A POSTFILTER TO MODIFY THE MODULATION SPECTRUM IN HMM-BASED SPEECH SYNTHESIS

Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti and Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology,
8916–5 Takayama-cho, Ikoma, Nara, 630-0192, Japan

ABSTRACT

In this paper, we propose a postfilter to compensate modulation spectrum in HMM-based speech synthesis. In order to alleviate over-smoothing effects which is a main cause of quality degradation in HMM-based speech synthesis, it is necessary to consider features that can capture over-smoothing. Global Variance (GV) is one well-known example of such a feature, and the effectiveness of parameter generation algorithm considering GV have been confirmed. However, the quality gap between natural speech and synthetic speech is still large. In this paper, we introduce the Modulation Spectrum (MS) of speech parameter trajectory as a new feature to effectively capture the over-smoothing effect, and we propose a postfilter based on the MS. The MS is represented as a power spectrum of the parameter trajectory. The generated speech parameter sequence is filtered to ensure that its MS has a pattern similar to natural speech. Experimental results show quality improvements when the proposed methods are applied to spectral and F_0 components, compared with conventional methods considering GV.

Index Terms— HMM-based speech synthesis, over-smoothing, global variance, modulation spectrum, postfilter

1. INTRODUCTION

Text-To-Speech (TTS) is a technology that can convert any text into speech, and it plays an important role in many speech applications. The demand of synthesis techniques that can synthesize natural-sounding speech is rapidly growing. One of the major reasons that HMM-based speech synthesis [1] have been an active research target is its voice control capability [2, 3, 4] based on the elegant framework of HMMs. On the other hand, the quality in synthetic speech is strongly degraded compared with natural speech and the synthetic speech sounds muffled [5]. This is because traditional generation frameworks generate over-smoothed parameter trajectories.

Global Variance (GV) [6] is one of the well-known features to capture this over-smoothing effect. Despite the fact that the GV is calculated in a simple form according to the second moment of parameters, generation algorithms considering the GV can efficiently alleviate the over-smoothing effect. However, the quality gap between natural speech and synthetic speech is still large.

In this paper, we introduce the Modulation Spectrum (MS) of parameter trajectory as a new feature to effectively capture the over-smoothing effect, and propose a postfilter to compensate MS. The MS is represented as the power spectrum of the temporal parameter sequence. The effectiveness of the MS in capturing the sound of speech has been noted in other research area, such as spectral cues of speech perception [7], and the use as acoustic features in HMM-based speech recognition [8]. Because the generated sequence is temporally smoothed, the MS of the synthetic speech tends to be degraded compared with that of the natural speech even when a generation algorithm considering GV is used. Therefore, the proposed method filters to fluctuate the generated parameter sequence. The postfilter is trained using training data consisting of natural and synthetic speech. Experimental results show quality improvements when the proposed methods is applied to spectral and F_0 components, compared with conventional method considering GV.

2. PARAMETER GENERATION IN HMM-BASED SPEECH SYNTHESIS

2.1. Maximizing HMM Likelihood [9]

In HMM-based speech synthesis, context-dependent HMMs are trained using natural speech parameters. In synthesis, sentence HMMs corresponding to input text to be synthesized are constructed, and speech parameter trajectory is generated to maximize HMM likelihood under a constraint on the relationship between static and dynamic features, which is as follows:

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}} P\left(\boldsymbol{W}\boldsymbol{c}|\boldsymbol{\lambda}\right), \tag{1}$$

where $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_t^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$ is a speech parameter vector sequence of T frames, $\boldsymbol{c}_t = \left[c_t\left(1\right), \cdots, c_t\left(d\right), \cdots, c_t\left(D\right)\right]^\top$ is a D-dimensional parameter vector at frame t, d is a dimensional index, $\boldsymbol{W}$ is the weighting matrix for calculating the dynamic features [9], $\boldsymbol{\lambda}$ is a HMM parameter set, respectively.

Parameter sequences generated with Eq. (1) tend to be over-smoothed, and the synthetic speech sounds muffled compared with the natural speech.

2.2. Maximizing HMM and GV Likelihood [6]

The GV is defined as second moment of the parameter trajectory, and is calculated as:

$$\boldsymbol{v}\left(\boldsymbol{c}\right) = \left[v\left(1\right), \cdots, v\left(d\right), \cdots, v\left(D\right)\right]^\top, \tag{2}$$

$$v\left(d\right) = \frac{1}{T}\sum_{t=1}^{T}\left(c_t\left(d\right) - \frac{1}{T}\sum_{\tau=1}^{T}c_\tau\left(d\right)\right)^2. \tag{3}$$

Speech parameter trajectory is generated to maximize both HMM and GV likelihoods.

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

where $\boldsymbol{\lambda}_v$ is a parameter set of GV and w is a weight of the GV likelihood. The statistics of the GV are trained from the natural speech parameters.

The GV generated using Eq. (1) is usually smaller than that of the natural speech parameters. Compensation of GV by this method improves speech quality, but the improvements are still limited.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

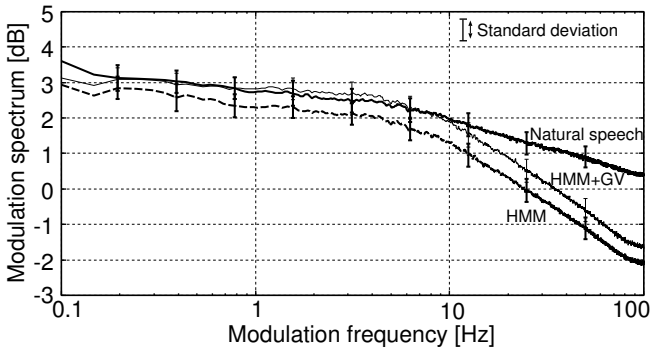

Fig. 1. Modulation spectrum of 9-th mel-cepstral coefficient sequence.

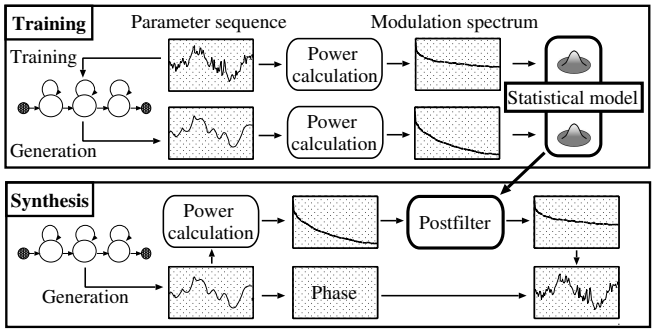

Fig. 2. Schematic diagram of the proposed method.

3. MODULATION SPECTRUM ANALYSIS

Though a Modulation Spectrum (MS) is traditionally defined as a value calculated using the Fourier transform of the parameter sequence [10], in this paper we define the MS as its log-scaled power spectrum. This feature, of course, can represent the temporal fluctuation of the parameter sequence. The MS $\boldsymbol{s}(\boldsymbol{c})$ of parameter sequence $\boldsymbol{c}$ is calculated as:

$$\boldsymbol{s}(\boldsymbol{c}) = \left[\boldsymbol{s}(1)^\top, \cdots, \boldsymbol{s}(d)^\top, \cdots, \boldsymbol{s}(D)^\top\right]^\top, \quad (5)$$

$$\boldsymbol{s}(d) = \left[s_d(0), \cdots, s_d(m), \cdots, s_d(M)\right]^\top, \quad (6)$$

where $s_d(m)$ is the m-th MS of the d-th dimension of the parameter sequence $[c_1(d), \cdots, c_t(d), \cdots, c_T(d)]^\top$, m is a modulation frequency index, M is one half number of the Discrete Fourier Transform (DFT) length. In this paper, the MS is calculated from a parameter sequence that is zero-padded to set its sequence length to $2M$.

Here, we analyze the MS of natural and synthetic speech. The MS mean of the 9-th mel-cepstral coefficient sequence generated using Eq. (1) ("HMM") and Eq. (4) ("HMM+GV") are shown in Fig. 1. Additionally, the MS of natural speech parameter sequence ("natural speech") is shown in same figure for comparison. It is observed that the MS of "HMM" is markedly degraded compared with that of "natural speech." This is because temporal fluctuation observed in the natural speech parameter sequence is lost in the HMM frameworks. We can also find that the MS of "HMM+GV" is closer to natural one but there is still a big gap between the MSs of "HMM+GV" and "natural speech."

From these result, we can expect further improvements in quality by directly accounting for this difference in the MS.

4. POSTFILTER BASED ON MODULATION SPECTRUM

In this section, we propose a postfilter to compensate the MS of the generated parameter sequence. The schematic diagram of the proposed method is shown in Fig. 2. Parameters of the proposed postfilter are trained using natural and generated parameter sequences in the training data.

4.1. Training Process

The following probability distribution function is estimated from natural speech parameter sequences:

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

where $\mathcal{N}\left(\cdot; \boldsymbol{\mu}^{(\mathrm{N})}, \boldsymbol{\Sigma}^{(\mathrm{N})}\right)$ is a Gaussian distribution of mean vector $\boldsymbol{\mu}^{(\mathrm{N})} = \left[\mu^{(\mathrm{N})}_{1,0}, \cdots, \mu^{(\mathrm{N})}_{D,M}\right]^\top$ and diagonal covariance matrix $\boldsymbol{\Sigma}^{(\mathrm{N})} = \mathrm{diag}\left[\left(\sigma^{(\mathrm{N})}_{1,0}\right)^2, \cdots, \left(\sigma^{(\mathrm{N})}_{D,M}\right)^2\right]$, $\mu^{(\mathrm{N})}_{d,m}$ and $\left(\sigma^{(\mathrm{N})}_{d,m}\right)^2$ is a mean and a variance of $s_d(m)$ 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 sequences generated with the generation method described in Section 2. To avoid the effect of the duration difference between natural and generated speech parameter sequence, the parameter sequence is generated under the natural speech duration.

4.2. Synthesis Process

The following filter is applied to the generated speech parameter sequence $\boldsymbol{c}$:

$$s'_d(m) = (1-k)s_d(m) + k\left[\frac{\sigma^{(\mathrm{N})}_{d,m}}{\sigma^{(\mathrm{G})}_{d,m}}\left(s_d(m) - \mu^{(\mathrm{G})}_{d,m}\right) + \mu^{(\mathrm{N})}_{d,m}\right], \quad (8)$$

where k is a postfilter emphasis coefficient valued between 0 and 1. If $k = 1$, the MS will be modified to be close to the MS of natural speech parameter sequences. On the other hand, if $k = 0$, the filtered sequence will be the same as the non-filtered sequence. The filtered parameter sequence is calculated from the MS and frequency phase characteristics of the parameter sequence, which are calculated before filtering.

4.3. Application to F_0 Component

While the proposed postfilter can be directly applied to the spectral component, additional process is required for application to the F_0 component because observed F_0 contours are not a continuous sequence. To solve this problem, we use continuous F_0 modeling [11] which can estimate F_0 values at the unvoiced frames. According to [12], we estimate the F_0 values of the unvoiced frames with spline-based interpolation. To avoid the MS's fluctuation of the continuous F_0 contour, we removed micro prosody by Low Pass Filter (LPF). We believe that the effect of micro prosody on speech quality is small as referring in [13]. Moreover, we subtracted utterance-level F_0 from original F_0 values before estimating continuous F_0 contours to avoid discontinuous transition in zero-padding process. Since the F_0 estimation quality is degraded by spline-based extrap-

olation, we calculate the MS from the non-silence frames[1].

In synthesis, the utterance-level mean and unvoiced/voiced regions of the generated F_0 contour are extracted before applying the proposed filter. After filtering, first, the filtered continuous F_0 contour is calculated in the same manner as the spectral component. Then, the filtered F_0 contour is calculated by adding the mean to the filtered continuous F_0 contour and restoring the unvoiced/voiced regions.

4.4. Relationship to GV-based Postfilter

A postfiltering process to ensure the GV of the generated parameter sequence is proposed in [14]. The generated speech parameter are linearly converted as follows:

$$\hat{c}_t(d) = \sqrt{\frac{\mu_d^{(\mathrm{GV,N})}}{\mu_d^{(\mathrm{GV,G})}}}\{c_t(d) - \langle c_t(d)\rangle\} + \langle c_t(d)\rangle, \tag{9}$$

where $\mu_d^{(\mathrm{GV,N})}$, $\mu_d^{(\mathrm{GV,G})}$ are the GV mean of d-th dimension of the natural and synthetic speech parameters in the training data, respectively, and $\langle c_t(d)\rangle$ is the mean of d-th dimension of the synthetic speech parameters. In this method, since only the variance of the sequence is considered, the MS degradation is not completely recovered, thus the generation of parameters including temporal fluctuation of natural speech parameters is difficult. On the other hand, the proposed method can recover this fluctuation because we directly consider the MS itself. Therefore, the proposed method can be expected to yield quality improvements.

According to the Perceval's theorem, the power of a temporal sequence is preserved during a DFT. the GV defined in Eq. (3) represents the power of the sequence except the bias component. Because the MS is defined as the power spectrum of the sequence, the sum of the MS over all modulation spectra except the bias component is equivalent to the GV[2]. In the GV-based postfiltering process, MSs of all modulation frequency except bias is converted in the same way. Namely, the GV-based conversion process is special case of the proposed MS-based conversion process under the following conditions:

$$\mu_{d,m}^{(\cdot)} = 0, \quad \sigma_{d,m}^{(\cdot)} = \begin{cases} 1 & m = 0 \\ \mu_d^{(\mathrm{GV},\cdot)} & \text{otherwise} \end{cases}, \tag{10}$$

in which the postfilter emphasis coefficient is set to 1. Conversely, the proposed method can convert MSs in each modulation spectrum individually.

5. EXPERIMENTAL EVALUATION

5.1. Experimental Conditions

We trained a context-dependent phoneme Hidden Semi-Markov Model (HSMM) [15] for a Japanese female speaker. We used 450 sentences for training and 53 sentences for evaluation from phonetically balanced 503 sentences included in the ATR Japanese speech database [16]. Speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 0th-through-24th mel-cepstral coefficients were extracted as spectral parameters and log-scaled F_0 and 5 band-aperiodicity [17, 18] were extracted as excitation parameters by the STRAIGHT analysis system [19]. The feature vector consisted of spectral and excitation parameters and their delta and delta-delta features. Five-state left-to-right HSMMs were used. The DFT length to calculate MS is set to 4096, which is over the maximum frame length in training and evaluation data. A 10 Hz-cutoff LPF is used to remove the micro prosody[3].

We conducted some evaluations with the following systems:

HMM : generated with Eq. (1)

HMM+MS : applied the proposed postfilter to "HMM"

HMM+GV : generated with Eq. (4)

HMM+GV+MS : applied the proposed postfilter to "HMM+GV"

Note that the postfilter of "HMM+GV+MS" is trained using parameter sequences generated with GV.

5.2. Objective Evaluation for Emphasis Coefficient

In order to determine the filter emphasis coefficient for the spectral and F_0 components, we calculate the HMM likelihood, GV likelihood, and MS likelihood for filtered parameter sequence for settings of the emphasis coefficient from 0 to 1. For comparison, the likelihood for natural speech parameter sequence is calculated.

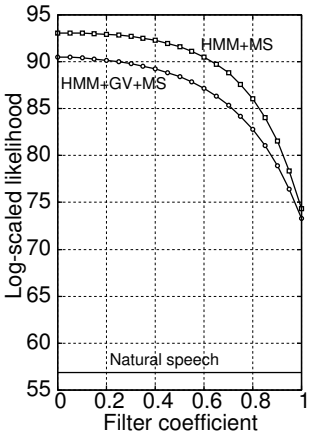

Fig. 3. HMM likelihood for the filtered spectral parameters.

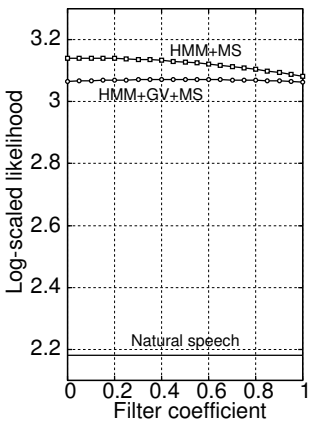

Fig. 4. HMM likelihood for the filtered F_0 contours.

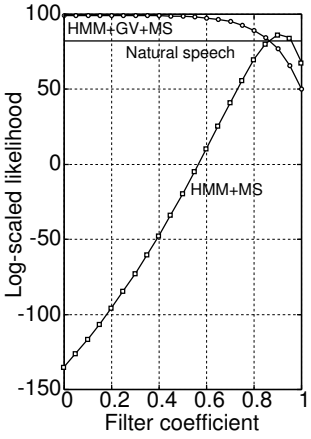

Fig. 5. GV likelihood for the filtered spectral parameters.

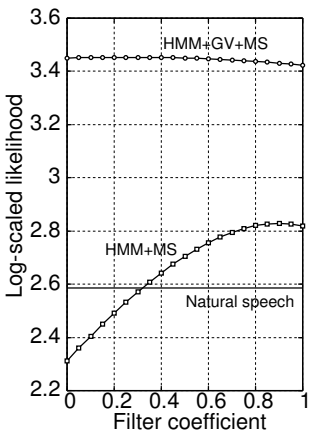

Fig. 6. GV likelihood for the filtered F_0 contours.

[1]We also considered simple approaches to estimate F_0 of silence such as the use of utterance-level mean of F_0 or the use of the F_0 value in the nearest voiced frame. However, we have confirmed that current method is better to model the MS.

[2]Properly describing, the sum of linear-scaled MS except bias is equivalent to GV.

[3]We evaluated training accuracy of MS likelihood in various cutoff frequencies, and we have confirmed that this setting is the best.

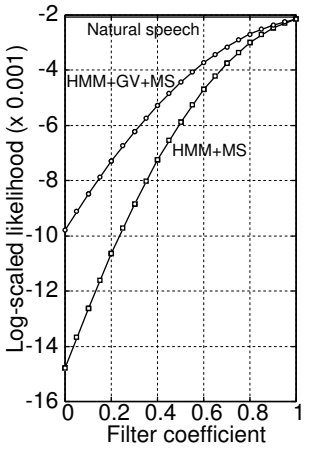

Fig. 7. MS likelihood for the filtered spectral parameters.

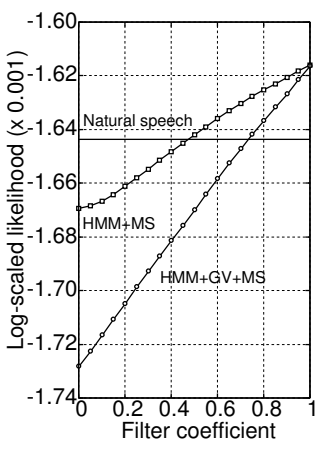

Fig. 8. MS likelihood for the filtered continuous F_0 contours.

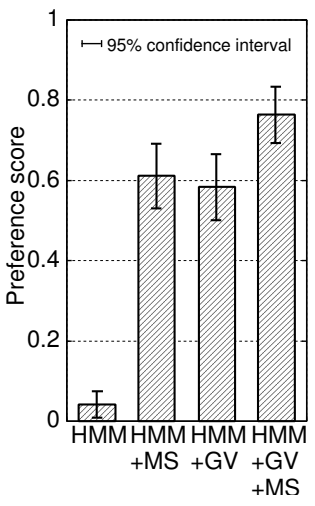

Fig. 9. Preference score for spectral component.

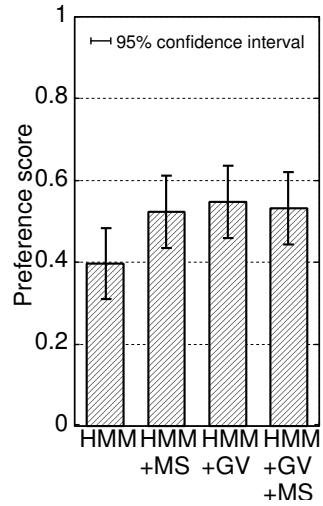

Fig. 10. Preference score for F_0 component.

The HMM likelihood, GV likelihood, and MS likelihood for filtered spectral parameters are shown in Fig. 3, Fig. 5 and Fig. 7, respectively. It is observed in Fig. 3 that the HMM likelihoods of "HMM+MS" and "HMM+GV+MS" decrease as the emphasis coefficient increases, and their values are always higher than that of "natural speech." In the GV likelihood shown in Fig. 5, we can see that these likelihoods cross that of "natural speech" at 0.85. On the other hand, MS likelihoods increase as the coefficient increases but their values always lower than "natural speech." From these results, we determined filter emphasis coefficient for spectral component to be 0.85.

Those likelihoods for the filtered F_0 contour are shown in Fig. 4, Fig. 6 and Fig. 8, respectively. The transition of these likelihoods as the coefficient changes show the same tendency as those for the spectral components except the relation with the likelihoods of "natural speech." We can find that all likelihoods of "HMM+MS" and "HMM+GV+MS" are higher than "natural speech" when setting the emphasis coefficient over 0.75, and we can also find that the coefficient 1.0 is the highest point of MS likelihood. From these results, we set the coefficient to 1.0.

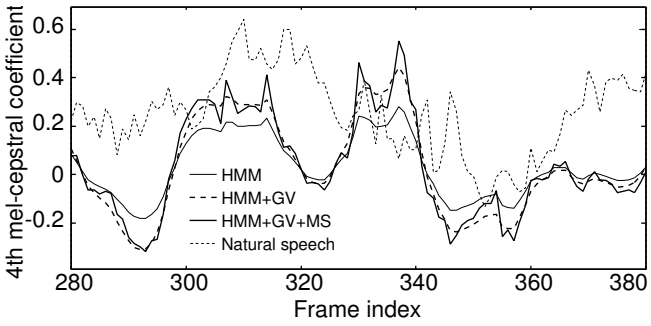

Fig. 11. Examples of the natural and generated 4th mel-cepstral coefficient sequence.

5.3. Subjective Evaluation for Speech Quality

To evaluate the quality improvements yielded by the proposed postfilter, we conducted a preference test (AB test) on speech quality by eight listeners for the spectral and F_0 components[4]. Every pair of these four types of synthetic speech was presented to listeners in random order. Listeners were asked which sample sounds better in terms of speech quality. The "HMM" system is used for the component that the proposed methods were not applied to.

The result of the preference test for spectral component is shown in Fig. 9, and an example of the spectral parameter sequence is shown in Fig. 11. We can see that the score of "HMM+MS" system dramatically increases over the "HMM" system, and achieves a similar score to the "HMM+GV" system. Additionally, further improvement by applying the proposed method to "HMM+GV" can be observed. From these result, the effectiveness of the proposed method in quality in the spectral component was yielded.

Similarly, the preference score for the F_0 component is shown in Fig. 10. Again, "HMM+MS" and "HMM+GV+MS" achieve a better score than "HMM," but there are not additional gains over when GV is considered. The reason why the score differences among conventional and proposed methods are smaller than those in the spectral components is that both natural and generated F_0 contour transitions are smoothly and these MSs are closer than those in spectral parameter sequence.

6. SUMMARY

In this paper, we proposed a postfilter to compensate the modulation spectrum of the generated parameter trajectory in HMM-based speech synthesis. The experimental results demonstrated that the quality improvements by the proposed method are yielded for both spectral and F_0 components. 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 22680016.

[4]Some samples are available from `http://isw3.naist.jp/~shinnosuke-t/sample_mspf.html`.

7. REFERENCES

[1] H. Zen, K. Tokuda, and A. Black. Statistical parametric speech synthesis. *Speech Commun.*, Vol. 51, No. 11, pp. 1039–1064, 2009.

[2] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura. Speaker interpolation for HMM-based speech synthesis system. *J. Acoust. Soc. Jpn. (E)*, Vol. 21, No. 4, pp. 199–206, 2000.

[3] J. Yamagishi and T. Kobayashi. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Trans., Inf. and Syst.*, Vol. E90-D, No. 2, pp. 533–543, 2007.

[4] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi. A style control technique for HMM-based expressive speech synthesis. *IEICE Trans., Inf. and Syst.*, Vol. E90-D, No. 9, pp. 1406–1413, 2007.

[5] S. King and V. Karaiskos. The blizzard challenge 2011. In *Proc. Blizzard Challenge workshop*, Turin, Italy, Sept. 2011.

[6] T. Toda and K. Tokuda. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans.*, Vol. E90-D, No. 5, pp. 816–824, 2007.

[7] 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.

[8] 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.

[9] 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.

[10] L. Atlas and S. A.Shamma. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing*, Vol. 7, pp. 668–675, 2003.

[11] K. Yu and S. Young. Continuous F0 modeling for HMM based statistical parametric speech synthesis. *IEEE Trans. Audio, Speech and Language*, Vol. 19, No. 5, pp. 1071–1079, 2011.

[12] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A hybrid approach to electrolaryngeal speech enhansement based on spectral subtraction and statistical voice conversion. In *Proc. INTERSPEECH*, pp. 3067–3071, Lyon, France, Sep. 2013.

[13] P. Taylor. *Text-To-Speech synthesis*. Cambridge Univ. Press, 2009.

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

[15] H. Zen, K. Tokuda, T. Kobayashi T. Masuko, and T. Kitamura. Hidden semi-markov model based speech synthesis system. *IEICE Trans., Inf. and Syst., E90-D*, No. 5, pp. 825–834, 2007.

[16] 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.

[17] 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.

[18] 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.

[19] 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.