Post-Filters to Modify the Modulation Spectrum for Statistical Parametric Speech Synthesis

Shinnosuke Takamichi, *Student Member, IEEE,* Tomoki Toda, *Member, IEEE,* Alan W. Black, *Member, IEEE*
Graham Neubig, *Nonmember, IEEE* Sakriani Sakti, *Member, IEEE* Satoshi Nakamura, *Fellow, IEEE*

Abstract—**This paper presents novel approaches based on Modulation Spectrum (MS) for high-quality statistical parametric speech synthesis, including Text-To-Speech (TTS) and Voice Conversion (VC). Although statistical parametric speech synthesis offers various advantages over concatenative speech synthesis, the synthetic speech quality is still not as good as that of concatenative speech synthesis or the quality of natural speech. One of the biggest issues causing the quality degradation is the over-smoothing effect often observed in the generated speech parameter trajectories. Global Variance (GV) is known as a feature well correlated with the over-smoothing effect, and the effectiveness of keeping the GV of the generated speech parameter trajectories similar to those of natural speech has been confirmed. However, the quality gap between natural speech and synthetic speech is still large. In this paper, we propose using the MS of the generated speech parameter trajectories as a new feature to effectively quantify the over-smoothing effect. Moreover, we propose post-filters to modify the MS utterance by utterance or segment by segment to make the MS of synthetic speech close to that of natural speech. The proposed post-filters are applicable to various synthesizers based on statistical parametric speech synthesis. We first perform an evaluation of the proposed method in the framework of Hidden Markov Model (HMM)-based TTS, examining its properties from different perspectives. Furthermore, effectiveness of the proposed post-filters are also evaluated in Gaussian Mixture Model (GMM)-based VC and Classification And Regression Trees (CART)-based TTS (a.k.a., CLUSTERGEN). The experimental results demonstrate that (1) the proposed utterance-level post-filter achieves quality comparable to the conventional generation algorithm considering the GV, and yields significant improvements by applying to the GV-based generation algorithm in HMM-based TTS. (2) the proposed segment-level post-filter capable of achieving low-delay synthesis also yields significant improvements in synthetic speech quality, and (3) the proposed post-filters are also effective in not only HMM-based TTS but also GMM-based VC and CLUSTERGEN.**

Index Terms—**Statistical parametric speech synthesis, over-smoothing, post-filter, global variance, modulation spectrum, HMM-based text-to-speech, GMM-based voice conversion, CLUSTERGEN**

I. Introduction

ALTHOUGH human beings naturally utilize their own speech as a communication tool, there exist many barriers in speech communication, such as vocal disorders [1], [2], [3], language differences [4], [5], [6], and physical constraints [3], [7]. Many speech technologies have been studied to break down these barriers. One of the promising technologies is parametric speech generation [8], including Text-To-Speech (TTS) synthesis [9] and Voice Conversion (VC) [10]. TTS is a technique to synthesize speech corresponding to a given text, and VC is a technique to convert non-/para-linguistic characteristics of input speech while preserving linguistic characteristics.

In parametric speech generation, statistical parametric speech synthesis [11] was established in the 1990s [12], [10], and has gained popularity in this decade. Nowadays, many technologies have been studied within this basic framework, including speech synthesis using Hidden Markov Models (HMM) [13], Gaussian Mixture Model (GMM) [14], Classification And Regression Trees (CART) [15], kernel regression [16], [17], and Deep Neural Nets (DNN) [18], [19]. Whereas concatenative speech synthesis [9], [20] directly uses waveform segments or natural speech parameter segments to generate a speech waveform, statistical parametric speech synthesis collects statistics from the speech parameter segments and utilizes these to generate speech parameters to be used in speech waveform generation. This statistical modeling and generation framework make it possible to build small footprint synthesizers [21], adapt existing voices to other target voices using only a small amount of speech data [22], and flexibly control voice characteristics of synthetic speech [23], [24].

On the other hand, a serious drawback of statistical parametric speech synthesis compared to concatenative speech synthesis is the lower quality of synthetic speech [25], [26]. There are three main reasons causing the quality degradation [11]: parameterization errors in the speech analysis/synthesis stage [27], [28], [29], inaccurate modeling in the training stage [30], [18], and over-smoothness of the generated speech parameters in the synthesis stage [25], [26]. In particular, the last factor, the over-smoothing effect usually makes synthetic speech sound muffled compared to concatenative speech synthesis or natural speech. One promising approach to alleviate the over-smoothing effect is to extract a specific feature to quantify the over-smoothing effect and to generate speech parameters so that their corresponding features become more similar to those of natural speech parameters. One widely known example of such a feature is Global Variance (GV) [31], [14], which is a second order moment of the speech parameter sequence. Considering the GV during parameter generation effectively works to alleviate the over-smoothing effect and to significantly improve synthetic speech quality. Currently, the

S. Takamichi, S. Sakti, G. Neubig and S. Nakamura are with the Graduate School of Information Science, Nara Institute of Science and Technology, Japan.

T. Toda is with Information Technology Center, Nagoya University, Japan.

A. W. Black is with Language Technologies Institute, Carnegie Mellon University, United States.

Manuscript received XXX XX, 20XX; revised XXX XX, 20XX.

GV-based parameter generation has been applied in a number of ways [32], [33], [34], [35]. However, the use of this metric in the parameter generation tends to additionally generate artificial sounds [32], [34] and the quality gap between natural and synthetic speech is still large.

In this paper, we propose a new feature more sensitively correlated to the over-smoothing effect than the GV, the Modulation Spectrum (MS). The linear-scaled MS of a speech parameter sequence is defined as the power spectrum of the sequence. The linear-scaled MS, like GV, is a second order moments of the parameter sequence. The effectiveness of the MS in capturing speech properties has been noted in other research areas, such as spectral cues of speech perception [36], the use as acoustic features in HMM-based speech recognition [37] and acoustic signal classification [38], and as a countermeasure to discriminate synthetic speech from natural speech in speaker verification [39]. Related to the perceptual effect, [40], [41] investigated the effect of the MS (especially, lower modulation frequency band) on the perceptual intelligibility. Because generated speech parameter sequences tend to be temporally smoothed by the parameter generation process, the MS of synthetic speech tends to be degraded compared to that of natural speech. This MS degradation is still observed even when GV is used in parameter generation. The post-filtering approach proposed in this paper remedies this problem by modifying the generated speech parameter sequence so that its MS becomes more similar to that of natural speech. The proposed post-filter modifies the MS utterance by utterance and can be automatically constructed using natural speech and synthetic speech as training data. This utterance-level post-filter is further extended to a segment-level post-filter to modify the MS segment by segment in order to achieve low-delay parameter generation [42], [43].

We first evaluate the proposed post-filters in HMM-based TTS from various perspectives. Then, we evaluate them in other speech synthesizers: the utterance-level post-filter in GMM-based VC [14] and the segment-level post-filter in CART-based TTS (a.k.a., CLUSTERGEN) [15]. The experimental results show that (1) the proposed post-filters effectively improve the naturalness of the spectrum, F_0, and HMM-state duration, yielding significant quality improvements in HMM-based TTS (as also shown in [44], [45]), (2) the proposed segment-level post-filter is also effective (as also shown in [45]), and (3) the proposed post-filters are effective in not only HMM-based TTS but also GMM-based VC (as also shown in [46]) and CLUSTERGEN.

The rest of this paper is organized as follows. Section II briefly reviews HMM-based TTS, GMM-based VC, and CLUSTERGEN. Section III presents the concept of our study, the MS, and Section IV proposes the MS-based post-filter. Section V and Section VI are the experimental evaluation and conclusion, respectively.

II. Speech Parameter Generation in Statistical Parametric Speech Synthesis

This section describes the speech parameter generation procedures in statistical parametric speech synthesis frameworks, such as HMM-based TTS, GMM-based VC, and CLUSTERGEN.

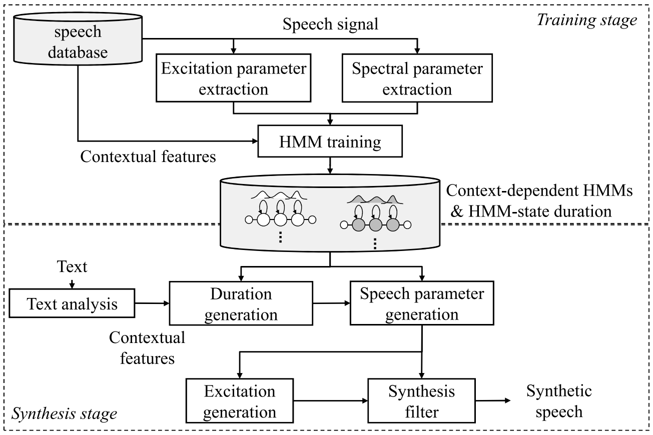


Fig. 1. An overview of HMM-based TTS.

A. HMM-Based TTS [13]

Figure 1 illustrates an overview of HMM-based TTS. The HMM models the T-frame output speech feature sequence $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ given the contextual factor sequence $\boldsymbol{X}$ of the input text, where $\boldsymbol{Y}_t$ is a speech feature vector at frame t. Speech features and HMM-state duration are simultaneously modeled in a unified framework [47]. In synthesis, given the contextual feature sequence $\boldsymbol{X}$, the corresponding HMM is first constructed, then the HMM-state sequence $\hat{\boldsymbol{q}} = [\hat{q}_1, \cdots, \hat{q}_t, \cdots, \hat{q}_T]$ is determined by maximizing the duration probability density function as follows:

$$\hat{\boldsymbol{q}} = \underset{\boldsymbol{q}}{\operatorname{argmax}}\, P\left(\boldsymbol{q} \middle| \boldsymbol{X}, \boldsymbol{\lambda}\right), \qquad (1)$$

where $\hat{q}_t$ is a HMM-state at frame t, and $\boldsymbol{\lambda}$ is the parameter set of the HMM. The synthetic speech parameter sequence is generated by the following algorithms.

Using HMMs [48]: The speech parameter sequence is generated by maximizing the following output probability density function under the explicit relationship between static and dynamic features.

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{y} \middle| \boldsymbol{q}, \boldsymbol{X}, \boldsymbol{\lambda}\right), \qquad (2)$$

where $\hat{\boldsymbol{y}} = \left[\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top\right]^\top$ is a speech parameter vector sequence of T frames, $\boldsymbol{y}_t = \left[y_t(1), \cdots, y_t(d), \cdots, y_t(D)\right]^\top$ is a D-dimensional speech feature vector at frame t, d is a dimension index, and $\boldsymbol{W}$ is the weighting matrix for calculating the dynamic features [11], where $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$.

Synthetic speech parameter sequences generated by Eq. (2) tend to be over-smoothed, and the synthetic speech sounds muffled compared to the natural speech.

Using HMMs and a GV model [31]:

The GV is defined as a second order moment of the parameter trajectory, which is calculated as

$$\boldsymbol{v}\left(\boldsymbol{y}\right) = \left[v\left(1\right), \cdots, v\left(d\right), \cdots, v\left(D\right)\right]^{\top}, \tag{3}$$

$$v\left(d\right) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t\left(d\right) - \frac{1}{T}\sum_{\tau=1}^{T} y_{\tau}\left(d\right)\right)^2. \tag{4}$$

The speech parameter sequence is generated by maximizing a weighted combination of the two probability density functions.

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{y}\middle|\, \boldsymbol{q}, \boldsymbol{X}, \boldsymbol{\lambda}\right) P\left(\boldsymbol{v}\left(\boldsymbol{y}\right)\middle|\, \boldsymbol{\lambda}_{\mathrm{v}}\right)^{w_{\mathrm{v}}}, \tag{5}$$

where $\boldsymbol{\lambda}_{\mathrm{v}}$ is the parameter set of the GV, and w_{v} is the weight of the GV probability density function. The probability density function of the GV is trained from the natural speech parameters in the training data.

B. GMM-Based VC [14]

In GMM-based VC, a GMM performs frame-level modeling using input and output speech feature sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$. The Dynamic Time Warping (DTW) algorithm is used to temporally align $\boldsymbol{X}$ and $\boldsymbol{Y}$. In synthesis, mixture component sequence $\hat{\boldsymbol{q}}$ is determined similarly to Eq. (1), where q_t is a mixture component at frame t. The synthetic speech parameter sequence $\hat{\boldsymbol{y}}$ is determined in the same manner as Eq. (5), where $P\left(\boldsymbol{W}\boldsymbol{y}\middle|\, \boldsymbol{q}, \boldsymbol{X}, \boldsymbol{\lambda}\right)$ is derived from the GMM.

C. CLUSTERGEN [15]

Whereas HMM-based TTS ties the probability density functions over multiple frames with the HMM-state-level probability density function, which is usually determined by decision tree clustering based on the Minimum Description Length (MDL) criterion [49], CLUSTERGEN predicts the probability density functions frame by frame in the CART framework. The output probability density function in Eq. (2) is calculated using the contextual factor sequence $\boldsymbol{X}$. The synthetic speech parameter sequence $\hat{\boldsymbol{y}}$ is determined in the same manner as Eq. (2).

III. Modulation Spectrum Analysis

Though the MS is traditionally defined as a value calculated using the Fourier transform of the parameter sequence [50], this paper defines the MS as its log-scaled power spectrum. The temporal fluctuation of the parameter sequence is modeled as power values of individual modulation frequency components of the parameter sequence. The MS $\boldsymbol{s}\left(\boldsymbol{y}\right)$ of the parameter sequence $\boldsymbol{y}$ is calculated as:

$$\boldsymbol{s}\left(\boldsymbol{y}\right) = \left[\boldsymbol{s}\left(1\right)^{\top}, \cdots, \boldsymbol{s}\left(d\right)^{\top}, \cdots, \boldsymbol{s}\left(D\right)^{\top}\right]^{\top}, \tag{6}$$

$$\boldsymbol{s}\left(d\right) = \left[s_d\left(0\right), \cdots, s_d\left(f\right), \cdots, s_d\left(D_{\mathrm{s}}\right)\right]^{\top}, \tag{7}$$

$$s_d\left(f\right) = \log\left(\left(\sum_{t=1}^{T} y_t\left(d\right)\cos mt\right)^2 + \left(\sum_{t=1}^{T} y_t\left(d\right)\sin mt\right)^2\right), \tag{8}$$

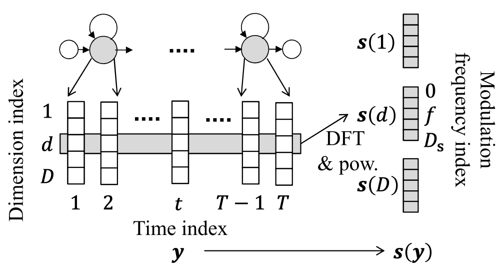


Fig. 2. Graphic representation of how to derive the MS $\boldsymbol{s}\left(\boldsymbol{y}\right)$ from the speech parameter sequence $\boldsymbol{y}$.

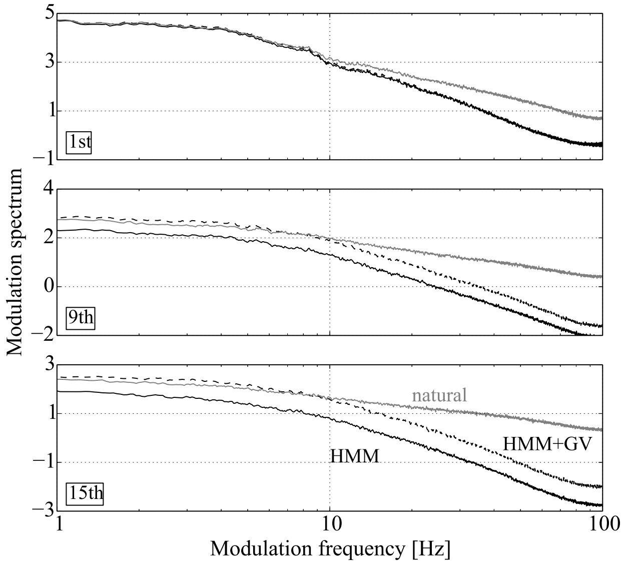


Fig. 3. Averaged MSs of the 1st, 9th and 15th mel-cepstral coefficient sequences from above in HMM-based TTS.

where f is a modulation frequency index, $m = -\pi f / D_{\mathrm{s}}$ is a modulation frequency, and D_{s} is one half of the Discrete Fourier Transform (DFT) length. The MS is calculated from zero-padded parameter sequences so its length is $2D_{\mathrm{s}}$. As shown in Fig. 2, $\boldsymbol{s}\left(\boldsymbol{y}\right)$ is given as a super vector consisting of the MSs corresponding to individual feature dimensions.

To demonstrate how the MS allows us to capture relevant frequency characteristics, we first demonstrate some characteristics of the MS of natural and synthetic speech. Figure 3 shows the MS mean of the mel-cepstral coefficient sequences generated using Eq. (2) ("HMM") and Eq. (5) ("HMM+GV") in HMM-based TTS. Additionally, the MS mean of a natural speech parameter sequence ("natural") is shown in the same figure for comparison. It can be observed that the MS of "HMM" is markedly degraded compared to that of "natural." This is because temporal fluctuation observed in the natural speech parameter sequences is lost in the HMM framework. We can also find that the MS of "HMM+GV" is closer to natural speech in lower modulation frequency bands but there is still a large gap between the MSs of "HMM+GV" and "natural speech" in higher modulation frequency bands (more than 10 Hz). From these results, we can expect that further

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2522655, IEEE/ACM Transactions on Audio, Speech, and Language Processing

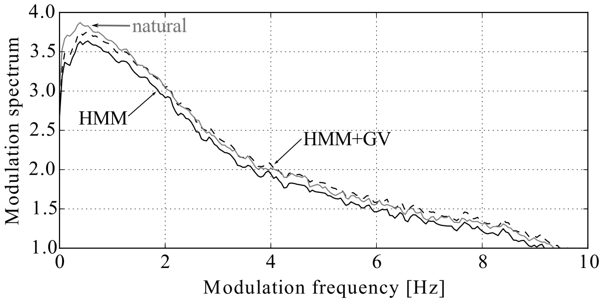


Fig. 4. Averaged MSs of log-F0 contours in HMM-based TTS. Note that the Nyquist frequency is 100 Hz similarly to the spectral parameters, but only < 10 Hz components are shown.

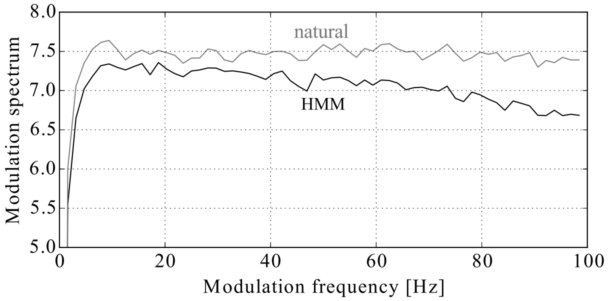


Fig. 5. Averaged MSs of phoneme-level duration in HMM-based TTS. Note that the pseudo Nyquist frequency is set to 100 Hz because we cannot define the Nyquist frequency for duration.

quality improvements will be yielded by compensating for these differences in the MS.

In addition, we consider the spectral tilt of the MS (defined as "MS tilt") which indicates the power difference between the lower and the higher modulation frequency components in Fig. 3. We can observe that the MS tilt of the natural mel-cepstrum tends to increase in the higher order mel-cepstral coefficients. On the other hand, the MS tilt of the generated mel-cepstrum is similar among different order mel-cepstral coefficients. Even when using the GV in the parameter generation "HMM+GV," the MS is just shifted and the MS tilt is not changed. These results show that the parameter generation process shown by Eq. (2) or Eq. (5) tends to constrain the MS tilt of the generated speech parameter sequence to be unnatural.

In addition to the cepstral coefficients, we can also calculate the MSs of the other features. as described in the following section. The MS of the F_0 contour shown in Fig. 4 is also degraded by the statistical process. Higher modulation frequency components of the generated MS are almost the same as those of natural speech, but lower components are slightly different. HMM-state duration determined by Eq. (1) is also affected by the over-smoothing effect due to the statistical averaging process implicit in conventional parameter generation, as in the spectrum and F_0 components [47], [51]. Figure 5 shows the MS mean of phoneme duration sequences. We can see that the generated MS is generally smaller than that of natural speech.

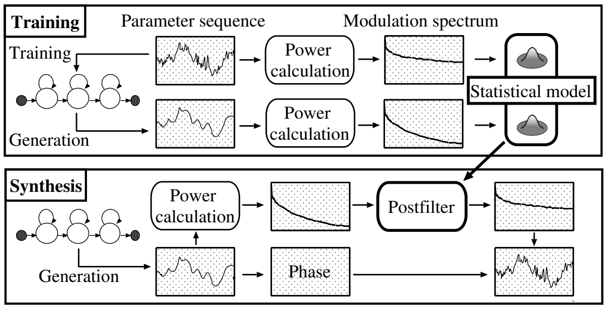


Fig. 6. A schematic diagram of the proposed MS-based post-filter to modify the MS of the generated parameter sequence in the case of HMM-based TTS.

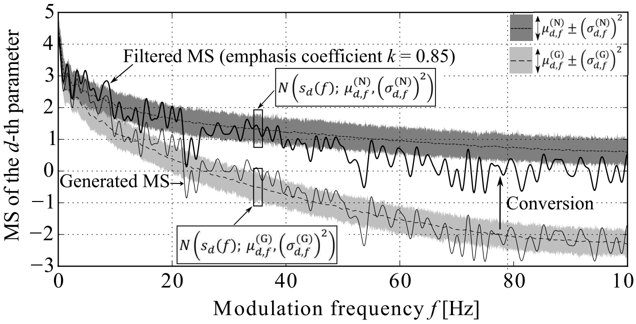


Fig. 7. An example of the MS conversion in the synthesis stage. Note that the MS envelope ("Generated MS" and "Filtered MS") is drawn instead of the MS itself for clear illustration. The MS envelope is calculated by low-pass liftering the cepstrum of MS.

IV. Proposed MS-Based Post-Filter

This section proposes post-filters to modify the MS of the generated parameter sequence. Figure 6 shows a schematic diagram of the proposed method. Parameters of the proposed post-filter are automatically trained using natural and generated speech parameter sequences in the training data. The speech parameters are generated by an individual speech synthesizer. First, the utterance-level MS-based post-filter is described for spectrum, F_0, and HMM-state duration. Then, the segment-level MS-based post-filter is derived by localizing the utterance-level post-filtering process.

A. Utterance-level MS-Based Post-Filter

The post-filter described in this section performs utterance-level MS filtering. The MS is calculated from a parameter sequence that is zero-padded to set its sequence length to $2D_{\mathrm{s}}$, and it is assumed that the sequence length in the training and synthesis is less than $2D_{\mathrm{s}}$.

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

$$P\left(\boldsymbol{s}|\boldsymbol{\lambda}_s\right) = \mathcal{N}\left(\boldsymbol{s}; \boldsymbol{\mu}_{\mathrm{s}}^{(\mathrm{N})}, \boldsymbol{\Sigma}_{\mathrm{s}}^{(\mathrm{N})}\right), \tag{9}$$

TABLE I
THE DETAILED PROCEDURE OF THE PROPOSED POST-FILTERING PROCESS.

1	Zero-pad the original parameter sequence.
2	Take the DFT and store the phase characteristics.
3	Calculate the log-scaled power spectrum (= MS).
4	Apply the post-filter to the MS.
5	Compute the power and add the original phase.
6	Take the inverse DFT.
7	Truncate the resulting signal to have an appropriate length.

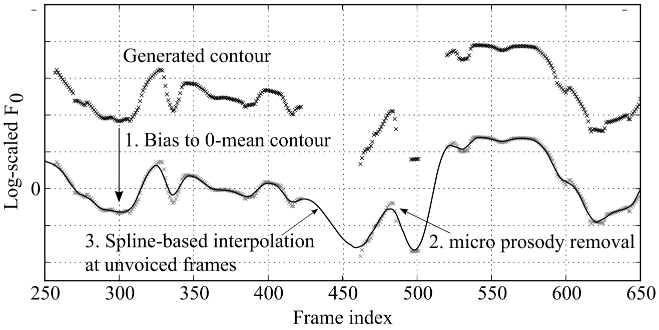


Fig. 8. An illustration of the pre-processing procedures to calculate the continuous F_0 contour from the original F_0 contour.

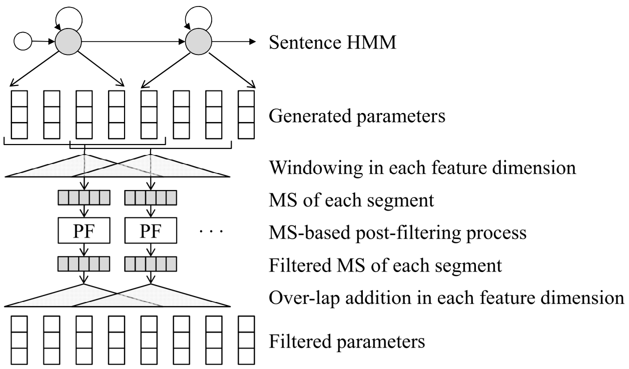

Sentence HMM
Generated parameters
Windowing in each feature dimension
MS of each segment
PF
PF
· · ·
MS-based post-filtering process
Filtered MS of each segment
Over-lap addition in each feature dimension
Filtered parameters

Fig. 9. Procedures of the segment-level MS-based post-filter in HMM-based TTS.

where $\boldsymbol{\mu}_\mathrm{s}^{(\mathrm{N})}$ and $\boldsymbol{\Sigma}_\mathrm{s}^{(\mathrm{N})}$ are the mean vector and the diagonal covariance matrix of the MS $\boldsymbol{s}$,

$$\boldsymbol{\mu}_\mathrm{s}^{(\mathrm{N})} = \left[\boldsymbol{\mu}_1^{(\mathrm{N})^\top}, \cdots, \boldsymbol{\mu}_d^{(\mathrm{N})^\top}, \cdots, \boldsymbol{\mu}_D^{(\mathrm{N})^\top}\right]^\top, \tag{10}$$

$$\boldsymbol{\Sigma}_\mathrm{s}^{(\mathrm{N})} = \mathrm{diag}\left[\boldsymbol{\Sigma}_1^{(\mathrm{N})}, \cdots, \boldsymbol{\Sigma}_d^{(\mathrm{N})}, \cdots, \boldsymbol{\Sigma}_D^{(\mathrm{N})}\right], \tag{11}$$

$$\boldsymbol{\mu}_d^{(\mathrm{N})} = \left[\mu_{d,0}^{(\mathrm{N})}, \cdots, \mu_{d,f}^{(\mathrm{N})}, \cdots, \mu_{d,D_\mathrm{s}}^{(\mathrm{N})}\right]^\top, \tag{12}$$

$$\boldsymbol{\Sigma}_d^{(\mathrm{N})} = \mathrm{diag}\left[\sigma_{d,0}^{(\mathrm{N})^2}, \cdots, \sigma_{d,f}^{(\mathrm{N})^2}, \cdots, \sigma_{d,D_\mathrm{s}}^{(\mathrm{N})^2}\right], \tag{13}$$

where $\mu_{d,f}^{(\mathrm{N})}$ and $\sigma_{d,f}^{(\mathrm{N})^2}$ are the mean and the variance of $s_d(f)$, respectively. $\boldsymbol{\lambda}_s$ is the parameter set of the MS. $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_\mathrm{s}^{(\mathrm{G})}, \boldsymbol{\Sigma}_\mathrm{s}^{(\mathrm{G})}\right)$ is also estimated in the same manner using the speech parameter sequences generated as described in Section II. To avoid the effect of the duration difference between natural and generated speech parameter sequences in HMM-based TTS, the parameter sequence is generated using the natural speech duration. In the case of GMM-based VC, temporally-aligned input speech parameter sequence $\boldsymbol{X}$ is used to generated the speech parameter sequence $\boldsymbol{y}$.

2) Synthesis Process: The following filter is applied to the generated speech parameter sequence $\boldsymbol{y}$ (see Fig. 7.):

$$\begin{aligned} s'_d(f) &= (1-k)s_d(f) \\ &+ k\left[\frac{\sigma_{d,f}^{(\mathrm{N})}}{\sigma_{d,f}^{(\mathrm{G})}}\left(s_d(f) - \mu_{d,f}^{(\mathrm{G})}\right) + \mu_{d,f}^{(\mathrm{N})}\right], \end{aligned} \tag{14}$$

where k is a post-filter emphasis coefficient 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 modified MS and original phase characteristics of the parameter sequence before filtering. The detailed procedure is shown in Table I.

We implemented the simple time-invariant filter as the yet another implementation to recover the MS, but the synthesized speech sounded discontinuous as reported in [45].

B. Application to Various Features

1) F_0 Contour: While the proposed post-filter can be directly applied to the spectral component, additional processing is required for its 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 [52] which also estimates F_0 values at the unvoiced frames. Following [53], F_0 values of the unvoiced frames are estimated with spline-based interpolation. Because the effect of micro prosody on speech quality is small [54] but the effect on the MS is not negligible, we remove it with a Low Pass Filter (LPF). Moreover, the utterance-level F_0 mean is subtracted from original F_0 values before estimating continuous F_0 contours to avoid discontinuous transitions in the zero-padding process. These procedures are shown in Fig. 8. Because spline-based methods are inappropriate for extrapolation, i.e., silence frames, 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 post-filter. 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.

2) HMM-state duration: The proposed utterance-level post-filter modifies the MS of the phoneme-level duration calculated from the state-level duration determined by Eq. (1). The phoneme-level duration sequence is filtered after excluding silence and its mean value is normalized as with the F_0 parameters. After restoring the utterance-level mean, the phoneme-

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2522655, IEEE/ACM Transactions on Audio, Speech, and Language Processing

level duration is revised if it is smaller than the number of states of the phoneme HMM. Finally, the HMM-state duration is updated by maximizing the state duration while fixing the phoneme duration to the filtered values.

C. Segment-Level Post-Filter

Because the proposed utterance-level MS-based post-filter calculates the MS utterance by utterance, the DFT length needs to be set large enough to cover various lengths of utterances. This MS calculation causes some problems: if the length of an utterance to be synthesized is longer than the previously determined DFT length, the MS can not be calculated accurately, and thus it is difficult to apply the utterance-level filtering process to a low-latency speech synthesis framework [42], [43] where frame-level or segment-level processing based on the recursive parameter generation is essential [12].

In order to handle these cases, we propose a segment-level post-filter that is effective on shorter segments. The segment-level post-filter is derived by localizing the post-filtering process as illustrated in Fig. 9. A part of the speech parameter sequence that is windowed by a triangular window with constant length is used as a segment to calculate the MS and its statistics. The window shift length is set to a half of the window length. The MS-based post-filtering process is performed segment by segment in the same manner as the trajectory-level post-filtering process. The filtered speech parameter sequence is generated by overlapping and adding the filtered segments. The Hanning window may also be used instead of the triangular window. Note that for the spectrum parameters, silence frames are removed in calculating the MS statistics to alleviate the over-fitting problem [18]. The segment-level post-filtering can be applicable to low-delay speech waveform generation. Moreover, it is possible to further implement context-dependent post-filtering.

D. Discussion

The proposed post-filters can be automatically constructed in a data-driven manner. Whereas conventional post-filtering processes [47], [55], [56], [57] requires the rule-based design [47], or manual tuning [55], the proposed post-filters enable automatic design and tuning.

Another data-driven approach is the post-filtering process to maintain the GV of the generated parameter sequence [35]. The generated speech parameters are linearly converted as follows:

$$\hat{y}_t(d) = \sqrt{\frac{\mu_d^{(\mathrm{GV,N})}}{\mu_d^{(\mathrm{GV,G})}}}\{y_t(d) - \langle y_t(d)\rangle\} + \langle y_t(d)\rangle, \qquad (15)$$

where $\mu_d^{(\mathrm{GV,N})}$ and $\mu_d^{(\mathrm{GV,G})}$ are the GV mean of the d-th dimension of the natural and synthetic speech parameters in the training data, respectively, and $\langle y_t(d)\rangle$ is the mean of the 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, temporal fluctuation of the generated speech parameters after filtering is still very different from that of natural speech. On the other hand, the proposed post-filters can recover this fluctuation because we directly consider the MS itself.

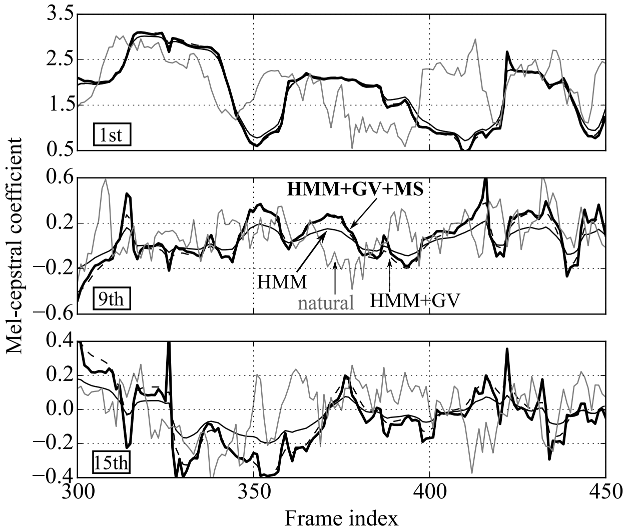


Fig. 10. An example of the 1st, 9th, and 15th mel-cepstral coefficient sequences from above in HMM-based TTS.

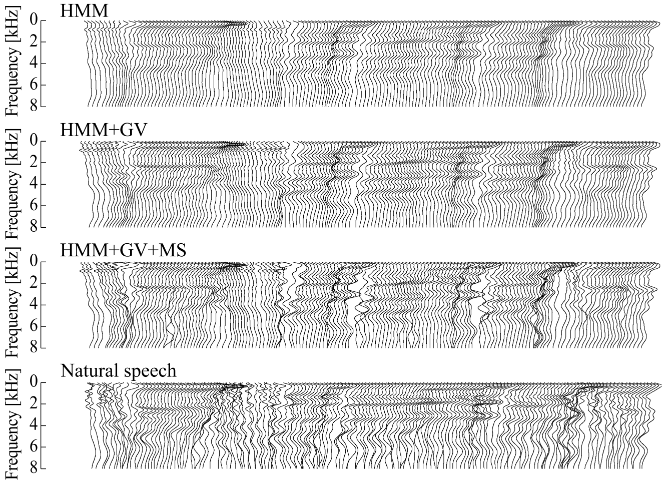


Fig. 11. An example of the spectrograms in HMM-based TTS.

According to the Parseval's theorem, the power of a temporal sequence is preserved during a DFT. The GV defined in Eq. (4) represents the power of the sequence excluding the bias component. Because the utterance-level MS is defined as the power spectrum of the sequence, the sum of the MS over all modulation frequencies excluding the bias component (frequency zero) is equal to the GV[2]. In the GV-based post-filtering process, MSs of all modulation frequencies other than the bias are converted in the same way. Namely, the GV-based post-filtering process is a special case of the proposed MS-

[2] Properly described, the sum of linear-scaled MS excluding the bias is proportional to GV.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2522655, IEEE/ACM Transactions on Audio, Speech, and Language Processing

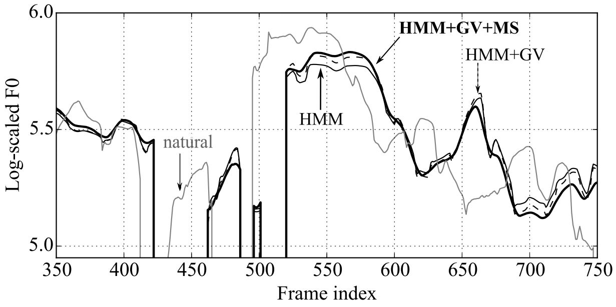


Fig. 12. An example of the F_0 contour in HMM-based TTS.

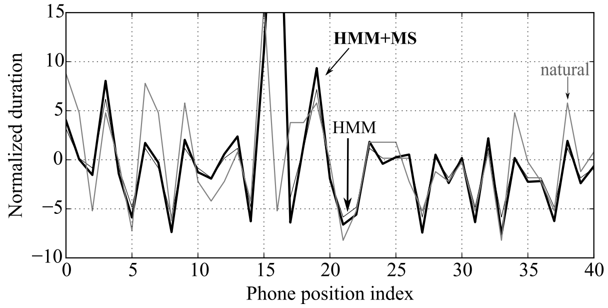


Fig. 13. An example of the phoneme-level duration in HMM-based TTS.

based post-filtering process under the following conditions:

$$\mu^{(\cdot)}_{d,f} = \log \mu^{(\mathrm{GV},\cdot)} \;\; (f > 0), \tag{16}$$

$$\mu^{(\mathrm{N})}_{d,f} = \mu^{(\mathrm{G})}_{d,f} \;\; (f = 0), \tag{17}$$

$$\sigma^{(\mathrm{N})}_{d,f} = \sigma^{(\mathrm{G})}_{d,f}, \tag{18}$$

in which the post-filter emphasis coefficient is set to 1. Namely, the GV-based post-filtering process only causes the constant MS shift as shown in Fig. 3[3]. On the other hand, the proposed methods can directly convert the MS components at individual modulation frequencies.

Figure 10 draws an example of the filtered/non-filtered mel-cepstral coefficient sequences. It is observed that the proposed post-filter generates the fluctuated parameter sequence, and the effect is larger in the higher order of the mel-cepstral coefficients. This is because the MS difference between natural and generated parameter sequences is larger in higher-order mel-cepstral coefficients as shown in Fig. 3. The effect of the proposed post-filter is also observed in the spectrogram shown in Fig. 11. We can find that the proposed post-filter produces a more fluctuated spectral sequence compared to the conventional approaches. Similarly, Fig. 12 and Fig. 13 show the F_0 contour and duration. We can also find the fluctuated parameter sequences are generated by the proposed post-filter.

Note that although these fluctuated parameter sequences are effective for improving naturalness of synthetic speech, they sometimes result in audible warbling in the synthesized speech.

V. Experimental Evaluation

First, we investigate the effects of the proposed utterance-level and segment-level post-filters from various perspectives in HMM-based TTS. Then, we evaluate them in other statistical parametric speech synthesis frameworks: the effect of the utterance-level post-filter in GMM-based VC and the effect of the segment-level post-filter in CLUSTERGEN.

A. Experimental Conditions for Evaluation in HMM-Based TTS

We trained a context-dependent phoneme Hidden Semi-Markov Model (HSMM) [60] for a Japanese female speaker for evaluation in HMM-based TTS. We used 450 sentences for training and 53 sentences for evaluation from the 503 phonetically balanced sentences included in the ATR Japanese speech database [61]. 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 [62], [63] were extracted as excitation parameters. The STRAIGHT analysis-synthesis system [27] was employed for parameter extraction and waveform generation. 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 proposed post-filter was trained in a context-independent manner. A 10 Hz-cutoff LPF was used to remove the micro prosody from the continuous F_0 contours[4].

We conducted evaluation with the following systems:

HMM: The spectrum and F_0 are generated with Eq. (2), and the HMM-state duration is determined with Eq. (1).

HMM+MS: The proposed post-filter is applied to "HMM."

HMM+GV: The spectrum and F_0 are generated with Eq. (5).

HMM+GV+MS: The proposed post-filter is applied to "HMM+GV."

Note that the post-filter of "HMM+GV+MS" was trained using parameter sequences generated with the GV. The "HMM" system was used for the components that the proposed methods were not applied to. The post-filters were not applied to the aperiodicity component because there is no quality gain achieved by the post-filters[5]. Sections V-B and V-C adopt the utterance-level and the segment-level MS-based post-filter, respectively.

B. Evaluation of Utterance-Level MS-Based Post-Filter

We investigate the effectiveness of the proposed utterance-level post-filter in HMM-based TTS. The filter emphasis coefficients for spectrum, F_0 and duration are first tuned by the likelihoods. The synthetic speech quality is then evaluated

[3] In Fig. 3, the parameter generation algorithm considering the GV rather than the GV-based post-filter is used. Although it tends to make the GV of the generated speech parameter sequence almost equal to the GV mean $\boldsymbol{\mu}_{\mathrm{v}}$ [58], [59], it still roughly results a MS shift in practical effect, although the amount of the MS shift changes utterance by utterance.

[4] We evaluated training accuracy of MS likelihood for various cutoff frequencies, and confirmed that this setting was the best.

[5] The same tendency is reported in the parameter generation algorithm considering the GV [63].

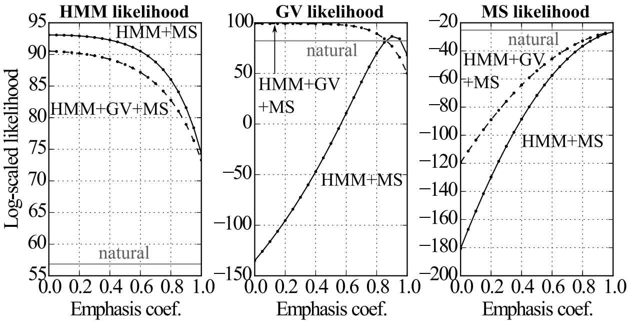


Fig. 14. HMM, GV, and MS likelihoods for the spectral parameter sequences filtered by the proposed utterance-level post-filter in HMM-based TTS.

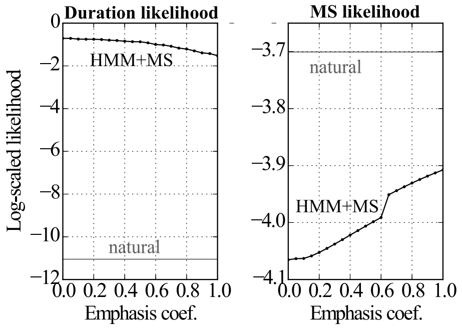


Fig. 16. HMM, GV, and MS likelihoods for the phoneme-level duration sequences filtered by the proposed utterance-level post-filter in HMM-based TTS.

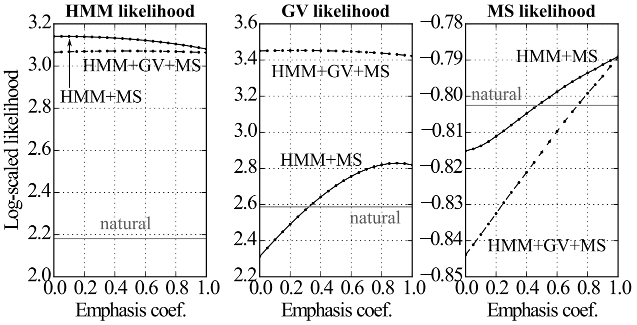


Fig. 15. HMM, GV, and MS likelihoods for the F_0 contours filtered by the proposed utterance-level post-filter in HMM-based TTS.

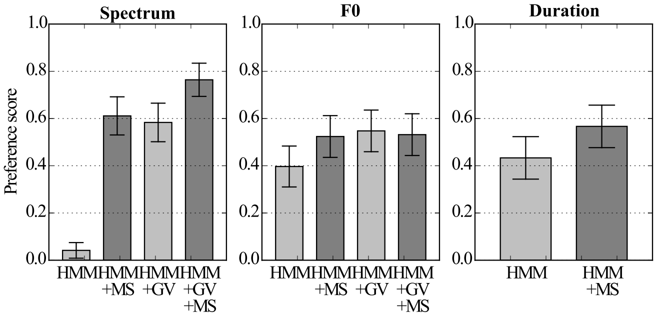


Fig. 17. Preference scores on speech quality with 95% confidence interval (proposed utterance-level post-filter).

using the tuned emphasis coefficients. The DFT length to calculate MS ($= 2D_\mathrm{s}$) was set to 4096, which is over the maximum frame length in training and evaluation data.

1) Tuning of the Emphasis Coefficients: In order to determine the filter emphasis coefficients, we calculated the HMM likelihood, GV likelihood, and MS likelihood for filtered spectrum, F_0, and HMM-state duration for settings of the emphasis coefficient from 0 to 1. The duration likelihood was calculated instead of the HMM likelihood when tuning the coefficient for duration. For comparison, the likelihood for natural speech parameter sequences was calculated, which was labeled as “natural.” Note that the HMM likelihood and the MS likelihood were normalized by the total number of frames T and one half of the DFT length D_s, respectively.

Figure 14 shows the likelihoods for the filtered spectral parameters. It is observed that the HMM likelihoods of “HMM+MS” and “HMM+GV+MS” decrease as the emphasis coefficient increases. Nevertheless, their values are always higher than that of “natural.” In the GV likelihood, we can see that these likelihoods cross that of “natural speech” at $k = 0.85$. On the other hand, MS likelihoods increase as the coefficient increases but their values always lower than “natural speech.” Considering these results, we determined the filter emphasis coefficient for spectral component to be 0.85.

Figure 15 shows the likelihoods for the filtered F_0 contour. The change of these likelihoods as the coefficient varies 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 $k = 0.75$, and we can also find that the coefficient $k = 1.0$ is the highest point of MS likelihood. From these results, we set the coefficient to 1.0.

Figure 16 shows the likelihoods for the filtered phoneme-level duration. The tendency of the likelihood change is similar to those of the spectrum and F_0, and the MS likelihood is the highest at $k = 1.0$. Therefore, we set the coefficient $k = 1.0$. We can also see discontinuous transitions of the MS likelihood. We expect that this was caused by the effect of rounding the filtered duration values into integer values after filtering.

2) Subjective Evaluation on Speech Quality: To investigate whether or not quality improvements are yielded by applying the proposed post-filter to the spectrum, F_0, and duration components, we conducted a preference AB test on speech quality. Every pair of these types of synthetic speech was presented to listeners in random order. Listeners were asked which sample sounded better in terms of speech quality. Evaluation for spectrum, F_0, and duration was conducted by 8, 8, and 6 listeners, respectively.

Figure 17 shows the preference test for the spectrum, F_0, and duration. For spectrum, we can see that the score of the “HMM+MS” system dramatically increases over the “HMM” system, and achieves a similar score to the ”HMM+GV” system. Additionally, further improvement can be observed by applying the proposed method to “HMM+GV.” From these

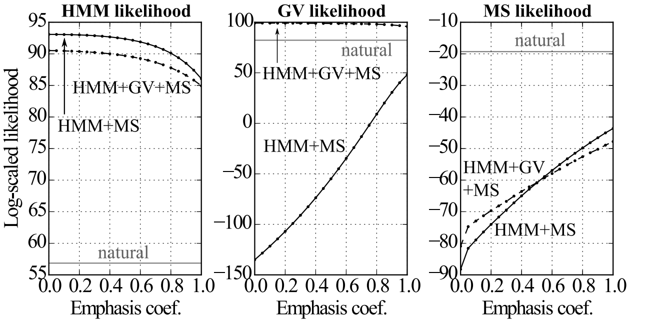


Fig. 18. HMM, GV, and MS likelihoods for the spectral parameter sequences filtered by the proposed segment-level post-filter in HMM-based TTS.

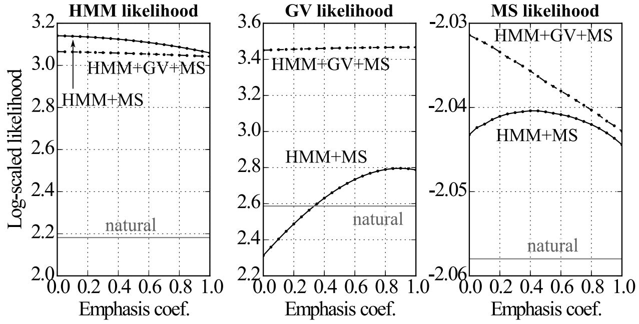


Fig. 19. HMM, GV, and MS likelihoods for the F_0 contours filtered by the proposed segment-level post-filter in HMM-based TTS.

results, the effectiveness of the proposed method for the spectral component is confirmed. For F_0, "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 the MS of the generated F_0 contours is quite close to that of the natural F_0 contours, as shown in Fig. 4, even if not applying the proposed post-filter. Finally, we can also see a slight improvement in quality for duration. These results demonstrate a quality gains by the proposed utterance-level post-filter for spectrum, F_0 and duration.

We have explained that the MS involves the GV, but it is shown that combining the GV and MS ("HMM+GV+MS") yields improvements compared to HMM+MS. This is because the post-filtering-based approach ignores the HMM probability density function. We expect that quality of parameter generation considering the MS will be comparable to that considering the GV and MS.

C. Evaluation of Segment-Level MS-Based Post-Filter

We evaluate the effectiveness of the segment-level post-filter in HMM-based TTS. The window length and window shift length were set to 125 ms (25 samples) and 60 ms (12 samples) [64]. A 64-taps DFT was used to calculate the MS. The tuning step and evaluation step were conducted in the same way as the evaluation of the proposed utterance-level post-filter. Note that the post-filter was not applied to the duration because we could not observe a large difference between filtered and non-filtered sequences.

1) Tuning the Emphasis Coefficients: The HMM likelihood, GV likelihood, and MS likelihood for the filtered spectral parameters and F_0 contours were calculated. The results are shown in Fig. 18 and Fig. 19. Their tendencies are similar to those of the utterance-level post-filter. Although the segment-level post-filtering process causes a degradation of the HMM likelihoods, they are still greater than those of natural parameters. Almost all likelihoods tend to increase as the filter coefficient approaches 1. We observed a degradation of the MS likelihood for F_0, but it is always greater than that of natural parameters. From these results, we tuned the emphasis coefficient to 1.0 for both spectrum and F_0. As the general tendency, the change of the MS likelihoods is smaller than that in the utterance-level post-filter.

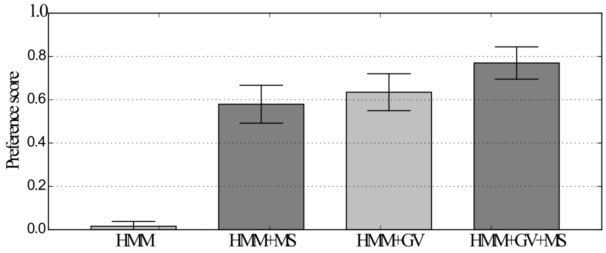


Fig. 20. Preference scores on speech quality with 95% confidence interval (proposed segment-level post-filter in HMM-based TTS).

2) Subjective Evaluation on Speech Quality: The preference AB test on speech quality by 7 listeners was conducted in the same manner as in the previous section. The post-filtering was applied to both spectrum and F_0.

The preference score is shown in Fig. 20. It is observed that a significant quality gain is yielded by "HMM+MS" compared to "HMM," and it is comparable to that yielded by "HMM+GV." Furthermore, we can see that an additional gain is yielded by "HMM+GV+MS" compared to "HMM+GV." This tendency is similar to that observed in the utterance-level post-filter. Note that the segment-level post-filter is applicable to speech parameter sequences of various lengths but the utterance-level post-filter is not.

3) Comparison of Utterance-Level and Segment-Level Post-Filters: We compare the proposed utterance-level and segment-level post-filters that are applied to "HMM+GV" for spectrum and F_0. We used the emphasis coefficients tuned in this and the previous section. The preference AB test on speech quality by 8 listeners was conducted.

Fig. 21 shows the result. Because there is no significant difference between two post-filters, we can find that the proposed post-filters have the same capability in the speech quality improvement.

D. Evaluation in Various Synthesizers

We confirm the effectiveness of the proposed post-filters in GMM-based VC and CLUSTERGEN.

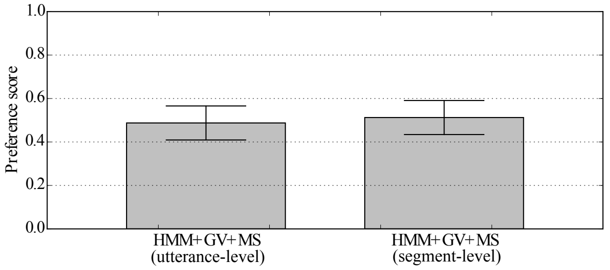


Fig. 21. Preference scores on speech quality with 95% confidence interval (proposed utterance-level and segment-level post-filters in HMM-based TTS).

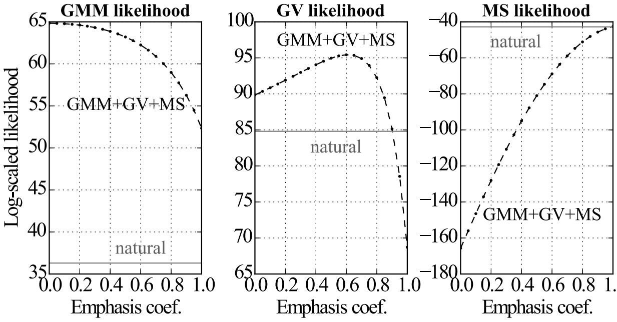


Fig. 22. GMM, GV, and MS likelihoods for the spectral parameters filtered by the proposed utterance-level post-filter in GMM-based VC.

1) GMM-Based VC: The proposed utterance-level post-filter was applied to GMM-based VC. Because this framework has a similar synthesis criterion as that described in Section II, the tuning step and evaluation step are conducted in the same manner as the evaluation for HMM-based TTS. Here, “HMM+GV” and “HMM+GV+MS” were relabeled as “GMM+GV” and “GMM+GV+MS,” respectively. The systems corresponding to “HMM” and “HMM+MS” were not used in the evaluation.

We prepared speech from two Japanese male and female speakers[6]. We selected 50 parallel sentences of subset A from the 503 phonetically balanced sentences included in the ATR Japanese speech database [61] for training, and 50 sentences of subset B for evaluation. We trained female-to-male GMMs. The speech features were the same as in the evaluations for HMM-based TTS. The spectral parameters and aperiodic components were converted with a 64-mixture GMM and a 16-mixture GMM, respectively. The log-scaled F_0 was linearly converted. The DFT length to calculate MS was set to 2048, which is over the maximum frame length in the training and evaluation data. The proposed utterance-level post-filter was applied to the spectral parameters.

The GMM likelihood, GV likelihood, and MS likelihood for the filtered spectral parameters were shown in Fig. 22. From this result, we can see that the tendency of the likelihood changes is almost the same as that in Fig. 14, but the GV likelihood of “GMM+GV+MS” starts to fall below “natural” at the emphasis coefficient $k = 0.90$. Therefore, the emphasis coefficient is set to 0.90.

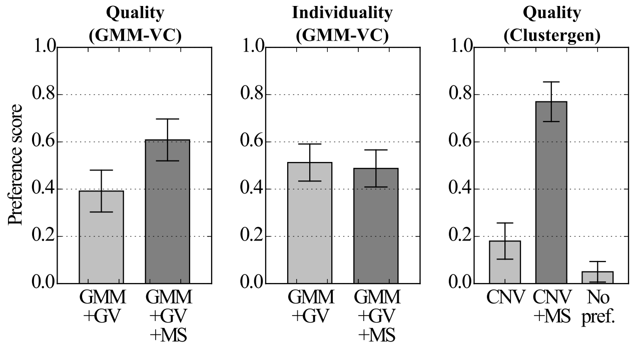


Fig. 23. Preference scores on speech quality with 95% confidence interval in GMM-based VC and CLUSTERGEN

We conducted a preference AB test on speech quality, and a preference XAB test on speaker individuality. We first presented an analysis-synthesized reference speech as ”X”, then we presented random-ordered synthesized speech. 7 listeners participated in each evaluation. Fig. 23 shows the results. In term of speech quality, a significant quality gain is observed. However, there is no significant difference in the preference score on speaker individuality. We expect that no cues for individuality are at higher modulation frequencies that are recovered by the MS-based post-filter.

2) CLUSTERGEN: The proposed segment-level post-filter was also applied to CLUSTERGEN. We also tuned the emphasis coefficient as in the previous experiments. We observed that the likelihoods didn’t vary very much as shown in Figs. 18 and 19. We also confirmed that a quality gain was yielded by setting k to 1.0. Here, the methods corresponding to “HMM” and “HMM+MS” were relabeled as “CNV” and “CNV+MS,” respectively.

We prepared an English female speaker. 418 and 46 sentences of news reader speech were used for training and evaluation, respectively. The speech features were the same as those in the evaluation for HMM-based TTS, but they were extracted by Speech signal Processing ToolKit (SPTK) [65] and the aperiodicity component was not used. The window length and window shift length of the segment-level post-filter were set to 125 ms (25 samples) and 60 ms (12 samples). A 64-taps DFT was used to calculate the MS. The segment-level post-filter was applied to both spectrum and F_0. parameters.

A preference AB test on speech quality was conducted by 6 listeners on the Amazon Mechanical Turk service [66]. Because many listening environments are expected, a no preference option was prepared. The right side of Fig. 23 shows the result. We can see that large improvements are yielded by the segment-level post-filter.

The results presented in this section suggest that the proposed MS-based post-filters are effective for a variety of statistical parametric speech synthesis frameworks.

[6] The female speaker here is a different person from the speaker we used in the evaluation for HMM-based TTS.

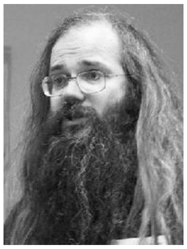
VI. Conclusion

This paper introduced the Modulation Spectrum (MS) of speech parameter trajectory as a new feature to effectively quantify the over-smoothing effect, which is the main cause of the synthetic speech quality degradation in statistical parametric speech synthesis. Moreover, this paper also proposed the MS-based post-filters on the utterance level and the segment level to improve the synthetic speech quality. Experimental evaluation was conducted using various statistical parametric speech synthesis methods, such as Hidden Markov Model (HMM)-based Text-To-Speech (TTS), Gaussian Mixture Model (GMM)-based Voice Conversion (VC), and Classification And Regression Trees (CART)-based TTS (a.k.a., CLUSTERGEN). The experimental results demonstrated that (1) the proposed utterance-level post-filter achieves better quality for spectrum, F_0, and HMM-state duration in HMM-based TTS, (2) the proposed segment-level post-filter capable of achieving low-delay synthesis also yields significant improvements in synthetic speech quality, and (3) the proposed post-filters are also effective in not only HMM-based TTS but also GMM-based VC and CLUSTERGEN. As future work, we plan to investigate which modulation frequency bands significantly affect synthetic speech quality, and integrate the MS into the speech synthesis metric.

Acknowledgements: Part of this work was supported by JSPS KAKENHI Grant Number 26280060 and Grant-in-Aid for JSPS Fellows Grant Number $26 \cdot 10354$, and part of this work was executed under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation."

Shinnosuke Takamichi received his B.E. from Nagaoka University of Technology, Japan, in 2011 and his M.E. degree from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan, in 2013. He was a short-time researcher at the NICT, Kyoto, Japan in 2013, and a visiting researcher of Carnegie Mellon University (CMU), in United States, from 2014 to 2015. He is currently a Ph.D. student of NAIST, and Research Fellow (DC2) of Japan Society for the Promotion of Science, Japan. He received the 7th Student Presentation Award from ASJ, the 35th Awaya Prize Young Researcher Award from ASJ, the 8th Outstanding Student Paper Award from IEEE Japan Chapter SPS, the Best Paper Award from APSIPA ASC 2014, the Student Paper Award from IEEE Kansai Section, the 30th TELECOM System Technology Award from TAF, and the 2014 ISS Young Researcher's Award in Speech Field from the IEICE. His research interests include electroacoustics, signal processing, and speech synthesis. He is a student member of ASJ and IEEE SPS, and a member of ISCA.

Tomoki Toda earned his B.E. degree from Nagoya University, Aichi, Japan, in 1999 and his M.E. and D.E. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003, respectively. He is a Professor at the Information Technology Center, Nagoya University. He was a Research Fellow of JSPS from 2003 to 2005. He was then an Assistant Professor (2005-2011) and an Associate Professor (2011-2015) at the Graduate School of Information Science, NAIST. His research interests include statistical approaches to speech processing. He received more than 10 paper/achievement awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal).



Alan W. Black is a Professor in the Language Technologies Institute at Carnegie Mellon University. Before joining the faculty at CMU in 1999, he worked in the Centre for Speech Technology Research at the University of Edinburgh, and before that at ATR in Japan. He is one of the principal authors of the free software Festival Speech Synthesis System, the FestVox voice building tools and CMU Flite, a small footprint speech synthesis engine, that is the basis for many research and commercial systems around the world. He also works in spoken dialog systems, the LetsGo Bus Information project and mobile speech-to-speech translation systems. Prof Black is an elected member of ISCA board (20072015). He has over 200 refereed publications and is one of the highest cited authors in his field.

Graham Neubig received his B.E. from University of Illinois, Urbana-Champaign in 2005, and his M.S. and Ph.D. in informatics from Kyoto University in 2010 and 2012 respectively. From 2012, he has been an assistant professor at the Nara Institute of Science and Technology, where he is pursuing research in machine translation and spoken language processing.

Sakriani Sakti received her B.E degree in Informatics (cum laude) received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR and U-STAR (2006-2011). In 2009-2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. From 2011, she has been an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015-2016, under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation". She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Satoshi Nakamura is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He also serves as a visiting professor of Collaborative Research Unit, National Institute of Informatics. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampoli Award 2012. He organized the International Workshop of Spoken Language Translation (IWSLT 2006) and Oriental Cocosda 2008 as a general chair. He also served as the program chair of INTERSPEECH 2010. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.

References

[1] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal diabilities: Voice banking and reconstruction," *Acoust. Sci. technol.*, vol. 33, pp. 1–5, 2012.

[2] 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*, Florence, Italy, May 2014, pp. 4521–4525.

[3] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system," in *Proc. APSIPA ASC*, Hollywood, U.S.A., Nov. 2012.

[4] 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*, Florence, Italy, Aug. 2011, pp. 2769–2772.

[5] P. K. Muthukumar and A. W. Black, "Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014.

[6] S. Sitaram, G. Anumanchipalli, J. Chiu, A. Parlikar, and A. W. Black, "Text to speech in new languages without a standardized orthography," in *Proc. SSW8*, Barcelona, Spain, Aug. 2013.

[7] 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*, Florence, Italy, May 2014, pp. 7954–7958.

[8] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.

[9] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

[10] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1988.

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

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

[13] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[14] 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.

[15] A. W. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Proc. INTERSPEECH*, Pittsburgh, U.S.A., Sep. 2006.

[16] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical parametric speech synthesis based on gaussian process regression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 173–183, Apr. 2014.

[17] E. Helander, T. V. H. Silen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, Mar. 2012.

[18] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 3872–3876.

[19] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 369–372.

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

[21] K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An investigation of implementation performance analysis of DNN based speech synthesis system," in *Proc. INTERSPEECH*, Brighton, U. K., 2014, pp. 577–582.

[22] 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.

[23] 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.

[24] L. Chen, M. J. F. Gales, L. Chen, K. Chin, K. Knull, and M. Akamine, "Exploring rich expressive information from audiobook data using cluster adaptive training," in *Proc. INTERSPEECH*, Portland, U.S.A., Sep. 2012.

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

[26] Y. Stylianou, "Voice transformation: A survey," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3585–3588.

[27] 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.

[28] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," in *Proc. SMAC*, Stockholm, Aug. 2013.

[29] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4230–4234.

[30] P. K. Muthukumar, A. W. Black, and H. T. Bunnell, "Optimizations and fitting procedures for the Liljencrants-Fant model for statistical parametric speech synthesis," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[31] 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.

[32] S. Takamichi, T. Toda, Y. Shiga, S. Sakti, G. Neubig, and S. Nakamura, "Parameter generation methods with rich context models for high-quality and flexible text-to-speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 239–250, 2014.

[33] T. Nose, V. Chunwijitra, and T. Kobayashi, "A parameter generation algorithm using local variance for HMM-based speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 221–228, 2014.

[34] M. Shannon and W. Byrne, "Fast, low-artifact speech synthesis considering global variance," in *Proc. ICASSP*, Vancouver, Canada, May. 2013, pp. 7869–7873.

[35] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," in *Proc. INTERSPEECH*, Portland, U.S.A., Sept. 2012.

[36] 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.

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

[38] S. Gergen, A. Nagathil, and R. Martin, "Reduction of reverberation effects in the MFCC modulation spectrum for improved classification of acoustic signals," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 1992–1995.

[39] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proc. ICASSP*, Vancouver, Canada, May. 2013, pp. 7234–7238.

[40] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech perception," *J. Acoust. Soc. of America*, vol. 95, pp. 1053–1064, 1994.

[41] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proc. ICSLP*, vol. 4, 1996, pp. 2490–2493.

[42] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," in *Proc. INTERSPEECH*, Brisbane, Australia, Sep. 2008, pp. 1076–1079.

[43] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4470–4474.

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

[45] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modified modulation spectrum-based post-filter for HMM-based speech synthesis," in *Proc. GlobalSIP*, Atlanta, United States, Dec. 2014, pp. 710–714.

[46] ——, "Modulation spectrum-based post-filter for GMM-based voice conversion," in *Proc. APSIPA ASC*, Siem Reap, Cambodia, Dec. 2014.

[47] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, Budapest, Hungary, Apr. 1999, pp. 2347–2350.

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

[49] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn.(E)*, vol. 28, no. 3, pp. 140–146, 2007.

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

[51] S. Pan, J. Tao, and Y. Wang, "A state duration generation algorithm considering global variance for HMM-based speech synthesis," in *Proc. APSIPA ASC*, Xi'an, China, 2011.

[52] 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.

[53] 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*, Lyon, France, Sep. 2013, pp. 3067–3071.

[54] P. Taylor, *Text-To-Speech Synthesis*. Cambridge Univ. Press, 2009.

[55] 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*, Florence, Italy, May 2014, pp. 275–279.

[56] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4455–4459.

[57] L.-H. Chen, T. Raitio, C. V.-Botinhao, J. Yamagishi, and Z.-H. Ling, "DNN-based stochastic postfilter for HMM-based speech synthesis," in *Proc. INTERSPEECH*, MAX Atria, Singapore, May 2014, pp. 1954–1958.

[58] H. Zen, M. J. F. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *IEEE Trans. on Audio, Speech, and Language processing*, vol. 20, no. 3, pp. 794–805, Mar. 2011.

[59] T. Nose and A. Ito, "Analysis of spectral enhancement using global variance in HMM-based speech synthesis," in *Proc. INTERSPEECH*, MAX Atria, Singapore, May 2014, pp. 2917–2921.

[60] H. Zen, K. Tokuda, T. K. 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.

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

[62] H. Kawahara, J. 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*, Firentze, Italy, Sept. 2001, pp. 1–6.

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

[64] V. Tyagi, I. McCowan, H. Misra, and H. Bourlard, "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR," in *Proc. ASRU*, MAX Atria, Singapore, Nov. 2003, pp. 399–404.

[65] "Speech signal processing toolkit (SPTK) http://sp-tk.sourceforge.net/."

[66] "Amazon mechanical turk https://www.mturk.com/."