Postfilters 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, Sakriani Sakti, Member, IEEE, and 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, the 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 postfilter 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 postfilter 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

A LTHOUGH 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 [10], [12], 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 [18], [30], 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) [14], [31], 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 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 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 speech synthesizers: the utterance-level post-filter in GMM-based TTS from various perspectives. Then, we evaluate them in other

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

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.

A. HMM-Based TTS [13]

Figure 1 illustrates an overview of HMM-based TTS. The HMM models the $T$-frame output speech feature sequence $Y = [Y_1^T, \ldots, Y_t^T, \ldots, Y_T^T]^\top$ given the contextual factor sequence $X$ of the input text, where $Y_t$ 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 $X$, the corresponding HMM is first constructed, then the HMM-state sequence $\hat{q} = [\hat{q}_1, \ldots, \hat{q}_t, \ldots, \hat{q}_T]$ is determined by maximizing the duration probability density function as follows:

$$
\hat{q} = \operatorname{argmax}_q P (q | X, \lambda),
$$

where $\hat{q}_t$ is a HMM-state at frame $t$, and $\lambda$ 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.

$$
\hat{y} = \operatorname{argmax}_y P (Wy | q, X, \lambda),
$$

where $\hat{y} = [y_1^T, \ldots, y_t^T, \ldots, y_T^T]^\top$ is a speech parameter vector sequence of $T$ frames, $y_t = [y_t(1), \ldots, y_t(d), \ldots, y_T(D)]^\top$ is a $D$-dimensional speech feature vector at frame $t$, $d$ is a dimension index, and $W$ is the weighting matrix for calculating the dynamic features [11], where $Y = Wy$. 

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on February 20,2020 at 13:14:26 UTC from IEEE Xplore. Restrictions apply.
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

\[ v(y) = [v(1), \ldots, v(d), \ldots, v(D)]^T, \quad (3) \]

\[ v(d) = \frac{1}{T} \sum_{t=1}^{T} (y_t - \frac{1}{T} \sum_{\tau=1}^{T} y_{\tau}(d))^2. \quad (4) \]

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

\[ \hat{y} = \arg\max_y P(W_y|q, X, \lambda) P(v(y)|\lambda) w_v, \quad (5) \]

where \( \lambda \) 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 \( X \) and \( Y \). The Dynamic Time Warping (DTW) algorithm is used to temporally align \( X \) and \( Y \). In synthesis, mixture component sequence \( \hat{q} \) is determined similarly to Eq. (1), where \( q_t \) is a mixture component at frame \( t \). The synthetic speech parameter sequence \( \hat{y} \) is determined in the same manner as Eq. (5), where \( P(W_y|q, X, \lambda) \) 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 \( X \). The synthetic speech parameter sequence \( \hat{y} \) 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 \( s(y) \) of the parameter sequence \( y \) is calculated as:

\[ s(y) = \left[ s(1)^T, \ldots, s(d)^T, \ldots, s(D)^T \right]^T, \quad (6) \]

\[ s(d) = [s_d(0), \ldots, s_d(f), \ldots, s_d(D_s)]^T, \quad (7) \]

**Fig. 2.** Graphic representation of how to derive the MS \( s(y) \) from the speech parameter sequence \( y \).

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

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

where \( f \) is a modulation frequency index, \( m = -\pi f/D_s \) 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 \( 2D_s \). As shown in Fig. 2, \( s(y) \) 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.
10 Hz components are shown.

Section II. To avoid the effect of the duration difference affected by the over-smoothing effect due to the statistical averaging process implicit in conventional parameter generation, as described for spectrum, synthesis is less than 2 and it is assumed that the sequence length in the training and

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 \( 2D_s \), and it is assumed that the sequence length in the training and synthesis is less than \( 2D_s \).

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

\[
P(s|\lambda_s) = \mathcal{N}(s; \mu^{(N)}_s, \Sigma^{(N)}_s) ,
\]

where \( \mu^{(N)}_s \) and \( \Sigma^{(N)}_s \) are the mean vector and the diagonal covariance matrix of the MS \( s \).

\[
\mu^{(N)}_s = \left[ \mu^{(N)}_1, \ldots, \mu^{(N)}_d, \ldots, \mu^{(N)}_D \right]^\top ,
\]

\[
\Sigma^{(N)}_s = \text{diag} \left[ \Sigma^{(N)}_1, \ldots, \Sigma^{(N)}_d, \ldots, \Sigma^{(N)}_D \right] ,
\]

\[
\mu^{(D)}_d = \left[ \mu^{(D)}_{d,0}, \ldots, \mu^{(D)}_{d,f}, \ldots, \mu^{(D)}_{d,D_s} \right]^\top ,
\]

\[
\Sigma^{(D)}_d = \text{diag} \left[ \sigma^{(D)}_{d,0}^2, \ldots, \sigma^{(D)}_{d,f}^2, \ldots, \sigma^{(D)}_{d,D_s}^2 \right] .
\]

where \( \mu^{(D)}_d \) and \( \sigma^{(D)}_{d,f}^2 \) are the mean and the variance of \( s_d(f) \), respectively. \( \lambda_s \) is the parameter set of the MS. \( \mathcal{N}(\cdot; \mu^{(G)}_s, \Sigma^{(G)}_s) \) 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 that the generated MS is generally smaller than that of natural speech.
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-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 cannot 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].

\(^1\)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.
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.

\[ \mu_d^{(f)} = \log \mu_d^{(GV, f)} \quad (f > 0), \]
\[ \mu_d^{(N)} = \mu_d^{(GV, f)} \quad (f = 0), \]
\[ \sigma_d^{(N)} = \sigma_d^{(GV, f)}. \]

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. 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

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

2 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 \( \mu_k \) [58], [59], it still roughly results a MS shift in practical effect, although the amount of the MS shift changes utterance by utterance.
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

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.

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. 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 using the tuned emphasis coefficients. The DFT length to calculate MS ($=2D_s$) 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

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

The same tendency is reported in the parameter generation algorithm considering the GV [63].
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 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
TAKAMICHI et al.: POSTFILTERS TO MODIFY MODULATION SPECTRUM

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.

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.

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. 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.

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 terms 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.

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 experiment 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.

REFERENCES


Shinnosuke Takamichi (S’14) received the B.E. degree from Nagaoka University of Technology, Nagaoka, Japan, in 2011, and the M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Ikoma, Japan, in 2013. He is currently pursuing the Ph.D. degree at NAIST. He was a short-time Researcher with the NICT, Kyoto, Japan, in 2013, and a Visiting Researcher of Carnegie Mellon University (CMU), in USA, from 2014 to 2015. He is a Research Fellow (DC2) of Japan Society for the Promotion of Science, in Japan. His research interests include electroacoustics, signal processing, and speech synthesis. He is a student member of ASJ and a member of ISCA. He was the recipient of the 7th Student Presentation Award from ASJ, the 35th Awaya Prize Young Researcher Award from ASJ, the 8th Outstanding Student Paper Award from the IEEE Japan Chapter SPS, the Best Paper Award from APSIPA ASC 2014, the Student Paper Award from the 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.

Tomoki Toda (M’05) received the B.E. degree from Nagoya University, Aichi, Japan, in 1999, and the M.E. and Ph.D. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003, respectively. He is a Professor with 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) with the Graduate School of Information Science, NAIST. His research interests include statistical approaches to speech processing. He was the recipient of 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 (M’03) is a Professor with the Language Technologies Institute, Carnegie Mellon University. Before joining the faculty at CMU in 1999, he was with the Centre for Speech Technology Research, University of Edinburgh, Edinburgh, U.K., 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 3D voice-to-speech translation systems. He is an elected member of ISCA Board (2007015). He has over 200 refereed publications and is one of the highest cited authors in his field.

Graham Neubig received the B.E. degree from the University of Illinois, Urbana-Champaign, Champaign, IL, USA, in 2005, and the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2010 and 2012, respectively. Since 2012, he has been an Assistant Professor with the Nara Institute of Science and Technology, Ikoma, Japan, where he is pursuing research in machine translation and spoken language processing.
Sakriani Sakti (M’13) received the B.E. degree (cum laude) in informatics from Bandung Institute of Technology, Bandung, Indonesia, in 1999, the M.Sc. degree in communication technology from the University of Ulm, Ulm, Germany, in 2002, and the Ph.D. degree from Dialog Systems Group, University of Ulm, in 2008. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003 and 2009, she worked as a Researcher with ATR SLC Labs, Japan, and from 2006 to 2011, she was with ATR-NICT, Japan. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003–2007), A-STAR, and U-STAR (2006–2011). Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework. From 2009 to 2011, she served as a Visiting Professor at Department of Computer Science, University of Indonesia (UI), Depok, Indonesia. Since 2011, she has been an Assistant Professor with the Augmented Human Communication Laboratory, NAIST, Ikoma, Japan. She served also as a Visiting Scientific Researcher at INRIA Paris-Rocquencourt, France, from 2015 to 2016, under “JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation.” She is a member of JNS, SFN, ASJ, ISCA, and IEICE. In 2000, she was the recipient of the DAAD-Siemens Program Asia 21st Century Award.

Satoshi Nakamura (F’16) received the B.S. degree from Kyoto Institute of Technology, Kyoto, Japan, in 1981, and the Ph.D. degree from Kyoto University, Kyoto, Japan, in 1992. He is Professor with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He was an Associate Professor with the Graduate School of Information Science, Nara Institute of Science and Technology, from 1994 to 2000. He was the Director of ATR Spoken Language Communication Research Laboratories from 2000 to 2008 and the Vice president of ATR from 2007 to 2008. He was the Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan, from 2009 to 2010. He is currently the Director of Augmented Human Communication laboratory and a Full Professor with the Graduate School of Information Science, Nara Institute of Science and Technology. He also serves as a Visiting Professor of Collaborative Research Unit, National Institute of Informatics. 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. His research interests include modeling and systems of speech-to-speech translation and speech recognition. He organized the International Workshop of Spoken Language Translation (IWSLT 2006) and Oriental Cocoda 2008, as the 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, the IEEE Signal Processing Magazine Editorial Board Member since April 2012, the IEEE SPS Speech and Language Technical Committee Member since 2013. He was the recipient of the Yamashita Research Award, the Kiyasu Award from the Information Processing Society of Japan, the Telecom System Award, the AAMT Nagao Award, the Docomo Mobile Science Award in 2007, the 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 was also the recipient of the LREC Antonio Zampoli Award 2012.