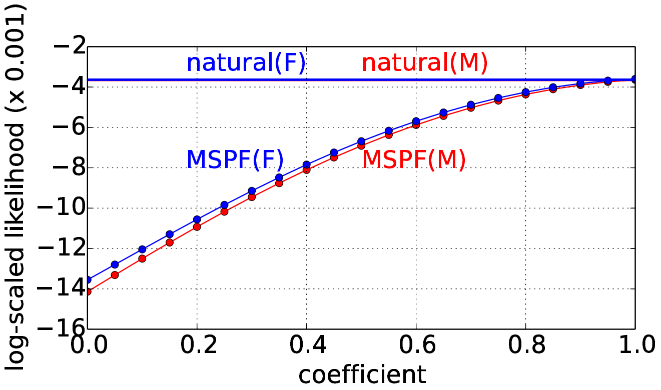

Fig. 4. MS likelihood for the filtered spectral parameters.

C. Subjective Evaluation on Speech Quality and Speaker Individuality

To investigate the effect of the MSPF on speech quality and speaker individuality, we first conducted a preference test (AB test) on speech quality. We presented every pair of converted speech of two systems in random order. Similarity, we conducted XAB test on speaker individuality. We first presented an analysis-synthesized reference speech as "X", then we presented random-ordered converted speech. We forced listeners to prefer speech sample. 7 listeners are prepared in each assessment.

The results of the preference tests on speech quality and speaker individuality are shown at left side and right side of Figure 5, respectively. Moreover, an example of the spectrograms is shown in Figure 6. More fluctuated spectrogram by the MSPF is observed in Figure 6. In term of speech quality, meaningful quality gain is observed in both speaker. Here, we could the effectiveness of MSPF on speech quality in both statistical TTS [13] and VC. However, unfortunately, there is no significant different in preference test on speaker individuality. We expect that no cues for individuality are at higher modulation frequency which is recovered by the MSPF.

V. Summary

In this paper, we applied the modulation spectrum-based post-filter to the GMM-based voice conversion. The post-filter filters parameter trajectory converted through maximum

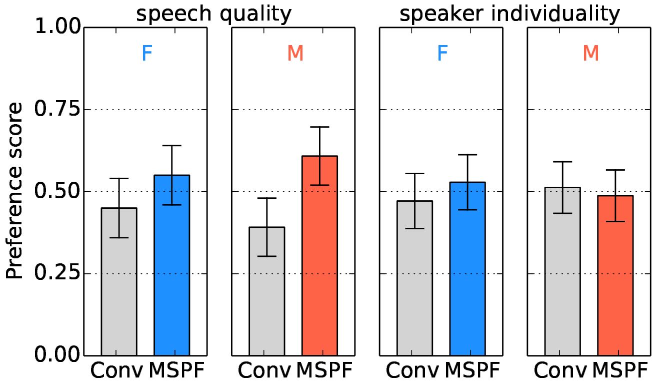

Fig. 5. Preference scores on speech quality and speaker individuality with 95% confidence interval bars.

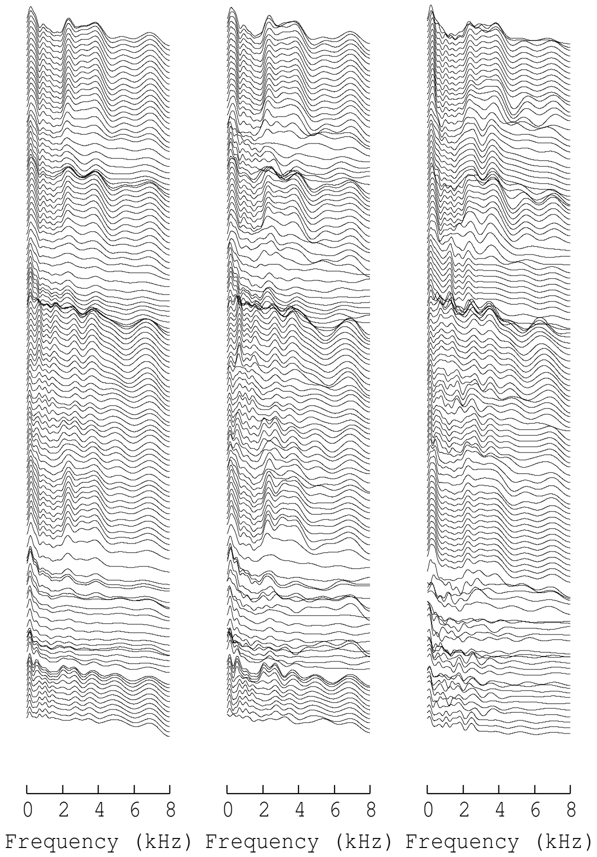

Fig. 6. An example of spectrograms representing "Conv," "MSPF," and "natural" from left.

likelihood estimation using Gaussian mixture models. We evaluated the post-filter on both speech quality and speaker individuality. The experimental results yielded the effectiveness of the post-filter in term of speech quality. As future work, we will incorporate the modulation spectrum to the parameter generation algorithm.

Acknowledgements: Part of this work was supported by JSPS KAKENHI Grant Number 26280060 and Grant-in-Aid for JSPS Fellows Grant Number 26 · 10354, and part of this work was executed under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation." The authors are grateful to Prof. Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method.

REFERENCES

[1] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. An evaluation of excitation feature prediction in a hybrid approach to electrolaryngeal speech enhancement. In *Proc. ICASSP*, pp. 4521–4525, Florence, Italy, May 2014.

[2] K. Kobayashi, T. Toda, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura. Regression approaches to perceptual age control in singing voice conversion. In *Proc. ICASSP*, pp. 7954–7958, Florence, Italy, May 2014.

[3] N. Hattori, T. Toda, H. Kawai, H. Saruwatari, and K. Shikano. Speaker-adaptive speech synthesis based on eigenvoice conversion and language-dependent prosodic conversion in speech-to-speech translation. In *Proc. INTERSPEECH*, pp. 2769–2772, Florence, Italy, Aug. 2011.

[4] S. Aryal and R. G.-Osuna. Can voice conversion be used to reduce non-native accents? In *Proc. ICASSP*, pp. 7929–7933, Florence, Italy, May 2014.

[5] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis. *IEICE Trans. Inf. Syst.*, Vol. J87-D-II, No. 8, pp. 1563–1571, 2004.

[7] F. Eyben and Y. Agiomyrgiannakis. A frequency-weighted post-filtering transform for compensation of the over-smoothing effect in HMM-based speech synthesis. In *Proc. ICASSP*, pp. 275–279, Florence, Italy, May 2014.

[8] T. Toda, T. Muramatsu, and H. Banno. Implementation of conputationally efficient real-time voice conversion. In *Proc. INTERSPEECH*, Portland, Oregon, U.S., Sept. 2012.

[9] Y. Jiao, X. Na, and M. Tu. Improving voice quality of HMM-based speech synthesis using voice conversion method. In *Proc. ICASSP*, pp. 7964–7968, Florence, Italy, May 2014.

[10] H. Hwang, Y. Tsao, H. Wang, Y. Wang, and S. Chen. Incorporating global variance in the training phase of GMM-based voice conversion. In *Proc. APSIPA*, pp. 1–6, Kaohsiung, Taiwan, Oct. 2013.

[11] R. Drullman, J .M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. of America*, Vol. 95, pp. 2670–2680, 1994.

[12] S. Thomas, S. Ganapathy, and H. Hermansky. Phoneme recgnition usng spectral envelop and modulation frequency features. In *Proc. ICASSP*, pp. 4453–4456, Taipei, Taiwan, April 2009.

[13] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A post-filter to modify modulation spectrum in HMM-based speech synthesis. In *Proc. ICASSP*, pp. 290–294, Florence, Italy, May 2014.

[14] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.

[15] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuawhara. A large-scale Japanese speech database. In *ICSLP90*, pp. 1089–1092, Kobe, Japan, Nov. 1990.

[16] H. Kawahara, Jo Estill, and O. Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT " . In *MAVEBA 2001*, pp. 1–6, Firentze, Italy, Sept. 2001.

[17] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In *Proc. INTERSPEECH*, pp. 2266–2269, Pittsburgh, U.S.A., Sep. 2006.

[18] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207, 1999.