MODULATION SPECTRUM COMPENSATION FOR HMM-BASED SPEECH SYNTHESIS USING LINE SPECTRAL PAIRS

Zhen-Hua Ling, Xiao-Hui Sun, Li-Rong Dai, Yu Hu

National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R.China

zhling@ustc.edu.cn, sunxh06@mail.ustc.edu.cn, lrdai@ustc.edu.cn, yuhu@iflytek.com

ABSTRACT

In previous work, a method to compensate the divergence between the distributions of natural and generated modulation spectra (MS) has been proposed for hidden Markov model (HMM) based speech synthesis. This method can alleviate the over-smoothing effect of parameter generation when mel-cepstral coefficients (MCC) are used as spectral features. This paper further investigates the MS compensation method for line spectral pairs (LSP). Four approaches to extract MS from LSPs are implemented and compared. These approaches calculate MS vectors using original LSP sequences, log power spectra (LPS) derived from LSPs, MCCs derived from LSPs, and MCCs derived from speech waveforms, respectively. Experimental results show that the naturalness of synthetic speech gets improved after MS compensation when LSPs are used as spectral features for HMM modeling. The degree of improvement depends on the type of spectral features for MS calculation significantly. MCCs derived from LSPs are more suitable for MS compensation than original LSPs and LPS derived from LSPs. Besides, using MCCs derived from speech waveforms also achieves satisfactory performance. This means that MS compensation can also be implemented as a post-filter to synthetic waveforms which does not rely on the type of spectral features and vocoders adopted in the synthesis system.

Index Terms— Speech synthesis, hidden Markov model, modulation spectrum, line spectral pair

1. INTRODUCTION

Hidden Markov model (HMM)-based speech synthesis has become a mainstream speech synthesis method in the last two decades [1, 2]. In this method, the spectral, F0 and segmental duration features are modeled simultaneously within a unified HMM framework [1]. At synthesis time, the acoustic features predicted from the estimated HMMs are sent into a vocoder to reconstruct speech waveforms. This method is able to synthesize highly intelligible and smooth speech sounds [3, 4]. However, the quality of its synthetic speech is degraded. One reason is that the generated spectral features are over-smoothed due to the conventional maximum output probability parameter generation algorithm [2].

This work is partially funded by the National Nature Science Foundation of China (Grant No.61273032), the Fundamental Research Funds for the Central Universities (Grant No. WK2350000001), the CAS Strategic Priority Research Program (Grant No. XDB02070006), and the Electronic Industry Development Fund of Ministry of Industry and Information Technology (Grant No. [2014]425).

In order to alleviate the over-smoothing effect, many improved parameter generation methods have been proposed, such as post-filtering after parameter generation [4, 5], modifying the parameter generation criterion by integrating a global variance (GV) model [6] or minimizing model divergences [7], etc. Recently, the modulation spectrum (MS) of an acoustic feature trajectory has been adopted as a new measurement to effectively capture the over-smoothing effect. A method of MS compensation has been proposed in [8] to alleviate the over-smoothing effect of parameter generation. At training stage of this method, the MS of natural feature sequences and generated feature sequences are extracted and modeled by Gaussian distributions respectively. At synthesis time, the divergences between these two distributions are compensated by modifying the MS of generated feature sequences. Experimental results show that this method can improve the naturalness of synthetic speech effectively. Further, the MS measurement has be successfully incorporated into joint parameter generation [9] and trajectory training of HMMs [10].

All the work in [8, 9, 10] adopts mel-cepstral coefficients (MCC) as spectral features for HMM training and MS calculation. Line spectral pair (LSP) is another popularly used spectral feature in HMM-based speech synthesis [4, 11]. Previous work on GV-based parameter generation [12] shows that the effectiveness of some over-smoothing measurements, such as GV, may depend on the type of spectral features used to derive them. Therefore, it is worthwhile to investigate the MS compensation method for HMM-based speech synthesis using LSPs. Four different approaches are implemented and compared in this paper. First, the conventional MS compensation method is applied to LSPs instead of MCCs directly. Second, inspired by our previous work in [12], log power spectra (LPS) derived from LSPs are used for MS calculation and compensation. Third, the LPS derived from LSPs are further transformed into MCCs for MS extraction. Fourth, a spectral-feature-independent and vocoder-independent MS compensation method is proposed, where the MCCs calculated from speech waveforms by short-time Fourier transform (STFT) are adopted for MS calculation and compensation. Experimental results demonstrate the superiority of calculating MS from MCCs, which will also be discussed from the aspect of neural mechanism for auditory perception in this paper.

This paper is organized as follows. Section 2 briefly reviews the existing MS compensation method. Section 3 introduces our proposed methods of MS compensation for LSPs. The experimental results are given in Section 4 and Section 5 is the conclusion.

2. CONVENTIONAL MS COMPENSATION METHOD

For an acoustic feature sequence $\boldsymbol{c} = [\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, ..., \boldsymbol{c}_T^\top]^\top$ where T is the number of frames, $\boldsymbol{c}_t = [c_{t,1}, c_{t,2}, ..., c_{t,D}]^\top$, and D is the

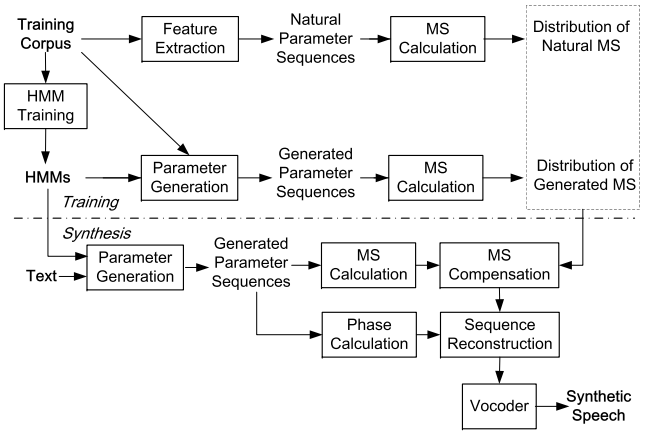

Fig. 1. Flowchart of the conventional MS compensation method.

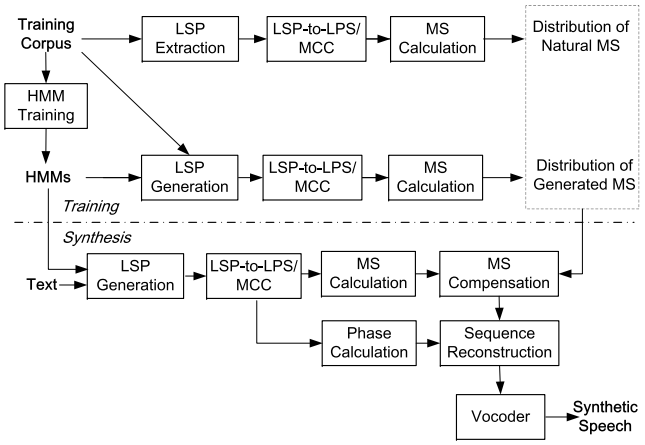

Fig. 2. Flowchart of the proposed MS compensation method for LSPs with LSP-to-LPS and LSP-to-MCC transformation.

dimension of feature vector at each frame, its MS is defined as

$$\boldsymbol{s}(\boldsymbol{c}) = [\boldsymbol{s}(\boldsymbol{c})_1, \boldsymbol{s}(\boldsymbol{c})_2, ..., \boldsymbol{s}(\boldsymbol{c})_D]^\top, \tag{1}$$

where $\boldsymbol{s}(\boldsymbol{c})_d = [s(\boldsymbol{c})_{d,1}, s(\boldsymbol{c})_{d,2}, ..., s(\boldsymbol{c})_{d,M}]^\top$ is the power spectrum of vector $[c_{1,d}, c_{2,d}, ..., c_{T,d}]^\top$ calculated by $2M$-point DFT.

The flowchart of the conventional MS compensation method [8] is shown in Fig. 1. At training stage, the acoustic parameter sequences of the sentences in the training corpus are first generated using the estimated HMMs and the conventional parameter generation algorithm [2]. Then, an MS vector can be extracted from the acoustic feature sequence of each natural and synthetic utterance following (1). Finally, two Gaussian distributions

$$p(\boldsymbol{s}(\boldsymbol{c})|\lambda^{(N)}) = \mathcal{N}\left(\boldsymbol{s}(\boldsymbol{c}); \boldsymbol{\mu}^{(N)}, \boldsymbol{\Sigma}^{(N)}\right), \tag{2}$$

$$p(\boldsymbol{s}(\boldsymbol{c})|\lambda^{(G)}) = \mathcal{N}\left(\boldsymbol{s}(\boldsymbol{c}); \boldsymbol{\mu}^{(G)}, \boldsymbol{\Sigma}^{(G)}\right) \tag{3}$$

are estimated using the natural and generated MS vectors, where $\boldsymbol{\mu}^{(N)} = [\mu_{1,1}^{(N)}, ..., \mu_{D,M}^{(N)}]^\top$ and $\boldsymbol{\mu}^{(G)} = [\mu_{1,1}^{(G)}, ..., \mu_{D,M}^{(G)}]^\top$ are mean vectors, $\boldsymbol{\Sigma}^{(N)} = diag\{(\sigma_{1,1}^{(N)})^2, ..., (\sigma_{D,M}^{(N)})^2\}$ and $\boldsymbol{\Sigma}^{(G)} = diag\{(\sigma_{1,1}^{(G)})^2, ..., (\sigma_{D,M}^{(G)})^2\}$ are diagonal covariance matrices.

At synthesis time, acoustic parameter sequences of the input text are first generated using the estimated HMMs and the context features extracted by text analysis. Then, the MS of generated parameter sequences $\boldsymbol{s}(\boldsymbol{c})$ are calculated by DFT and the phase components are also preserved. A post-filter is designed to modify $\boldsymbol{s}(\boldsymbol{c})$ in order to compensate the divergence between the two estimated distributions in (2) and (3). The modified MS $\boldsymbol{s}(\boldsymbol{c})'$ is calculated as

$$s(\boldsymbol{c})'_{d,m} = \alpha \left[\frac{\sigma_{d,m}^{(N)}}{\sigma_{d,m}^{(G)}}\left(s(\boldsymbol{c})_{d,m} - \mu_{d,m}^{(G)}\right) + \mu_{d,m}^{(N)}\right] + (1-\alpha)s(\boldsymbol{c})_{d,m}, \tag{4}$$

where α is an interpolation coefficient which controls the degree of post-filtering. Using the modified MS $s(\boldsymbol{c})'_{d,m}$ and the preserved phase spectra, the acoustic parameter sequences can be reconstructed. Finally, these sequences are sent into a vocoder to recover speech waveforms. Experimental results in [8] show that this MS compensation method is effective in alleviating the over-smoothing effects and improving the naturalness of synthetic speech when applied to MCC and F0 sequences.

3. MS COMPENSATION FOR LSPS

In this paper, we apply MS compensation to HMM-based speech synthesis when LSPs are used as spectral features. After training HMMs using LSPs as spectral observations, it is straightforward to calculate MS from natural and generated LSP sequences and apply the conventional MS compensation method introduced in Section 2 to LSPs directly. The training and synthesis process is the same as the one shown in Fig. 1. In our previous work on GV-based parameter generation [12], modeling GV vectors of LPS derived from LSPs achieved significantly better performance than modeling GV vectors of LSPs directly. Here, we adopt similar idea to investigate MS compensation for LSPs considering that GV can be approximately considered as a simplified representation of MS [8]. Three spectral features derived from LSPs or waveforms are utilized for MS compensation in this paper.

3.1. LPS derived from LSPs

The training and synthesis process of this method is shown in Fig. 2. At training time, the LSPs extracted from natural recordings and the LSPs generated from estimated HMMs are firstly transformed into log power spectrum (LPS) sequences according to the definition of LSPs [12]. The LPS of one frame is a K-dimension spectral envelope where K is the number of sampling points within frequency range $[0, \pi]$. Then, MS vectors are calculated from these derived LPS sequences following the method introduced in Section 2 and the distributions of natural and generated MS can be estimated. At synthesis time, predicted LSPs are converted into LPS at first. The MS of these derived LPS are modified according to (4). Finally, new LPS parameters are reconstructed from the modified MS, and are converted back into LSPs or sent into vocoder directly for waveform synthesis.

3.2. MCCs derived from LSPs

Using MCCs derived from LSPs for MS compensation is similar to using LPS derived from LSPs and follows the flowchart in Fig. 2. The difference is that LSPs are converted into MCCs instead of LPS for MS calculation. The conversion from LSPs to MCCs is achieved by first transforming LSPs into LPS and then transforming

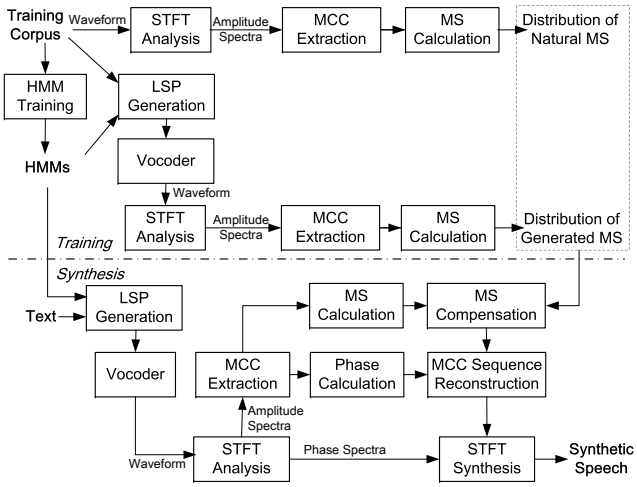

Fig. 3. Flowchart of the proposed MS compensation method using MCCs derived from speech waveforms.

LPS into MCCs. The latter one is achieved by applying DFT to frequency-warped LPS and preserving the first N dimensions of the DFT outputs.

3.3. MCCs derived from speech waveforms

Here, MCCs derived from speech waveforms are used as the spectral features for MS calculation. Its aim is to achieve an MS compensation method which is independent on the type of spectral features and vocoders used in HMM-based speech synthesis. The flowchart of this method is shown in Fig. 3. At training time, each sentence in the training corpus is synthesized using the natural segmental durations, natural F0 features and generated LSPs. STFT analysis is applied to the waveforms of each natural and synthetic utterance. Then, MCCs are extracted from the log amplitude spectrum of each frame by frequency warping and DFT. The first N dimensions of the DFT outputs compose the MCC vector at each frame. These MCCs are used for MS calculation and distribution estimation. At synthesis time, MS compensation is implemented as a post-filter to the synthetic waveforms as shown in Fig. 3. Being consistent with the training process, MCCs are extracted from synthetic waveforms through STFT analysis. The phase spectra of STFT are also preserved. The MS of extracted MCCs are modified according to (4) and then used to reconstruct MCC sequences and amplitude spectra of speech. These modified amplitude spectra are combined with the preserved phase spectra to reconstruct speech waveforms by short-time Fourier synthesis [13].

4. EXPERIMENTS

4.1. Experimental Conditions

Two American English speech databases (male speaker *RMS* and female speaker *SLT* in CMU ARCTIC databases [14]) were used in our experiments. For each database, 1076 utterances were used for model training and the remaining 56 utterances were used for test. The waveforms were in 16 kHz PCM format with 16 bit precision. The acoustic features used for training HMM-based systems were composed of F0, spectral parameters, and their delta and acceleration components. STRAIGHT [15] was adopted as the vocoder to extract acoustic features and to reconstruct speech waveforms.

Table 1. Configuration of systems in our experiments.

System Name	Spectral Features	Post Filtering
MCC_BS	MCC	none
MCC_MS	MCC	conventional MS compensation [8]
LSP_BS	LSP	none
LSP_PF	LSP	formant enhancement [4]
LSP_MS	LSP	conventional MS compensation [8]
LPS_MS	LSP	MS compensation using 513-order LPS derived from LSPs
MCC40_MS	LSP	MS compensation using 40-order MCCs derived from LSPs
MCC513_MS	LSP	MS compensation using 513-order MCCs derived from LSPs
WAV_MS	LSP	MS compensation using 513-order MCCs derived from waveforms

For each speaker, nine systems were constructed and compared in our experiments.[1] These systems adopted either 41-order MCCs or 40-order LSPs plus a gain dimension as spectral features for HMM modeling. Different post-filtering techniques were adopted by these systems. The detailed configurations of these systems are shown in Table 1. For calculating MS from parameter sequences, the FFT length was set to $2M = 4096$ according to the maximum length of utterances in the training corpus. In the *WAV_MS* system, Hamming window was used for STFT analysis. The FFT length of STFT analysis was 1024. The frame length and frame shift was set to 20 ms and 5 ms respectively. After heuristic parameter tuning, the interpolation coefficient α in (4) was set to 0.85 for *LPS_MS*, *MCC40_MS*, and *MCC513_MS* systems, and was set to 1.0 for other systems using MS compensation.

4.2. Subjective Evaluation

The first experiment evaluated the systems using LSPs as spectral features and applying MS compensation. The listening test was conducted by crowdsourcing on Amazon Mechanical Turk (AMT)[2] with anti-cheating considerations [16]. 10 sentences were randomly selected from the test set and were synthesize by *LSP_BS*, *LSP_MS*, *LPS_MS*, *MCC40_MS*, *MCC513_MS*, *WAV_MS* systems of each speaker. A MUSHRA test [17] was conducted to evaluate the naturalness of these systems for each speaker. 20 English-native listeners took part in the tests by rating the utterances synthesized by all systems in parallel using a scale from 0 to 100. Natural recordings were used as reference stimuli. The average naturalness scores of all systems are calculated and shown in Fig. 4. Results of paired *t*-test show that the differences between every pair of systems are significant at 0.05 significance level for both speakers, except the differences among *MCC40_MS*, *MCC513_MS* and *WAV_MS* systems for speaker *SLT*. Comparing *LSP_BS* with *LSP_MS* in Fig. 4, we can see that the naturalness of synthetic speech can be slightly improved if we apply MS compensation to LSPs directly. On the other hand, the *LPS_MS*, *MCC40_MS*, and *MCC513_MS* systems achieved much

[1] Some demos of speech synthesized using these systems can be found at http://home.ustc.edu.cn/~sunxh06/LSP_MS/demo.html.

[2] http://www.mturk.com/

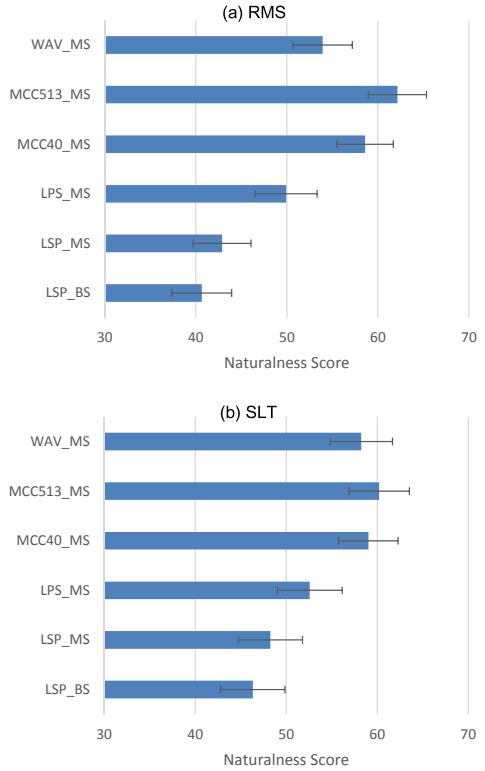

Fig. 4. The average naturalness scores of different systems and their 95% confidence interval for (a) speaker *RMS* and (b) speaker *SLT*.

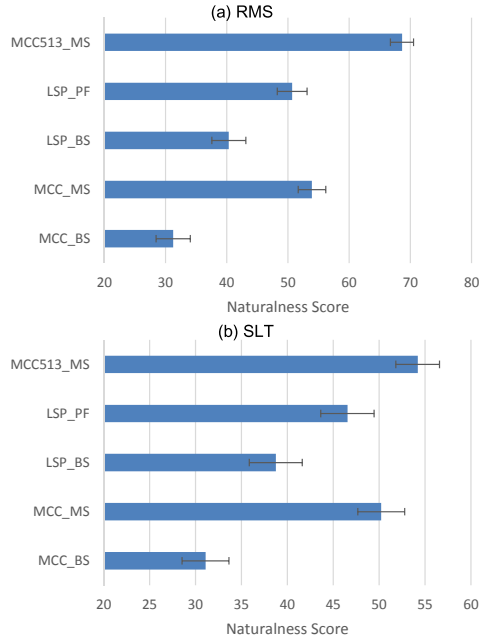

Fig. 5. The average naturalness scores of different systems and their 95% confidence interval for (a) speaker *RMS* and (b) speaker *SLT*.

better performance than *LSP_MS*. This indicates that the effect of MS compensation depends on the type of spectral features used for MS calculation. Using MCCs derived from LSPs achieved the best performance for both speakers. For speaker *RMS*, using higher MCC orders led to better naturalness score. One possible reason is that using 40-order MCCs may lose some details of LPS, especially for the low-frequency band of male speakers. Using MCCs derived from speech waveforms for MS calculation also worked well. For speaker *SLT*, *WAV_MS* is one of the best systems.

Our second experiment compared *MCC_BS*, *MCC_MS*, *LSP_BS*, *LSP_PF*, and *MCC513_MS* systems by a MUSHRA test, where *MCC513_MS* achieved the best performance in previous experiment. The evaluation conditions were the same as previous one and the results are shown in Fig. 5. Results of paired t-test show that the differences between every pairs of systems are significant at 0.05 significance level for both speakers. We can see that using LSPs as spectral features led to better naturalness of synthetic speech than using MCCs in the baseline systems. This is consistent with the results of previous work [18]. Comparing *LSP_PF* with *MCC513_MS*, we can see that applying MS compensation to the MCCs derived from LSPs can achieve better post-filtering performance than the formant enhancement algorithm for LSPs [4]. Furthermore, *MCC513_MS* also had higher naturalness score than *MCC_MS*, which adopted MCCs for both spectral modeling and MCC compensation.

4.3. Discussions

Fig. 4 shows MCCs are more suitable for MS compensation than LSPs and LPS. One possible reason is the similarity between the calculation of MS for MCCs and neural properties of the primary auditory cortex (A1). Previous work on measuring Spectro-Temporal Response Fields (STRF) of mammalian A1 cells [19, 20] shows that an A1 cell is usually selective to a broadband signal whose spectro-temporal envelopes are sinusoidally modulated with particular modulation parameters [21]. This means the STFR of an A1 cell can be considered as a filter specific to a particular range of spectral and temporal modulation frequencies. This is similar to the process of calculating MS of MCCs, which consists of two Fourier transforms to speech spectrogram along frequency and temporal axes respectively. The transformation along frequency axis converts warped LPS into MCCs. The transformation along temporal axis derives MS from MCC sequences.

The *WAV_MS* systems of both speakers achieved satisfactory performance in our experiment. This demonstrates the feasibility of implementing the idea of MS compensation in a spectral-feature-independent and vocoder-independent way. The conventional MS compensation method only works on spectral parameters, which ignores the effect of excitation on the MS of synthetic speech as discussed in [22]. Using MCCs derived from speech waveforms for MS compensation provides an alternative to avoid this issue.

5. CONCLUSIONS

This paper explores the MS compensation method for HMM-based speech synthesis when LSPs are used as spectral features. The effectiveness of compensating the MS of LSPs and other spectral features derived from LSPs are evaluated by subjective evaluation, which shows the superiority of adopting MCCs for MS compensation. An MS compensation method using MCCs derived from speech waveforms has also been proposed to achieve MS compensation in a spectral-feature-independent and vocoder-independent way. To extend current work from MS compensation to joint parameter generation and to develop more perception-related features based on MS will be the tasks of our future work.

6. REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Eurospeech*, 1999, pp. 2347–2350.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *ICASSP*, 2000, vol. 3, pp. 1315–1318.

[3] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.

[4] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006: an improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Eurospeech*, 2001, pp. 2263–2266.

[6] T. Toda and T. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 5, pp. 816–824, 2007.

[7] Z.-H. Ling and L.-R. Dai, "Minimum Kullback-Leibler divergence parameter generation for HMM-based speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1492–1502, 2012.

[8] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 290–294.

[9] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis," in *Proc. ICASSP*. IEEE, 2015, pp. 4210–4214.

[10] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-constrained trajectory training algorithm for HMM-based speech synthesis," in *Proc. Interspeech*, 2015, pp. 1206–1210.

[11] Z.-J. Yan, Y. Qian, and F. K. Soong, "Rich context modeling for high quality HMM-based TTS," in *Interspeech*, 2009, pp. 1755–1758.

[12] Z.-H. Ling, Y. Hu, and L.-R. Dai, "Global variance modeling on the log power spectrum of LSPs for HMM-based speech synthesis," in *Interspeech*, 2010, pp. 825–828.

[13] L. R. Rabiner and R. W. Schafer, *Theory and Applications of DigitalSpeech Processing*, Prentice Hall, 2011.

[14] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[16] Sabine Buchholz and Javier Latorre, "Crowdsourcing preference tests, and how to detect cheating," in *Proc. Interspeech*, 2011, pp. 3053–3056.

[17] BS. 1534-1. Recommendation, ITUR, "Method for the subjective assessment of intermediate sound quality (MUSHRA)," *International Telecommunications Union, Geneva*, 2001.

[18] Y.-J. Wu and R.-H. Wang, "HMM-based trainable speech synthesis for Chinese," *Journal of Chinese Information Processing*, vol. 20, no. 4, pp. 75–81, 2006.

[19] S. A. Shamma, H. Versnel, and N. Kowalski, "Ripple analysis in ferret primary auditory cortex. 1. response characteristics of single units to sinusoidally rippled spectra," Tech. Rep., DTIC Document, 1994.

[20] N. Kowalski, D. A. Depireux, and S. A. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex. I. characteristics of single-unit responses to moving ripple spectra," *Journal of neurophysiology*, vol. 76, no. 5, pp. 3503–3523, 1996.

[21] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[22] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, Z.-H. Ling, and J. Yamagishi, "A deep generative architecture for postfiltering in statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 11, pp. 2003–2014, 2015.