Modified Post-filter to Recover Modulation Spectrum for HMM-based Speech Synthesis

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

***Abstract*—This paper proposes a modified post-filter to recover a Modulation Spectrum (MS) in HMM-based speech synthesis. To alleviate the over-smoothing effect which is one of the major problems in HMM-based speech synthesis, the MS-based post-filter has been proposed. It recovers the utterance-level MS of the generated speech trajectory, and we have reported its benefit to the quality improvement. However, this post-filter is not applicable to various lengths of speech parameter trajectories, such as phrases or segments, which are shorter than an utterance. To address this problem, we propose two modified post-filters, (1) the time-invariant filter with a simplified conversion form and (2) the segment-level post-filter which applicable to a short-term parameter sequence. Furthermore, we also propose (3) the post-filter to recover the phoneme-level MS of HMM-state duration. Experimental results show that the modified post-filters also yield significant quality improvements in synthetic speech as yielded by the conventional post-filter.**

***Index Terms*—HMM-based speech synthesis, modulation spectrum, post-filter, over-smoothing**

I. Introduction

Parametric speech synthesis based on Hidden Markov Models (HMMs) [1] is an effective framework for generating diverse synthetic speech. In HMM-based speech synthesis, speech parameters, i.e., spectral and excitation features and HMM-state duration are simultaneously modeled with context-dependent HMMs in a unified framework. This approach allows us not only to produce smooth speech parameter trajectories under a small footprint [2] but also to apply several techniques for flexibly controlling them [3], [4], [5] to various speech-based systems [6], [7].

One of the critical problems of HMM-based speech synthesis is that the trajectories generated from the trained HMMs are often over-smoothed. This phenomenon causes the degradation of perceptual quality, and synthetic speech sounds muffled [8]. One approach to addressing this problem is to combine a unit selection framework [9], [10], and the other approach is to enhance specific features not well reproduced from the traditional HMMs due to the over-smoothing effect [11], [12]. The latter approach can achieve the production of high-quality speech while preserving its small footprint. As one of the methods based on the latter approach, we have proposed the Modulation Spectrum (MS)-based post-filter [13]. The MS is known as a perceptual cues [14], [15], and the proposed post-filter can improve the quality by recovering the utterance-level MS of the generated speech parameters. However the post-filtering process needs to calculate the MS of the fixed length of speech parameter trajectories, and therefore, it is not applicable to various lengths of speech parameter trajectories, such as phrases or segments. This constraint causes some limitations; e.g., it prevents a recursive speech parameter generation algorithm [16] from being used for low-delay speech waveform generation.

In this paper, we propose two modified post-filters capable of being widely used by relaxing the constraint: (1) the time-invariant filter and (2) the segment-level post-filter. The time-invariant filter makes the filtering process independent of the length of the generated trajectories. The segment-level filter achieves a segment-by-segment filtering process to recover the MS of a shorter length of speech parameter trajectories compared to the conventional utterance-level filter. Furthermore, to further improve naturalness of synthetic speech, (3) we propose the post-filter for HMM-state duration to recover the MS of a phone-level duration sequence in a similar manner to in the conventional post-filter. We evaluate performance of the individual proposed methods separately to investigate the effect of them on naturalness of synthetic speech.

II. Parameter Generation

In synthesis, HMMs corresponding to input text are constructed from context-dependent HMMs build using natural speech parameters in training. . After determining the HMM-state sequence $\boldsymbol{q} = [q_1, \cdots, q_T]$ to maximize the duration likelihood, the parameter trajectory is generated to maximize HMM likelihood under a constraint on the relationship between static and dynamic features as follows:

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

where $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$ is a speech parameter vector sequence of T frames, $\boldsymbol{c}_t = \left[c_t\left(1\right), \cdots, c_t\left(d\right), \cdots, c_t\left(D\right)\right]^\top$ is a D-dimensional parameter vector at frame t, d is a dimensional index, $\boldsymbol{W}$ is the weighting matrix for calculating the dynamic features [17], q_t is a HMM-state index at frame t, and $\boldsymbol{\lambda}$ is a HMM parameter set. To alleviate the over-smoothness of the generated parameters, Global Variance (GV) [11] can be also considered in parameter generation.

III. Conventional MS-based Post-Filter [13]

A. MS-based Post-Filtering Process

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

978-1-4799-7088-9/14/$31.00 ©2014 IEEE

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

$$\boldsymbol{s}(d) = \left[s_d(0), \cdots, s_d(f), \cdots, s_d(F_s)\right]^\top, \quad (3)$$

where $s_d(f)$ is the f-th MS of the d-th dimension of the parameter sequence $[c_1(d), \cdots, c_T(d)]^\top$, f is a modulation frequency index, F_s is a half number of the DFT length.

In synthesis, the speech parameter sequence generated from the HMM is transformed to the modulation frequency domain. Then, its MS is converted as follows:

$$s'_d(f) = (1-k)s_d(f) + k\left[\frac{\sigma^{(\mathrm{N})}_{d,f}}{\sigma^{(\mathrm{G})}_{d,f}}\left(s_d(f) - \mu^{(\mathrm{G})}_{d,f}\right) + \mu^{(\mathrm{N})}_{d,f}\right], \quad (4)$$

where $\mu^{(\cdot)}_{d,f}$ and $\sigma^{(\cdot)}_{d,f}$ are mean and standard deviation of $s_d(f)$, N and G indicate of MS of the natural parameter and the generated speech parameter sequence, respectively. The MS statistics are estimated in advance from natural and generated speech parameter sequences for training data. The coefficient k is a parameter to control the degree of emphasis, which is determined manually. Finally, the filtered speech parameter sequence is generated from the converted MS and its original phase.

B. Problems

In [13], the MS is calculated utterance by utterance, The DFT length for the MS calculation 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; the utterance-level filtering process is hard to be applied to a low-latency speech synthesis frame work [18] where a frame-level or segment-level processing based on the recursive parameter generation [16] is essential.

Moreover, it has been reported that post-processing to enhance speech parameters, such as the GV-based parameter generation, is also effective for not only spectral and F_0 parameters but also HMM-state duration [19]. Although we have applied the MS-based post-filter to only spectrum and F_0, it is worthwhile to also apply the MS-based post-filter to the HMM-state duration and investigate its effectiveness.

IV. Proposed Modification Methods for MS-Based Post-Filter

To address the problems in the conventional MS-based post-filter, we propose two modification methods for the MS-based post-filter. Moreover, we also propose the MS-based post-filter for the HMM-state duration.

A. Method 1: Time-Invariant Post-Filter

A time-invariant post-filter is derived by assuming that $\sigma^{(\mathrm{N})}_{d,f}$ is equal to $\sigma^{(\mathrm{G})}_{d,f}$ in Eq. (4) as follows:

$$\begin{aligned} s'_d(f) &= (1-k)s_d(f) + k\left[s_d(f) - \mu^{(\mathrm{G})}_{d,f} + \mu^{(\mathrm{N})}_{d,f}\right] \\ &= s_d(f) + k\left[\mu^{(\mathrm{N})}_{d,f} - \mu^{(\mathrm{G})}_{d,f}\right]. \end{aligned} \quad (5)$$

Because the second term in R.H.S. is independent of $s_d(f)$, this conversion process can be represented as a filtering process for the generated speech parameter sequence with a time-invariant FIR filter.

B. Method 2: Segment-Level Post-Filter

A segment-level post-filter is derived by localizing the post-filtering process as illustrated in the left-hand side of Figure 1. 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 conventional filtering. 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 parameter, silence frames are removed in calculating the MS statistics to alleviate the over-fitting problem [20]. For F_0, continuous F_0 pattern [21] is used [13]. The segment-level post-filtering can be applicable to the low-delay speech waveform generation. Moreover, it is possible to further implement context-dependent post-filtering.

C. Method 3: MS-Based Post-Filter for Duration

Although the state duration is not an actual parameter trajectory, it is affected by the over-smoothing effect due to a statistical averaging process as in spectrum and F_0 parameters [22]. As illustrated in Figure 2, we can interestingly find the MS degradation of the modulation frequency[1] of phoneme-level duration sequences. Therefore, it is expected that quality improvements in synthetic speech are yielded by recovering their MS.

The overview of the proposed method is illustrated in the right side of Figure 1. First, phoneme-level duration is calculated from the determined state-level duration. Then, a phoneme-level duration sequence over an utterance is constructed by excluding the silence parts and its mean value is normalized as in F_0 parameters [13]. The resulting sequence is used to calculate the MS and is also filtered in the same manner as the conventional post-filtering. 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.

V. Experimental Evaluations

A. Experimental Conditions

We trained a context-dependent phoneme Hidden Semi-Markov Model (HSMM) [23] for a Japanese female speaker.

[1]Nyquist frequency is set to 1.0.

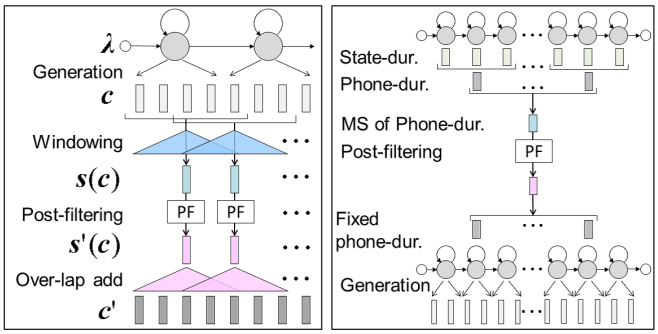

Fig. 1. An overview of the proposed methods (left: the segment-level post-filter, right: the post-filter for duration).

We used 450 sentences for training and 53 sentences for evaluation from phonetically balanced 503 sentences included in the ATR Japanese speech database [24]. 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 a spectral parameter and log-scaled F_0 and 5 band-aperiodicity [25], [26] 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.

B. Evaluation 1: Time-Invariant Post-Filter

To confirm the effect by the time-invariant filter, we conducted the subjective evaluation to compare the following speech samples:

HMM: original parameter by Eq. (1)
HMM+MS (ti): parameters filtered by the time-invariant filter
HMM+MS: parameters filtered by the conventional filter

From [13], the emphasis coefficient and DFT length were set to 0.85 and 4096, respectively. We applied the MS-based post-filter to both spectrum and F_0. We conducted a preference test (AB test) on speech quality. Every pair of three types of synthetic speech was presented to listeners in random order. 6 listeners were asked which sample sounds better in terms of speech quality.

The preference result is shown in Figure 3. We can see that a significant quality improvement is yielded by applying the time-invariant post-filter to the generated speech parameters. Although the improved quality is not comparable to that yielded by the conventional post-filter, the time-invariant post-filter is applicable to various lengths of speech parameter sequences.

C. Evaluation 2: Segment-Level Post-Filter

The window length and window shift length were set to 125 ms (25 samples) [28] and 60 ms (12 samples), respectively. 64-taps FFT was used. We compared the following speech samples:

HMM: original parameter generated by Eq. (1)

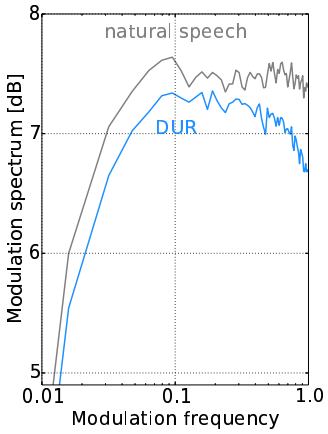

Fig. 2. Averaged MSs of phone-level duration sequences ("DUR": generated duration).

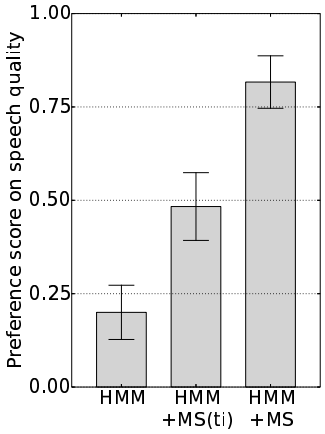

Fig. 3. Preference scores with 95% confidence intervals (the time-invariant post-filter).

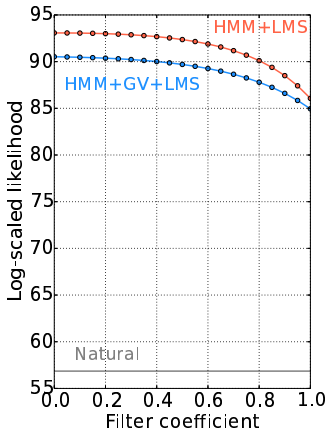

Fig. 4. HMM likelihoods for the filtered spectrum.

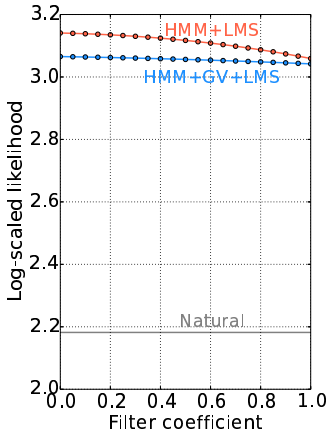

Fig. 5. HMM likelihoods for the filtered F_0.

HMM+LMS: "HMM" parameters filtered by the segment-level filter
HMM+GV: parameters generated by Eq. (1) with the GV
HMM+GV+LMS: "HMM+GV" parameters filtered by the segment-level filter

1) Tuning Emphasis Coefficient: We calculated the HMM likelihood, GV likelihood, and MS likelihood for the filtered both spectral parameters and F_0 contours while varying the emphasis coefficient from 0 to 1. For comparison, the likelihood for natural speech parameters was calculated, which is labeled as **"Natural."**

The results are shown in Figs. 4 to 9. Their tendencies are similar to those of the conventional post-filter as reported in [13]. The degradation of HMM likelihoods by the post-filtering process, but they are sill greater than that of natural parameters. Almost likelihoods tend to increase as the filter coefficient is close to 1. we observed the 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.

2) Subjective Assessment on Speech Quality: AB test using the above 4 methods on speech quality by 7 listeners was

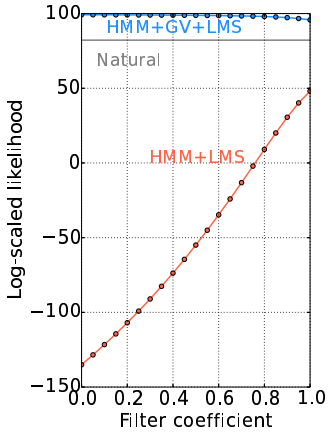

Fig. 6. GV likelihoods for the filtered spectrum.

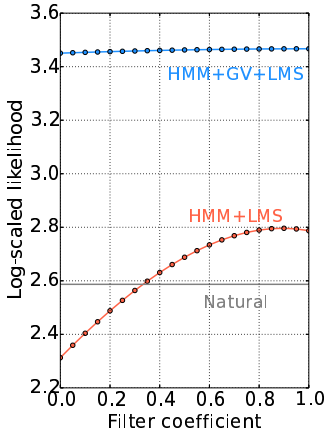

Fig. 7. GV likelihoods for the filtered F_0.

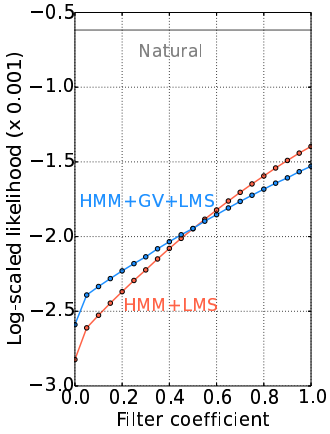

Fig. 8. MS likelihoods for the filtered spectrum.

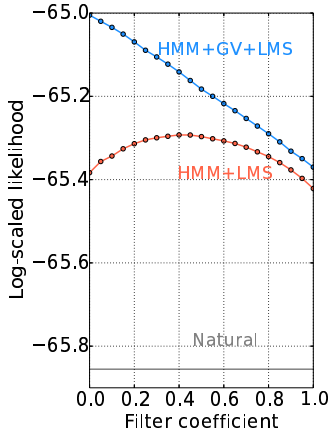

Fig. 9. MS likelihoods for the filtered F_0.

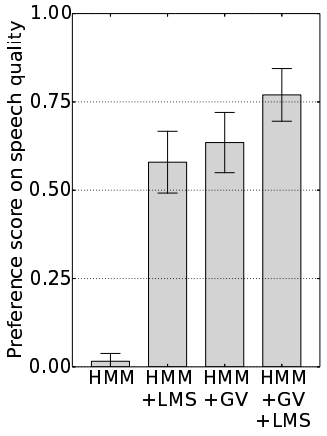

Fig. 10. Preference scores with 95% confidence intervals (local MS-based post-filter).

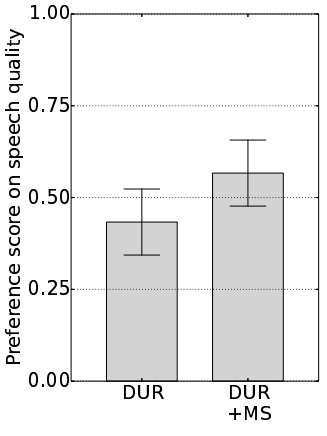

Fig. 11. Preference scores with 95% confidence intervals (post-filter for duration) .

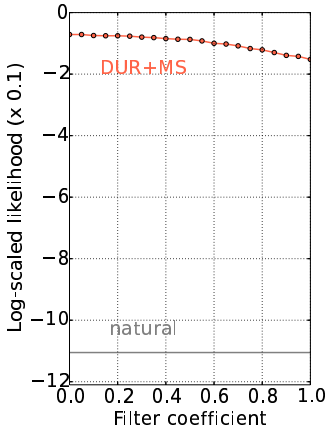

Fig. 12. Duration likelihoods for the filtered duration.

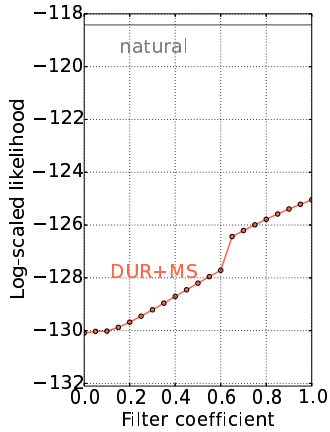

Fig. 13. MS likelihoods for the filtered duration.

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 Figure 10. It is observed that the significant quality gain is yielded by "HMM+LMS" compared to "HMM," and its comparable to that yielded by "HMM+GV." Furthermore, we can see that the additional gain is yielded by "HMM+GV+LMS" compared to "HMM+GV." This tendency is similar to that observed in the conventional post-filter as reported in [13]. Please note that the segment-level post-filter is applicable to various lengths of a speech parameter sequence but the conventional one cannot.

D. Evaluation 3: MS-Based Post-Filtering for Duration

We evaluated the effectiveness of the post-filter for duration. 64-taps FFT was used. The spectrum and F_0 is not filtered. Compared speech samples are below:

DUR: original duration

DUR+MS: duration filtered by the proposed the post-filter

The duration likelihood and MS likelihood are shown in Figure 12 and Figure 13, respectively. We can see that the MS likelihood increases as the filter coefficient is close to 1 while preserving the duration likelihood high enough. Therefore, the emphasis coefficient was set to 1.0 in the subjective evaluation. We can also see discontinuous transition of the MS likelihood. We expect that this was caused by the effect of rounding the filtered duration values into integer values after filtering.

The result of AB test by 6 listeners is shown in Figure 11. We can see that te MS-based post-filter for duration tends to slightly improve speech quality.

VI. Summary

This paper have proposed the modified Modulation Spectrum (MS)-based post-filters in HMM-based speech synthesis. We have reported that the post-filters can avoid the conventional limitation while preserving the quality gain. Furthermore, we have applied the MS-based post-filter to phone-level duration, and have yielded the effectiveness on speech quality. We will investigate the benefits of the post-filter and MS itself on various situation.

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."

REFERENCES

[1] 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.

[2] K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda. Tying covariance matrices to reduce the footprint of HMM-based speech synthesis systems. In *Proc. INTERSPEECH*, pp. 1759–1762, Brighton, U. K., 2009.

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

[4] 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.

[5] 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.

[6] K. Shirota, K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda. Integration of speaker and pitch adaptive training for HMM-based singing voice synthesis. In *Proc. ICASSP*, pp. 2578–2582, Florence, Italy, May 2014.

[7] 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.

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

[9] Z. Ling, L. Qin, H. Lu, Y. Gao, L. Dai, R. Wang, Y. Jiang, Z. Zhao, J. Yang, J. Chen, and G. Hu. The USTC and iFlytek speech synthesis systems for blizzard challenge 2007. In *Proc. Blizzard Challenge workshop*, Bonn, Germany, Aug. 2007.

[10] 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, May 2014.

[11] 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.

[12] 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.

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

[14] 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.

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

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

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

[18] T. Baumann and D. Schlangen. INPRO_iSS A component for just-in-time incremental speech synthesis. *Proc. ACL*, pp. 103–108, Jul. 2012.

[19] S. Pan, Y. Nankaku, K. Tokuda, and J.Tao. Global variance modelinf on the log power spectrum of lsps for HMM-based speech synthesis. In *Proc. ICASSP*, pp. 4716–4719, Prague, Czech Republic, 2011.

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

[21] 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.

[22] 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*, pp. 2347–2350, Budapest, Hungary, Apr. 1999.

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

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

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

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

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