PARAMETER GENERATION ALGORITHM CONSIDERING 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
‡ Language Technologies Institute, Carnegie Mellon University (CMU), U.S.A
Email: shinnosuke-t@is.naist.jp

ABSTRACT

This paper proposes a novel parameter generation algorithm for high-quality speech generation in Hidden Markov Model (HMM)-based speech synthesis. One of the biggest issues causing significant quality degradation is the over-smoothing effect often observed in generated parameter trajectories. Global Variance (GV) is known as a feature well correlated with the over-smoothing effect and a metric on the GV of the generated parameters is effectively used as a penalty term in the conventional parameter generation. However, the quality of the synthetic speech is far from that of the natural speech. Recently, we have found that a Modulation Spectrum (MS) of the generated parameters, which is also regarded as an extension of the GV, is more sensitively correlated with the over-smoothing effect than the GV. This paper incorporates a metric on the MS as a new penalty term in the proposed parameter generation algorithm. The experimental results demonstrate that the proposed parameter generation algorithm considering the MS yields significant improvements in synthetic speech quality compared to the conventional parameter generation algorithm considering the GV.

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

1. INTRODUCTION

Statistical parametric speech synthesis based on Hidden Markov Models (HMMs) [1] is an effective framework for generating diverse types of synthetic speech. Speech parameters, i.e., spectral and excitation features and HMM-state duration are simultaneously modeled with context-dependent HMMs in a unified framework [2]. In synthesis, the speech parameter trajectories are generated by maximizing the likelihood of the HMMs [3]. This approach allows us not only to produce smoothly varying speech parameter trajectories under a small footprint [4] but also to apply several techniques for flexibly controlling synthetic speech [5, 6, 7] to various speech-based systems [8, 9].

One of the critical problems in HMM-based speech synthesis is that the parameter trajectories generated from the HMMs are often over-smoothed. This phenomenon causes significant degradation of the perceptual quality and makes synthetic speech sound muffled [10]. To address this over-smoothing problem, Toda and Tokuda [11] proposed a parameter generation algorithm considering Global Variance (GV). The GV is defined as the 2nd order moment of the parameter trajectory. The GV of the generated trajectories is often smaller than that of natural speech parameter trajectories. They integrated a metric on the GV into the parameter generation algorithm to keep the GV close to natural one, and they reported that the speech quality was significantly improved. This conventional method has been widely studied [12, 13, 14] because of not only its practical advantage for improving synthetic speech quality but also its mathematically tractable formulation. However, the synthetic speech quality still sounds unnatural compared to the natural one.

Recently, Takamichi *et al.,* [15] have found that a Modulation Spectrum (MS) of the generated parameter trajectory is more sensitively correlated with the over-smoothing effect than the GV. Because the MS is regarded as an extension of the GV, it enables to quantify more characteristics of the parameter trajectory than the GV. [15] has reported compensating the MS of the generated trajectories using a simple post-filter process. However, this framework based on the post-filtering possibly causes adverse effects due to completely ignoring the HMM likelihood. Moreover, it is not straightforward to further apply various extensions, such as model adaptation [6] and joint optimization in training [16], to this framework as it is not well formulated mathematically.

This paper proposes a speech parameter generation algorithm considering the MS. The proposed algorithm generates the parameter trajectories by maximizing a novel objective function consisting of the HMM and MS likelihoods. The MS likelihood works as a penalty term to make the MS of the generated parameters close to that of natural ones. The proposed algorithm is applied to both spectral and F_0 components. The experimental results demonstrate that the proposed algorithm significantly outperforms the conventional algorithm considering the GV.

2. CONVENTIONAL PARAMETER GENERATION ALGORITHM CONSIDERING GLOBAL VARIANCE

Context-dependent HMMs are trained using natural speech parameters. In synthesis, after constructing a sentence HMMs corresponding to input text to be synthesized, the HMM state sequence $\boldsymbol{q} = [q_1, \cdots, q_t, \cdots, q_T]^\top$ is determined , where q_t is state index at frame t. Then, the parameter trajectory is generated by maximizing the objective function consisting of both HMM and GV likelihood as follows:

$$\hat{\boldsymbol{c}} = \operatorname*{argmax}_{\boldsymbol{c}} \log \mathcal{N}\left(\boldsymbol{W}\boldsymbol{c}; \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}\right)^{\omega_{\mathrm{v}}} \mathcal{N}\left(\boldsymbol{v}\left(\boldsymbol{c}\right); \boldsymbol{\mu}_{\mathrm{v}}, \boldsymbol{\Sigma}_{\mathrm{v}}\right), \tag{1}$$

where $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_t^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$ is a speech parameter vector sequence of T frames, $\boldsymbol{c}_t = \left[c_t\left(1\right), \cdots, c_t\left(d\right), \cdots, c_t\left(D\right)\right]^\top$ is a D-dimensional parameter vector at frame t, d is a dimensional index, $\boldsymbol{W}$ is the weighting matrix for calculating the dynamic features [3], and $\mathcal{N}\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denotes Gaussian distribution of a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{\mu}_{\boldsymbol{q}} = \left[\boldsymbol{\mu}_{q_1}^\top, \cdots, \boldsymbol{\mu}_{q_T}^\top\right]^\top$ and $\boldsymbol{\Sigma}_{\boldsymbol{q}} = \mathrm{diag}\left[\boldsymbol{\Sigma}_{q_1}, \cdots, \boldsymbol{\Sigma}_{q_T}\right]$ are the mean vector and the covariance matrix, where $\boldsymbol{\mu}_{q_t}$ and $\boldsymbol{\Sigma}_{q_t}$ are a mean vector and a covariance matrix of the state q_t, respectively. $\boldsymbol{v}\left(\boldsymbol{c}\right) = \left[v\left(1\right), \cdots, v\left(d\right), \cdots, v\left(D\right)\right]^\top$ is

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

the GV vector [11], where the d-th component $v(d)$ is given by:

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(c_t(d) - \bar{c}(d)\right)^2, \tag{2}$$

$$\bar{c}(d) = \frac{1}{T}\sum_{\tau=1}^{T} c_\tau(d), \tag{3}$$

where $\boldsymbol{\mu}_{\mathrm{v}}$ and $\boldsymbol{\Sigma}_{\mathrm{v}}$ are the mean vector and the covariance matrix of the GV, respectively. ω_{v} is a weight of the GV likelihood. The maximization of the objective function is performed using the gradient method. To determine the initial parameter trajectories, the parameter trajectories analytically determined by maximizing only the HMM likelihood are linearly transformed to maximize the GV likelihood [11]. Then, they are iteratively updated by maximizing the objective function.

3. PROPOSED PARAMETER GENERATION ALGORITHM CONSIDERING MODULATION SPECTRUM

3.1. Modulation Spectrum Definition

Whereas the GV represents the temporal scaling of the parameter trajectory as a scalar value in each feature dimension, the MS explicitly represents the temporal fluctuation as a vector, and its sum over all modulation frequency except bias is equal to the GV [15]. In this paper, we define the MS $\boldsymbol{s}(\boldsymbol{c})$ of the parameter trajectory $\boldsymbol{c}$ as follows:

$$\boldsymbol{s}(\boldsymbol{c}) = \left[\boldsymbol{s}(1)^\top, \cdots, \boldsymbol{s}(d)^\top, \cdots, \boldsymbol{s}(D)^\top\right]^\top, \tag{4}$$

$$\boldsymbol{s}(d) = \left[s_d(0), \cdots, s_d(m), \cdots, s_d(M-1)\right]^\top, \tag{5}$$

$$s_d(m) = R_{d,m}^2 + I_{d,m}^2, \tag{6}$$

$$= \left(\sum_{t=0}^{T} c_t(d)\cos kt\right)^2 + \left(\sum_{t=0}^{T} c_t(d)\sin kt\right)^2, \tag{7}$$

where $2M$ is a length of Discrete Fourier Transform (DFT). $k = -\pi m/M$ is a modulation frequency. The MS likelihood is defined as $\mathcal{N}(\boldsymbol{s}(\boldsymbol{c}); \boldsymbol{\mu}_{\mathrm{s}}, \boldsymbol{\Sigma}_{\mathrm{s}})$ where $\boldsymbol{\mu}_{\mathrm{s}}$ and $\boldsymbol{\Sigma}_{\mathrm{s}}$ are a DM-by-1 mean vector and a DM-by-DM covariance matrix, respectively. $\boldsymbol{\Sigma}_{\mathrm{s}}^{-1}$ is represented as $\left[\boldsymbol{p}_{\mathrm{s}}^{(1)}, \cdots, \boldsymbol{p}_{\mathrm{s}}^{(d)}, \cdots, \boldsymbol{p}_{\mathrm{s}}^{(D)}\right]$ where $\boldsymbol{p}_{\mathrm{s}}^{(d)}$ is DM-by-M matrix of which columns correspond to $\boldsymbol{s}(d)$. In this paper, the MS is calculated utterance by utterance and its mean vector and covariance matrix are calculated from the whole utterances of the training data.

3.2. Generation Algorithm

We define an objective function L_{s} maximized in the proposed parameter generation algorithm as follows:

$$\hat{\boldsymbol{c}} = \operatorname*{argmax}_{\boldsymbol{c}} L_{\mathrm{s}}, \tag{8}$$

$$L_{\mathrm{s}} = \log \mathcal{N}\left(\boldsymbol{W}\boldsymbol{c}; \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}\right)^{\omega_{\mathrm{s}}} \mathcal{N}\left(\boldsymbol{s}(\boldsymbol{c}); \boldsymbol{\mu}_{\mathrm{s}}, \boldsymbol{\Sigma}_{\mathrm{s}}\right), \tag{9}$$

$$= \omega_{\mathrm{s}}\left(-\frac{1}{2}\boldsymbol{c}^\top\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}\boldsymbol{c} + \boldsymbol{c}^\top\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}}\right)$$

$$-\frac{1}{2}\boldsymbol{s}(\boldsymbol{c})^\top\boldsymbol{\Sigma}_{\mathrm{s}}^{-1}\boldsymbol{s}(\boldsymbol{c}) + \boldsymbol{s}(\boldsymbol{c})^\top\boldsymbol{\Sigma}_{\mathrm{s}}^{-1}\boldsymbol{\mu}_{\mathrm{s}} + \bar{K}, \tag{10}$$

where ω_{s} denotes the MS weight for controlling balance between the HMM and MS likelihoods, which is set to the ratio between the number of dimensions, i.e., $\omega_{\mathrm{s}} = M/(3T)$ in this paper. $\bar{K}$ is a constant value to $\boldsymbol{c}$. Because the objective function is given by a quartic form like that in the conventional algorithm considering the GV, it is hard to analytically solve its maximization problem. We employ the steepest descent algorithm to iteratively update the generated parameter trajectory as follows:

$$\hat{\boldsymbol{c}}^{(i+1)} = \hat{\boldsymbol{c}}^{(i)} + \alpha \left.\frac{\partial L_{\mathrm{s}}}{\partial \boldsymbol{c}}\right|_{\boldsymbol{c}=\hat{\boldsymbol{c}}^{(i)}}, \tag{11}$$

where α is the step size parameter, and i is an iteration index. The first derivative of L_{s} is calculated as

$$\frac{\partial L_{\mathrm{s}}}{\partial \boldsymbol{c}} = \omega_{\mathrm{s}}\left(-\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}\boldsymbol{c} + \boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}}\right),$$

$$+\left[{\boldsymbol{s}'_1}^\top, \cdots, {\boldsymbol{s}'_t}^\top, \cdots, {\boldsymbol{s}'_T}^\top\right]^\top \tag{12}$$

$$\boldsymbol{s}'_t = \left[s_t(1), \cdots, s_t(d), \cdots, s_t(D)\right]^\top, \tag{13}$$

$$s_t(d) = \left(\boldsymbol{s}(\boldsymbol{c}) - \boldsymbol{\mu}_{\mathrm{s}}\right)^\top \boldsymbol{p}_{\mathrm{s}}^{(d)} \boldsymbol{f}_t(d), \tag{14}$$

$$\boldsymbol{f}_t(d) = \left[f_{t,d}(0), \cdots, f_{t,d}(m), \cdots, f_{t,d}(M-1)\right]^\top, \tag{15}$$

$$f_{t,d}(m) = -2\left(R_{d,m}\cos kt + I_{d,m}\sin kt\right). \tag{16}$$

3.3. Initialization

For initialization, we can basically use the same idea as used in the conventional algorithm, i.e., first generating the parameter trajectory by maximizing only the HMM likelihood and then it is further transformed to maximize the other likelihood. To transform the parameter trajectory so that the MS likelihood is increased, we use the MS-based post-filter [15], which is given by

$$s_d(m)' = \frac{\sigma_{d,m}}{\sigma'_{d,m}}\left(s_d(m) - \mu'_{d,m}\right) + \mu_{d,m}, \tag{17}$$

where $\mu_{d,m}$ and $\sigma_{d,m}$ are a mean and a standard deviation of $s_d(m)$, respectively. $\mu'_{d,m}$ and $\sigma'_{d,m}$ are those of MS of generated trajectories. We in advance estimate $\mu'_{d,m}$ and $\sigma'_{d,m}$ using the generated trajectories included in the training data. Finally, the initial parameter trajectory $\boldsymbol{c}^{(0)}$ is determined using the filtered MS and the original phase components of the parameter trajectory before the filtering.

3.4. Application to F0 Contour Generation

The proposed parameter generation is also applied to the F_0 components modeled with Multi-Space probability Distribution (MSD)-HMM [17]. In the F_0 parameter generation, after unvoiced/voiced determination, F_0 values at only voiced frames are generated from the corresponding probability density functions while setting the precision matrices (inverse matrix of the covariance matrix) at the unvoiced/voiced boundaries to zero matrices to allow discontinuous transitions. Therefore, the MS is calculated from the concatenated voiced frames in this paper. Moreover, we reform $c_t(d)$ of Eq. (7) as $c_t(d) - \bar{c}(d)$ as pointed out in [15]. In this implementation, the MS is directly affected by the discontinuous transitions at the unvoiced/voiced boundaries. This causes some adverse effects in the MS-based post-filtering process. To avoid this, we adopt the initialization method used in the conventional GV-based algorithm rather than the MS-based post-filtering in this paper.

3.5. Discussion

Though we can also integrate the GV term into the proposed objective function, i.e., a product of the HMM, GV, and MS likelihoods, the proposed objective function effectively recovers the GV likelihood as well without it because the MS involves the GV as we described. Figures 1 and 2 illustrate an example of the GV and the MS of the generated parameter trajectories. “HMM,” “GV,” and “MS” indicate the results of the generated parameter trajectories using the traditional algorithm without the GV [3], the conventional algorithm

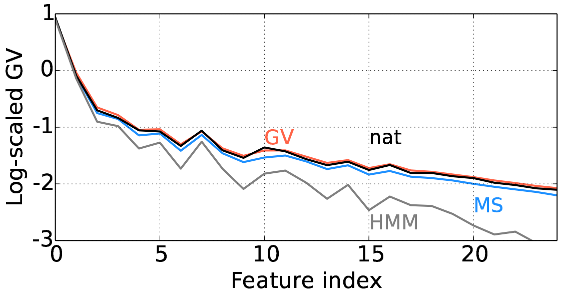

Fig. 1. Examples of GV of generated mel-ceptral coefficients.

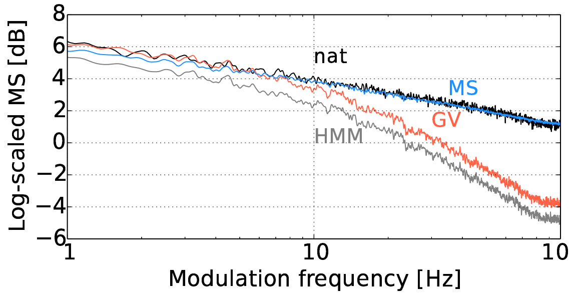

Fig. 2. Examples of MS of generated the 9-th mel-ceptral coefficient.

with the GV [11], and the proposed algorithm with the MS, respectively. "nat" indicates those of natural speech parameter trajectories. We can find that the proposed generation algorithm well recovers not only the MS but also the GV. On the other hand, "GV" cannot recover the MS appropriately. It makes the MS slightly larger but the resulting MS is still very different from the natural one. This is because the GV can model only an average value of the MS components over modulation frequencies.

The footprint of the synthesis system using the proposed algorithm is slightly larger compared to that using the conventional algorithm with the GV because the MS is MD-dimensional vector whereas the GV is D-dimensional vector. We may reduce the footprint by considering only low modulation frequency components which has more dominant effect on speech perception [18].

Finally, the MS-based post-filter [15] tends to generate the over-transformed trajectory and synthesize over-emphasized speech because it completely ignores the criterion like the HMM likelihood. On the other hand, the proposed algorithm can effectively generate the naturally-fluctuated parameter trajectory by jointly maximizing both the HMM and MS likelihoods. In addition, as the proposed algorithm is well formulated mathematically, it is straightforward to apply various techniques, such as context dependent modeling, adaptation, and joint optimization in training.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Condition

We trained a context-dependent phoneme Hidden Semi-Markov Model (HSMM) [19] for an English male speaker "RMS" and an English female speaker "SLT" from the CMU ARCTIC database [20]. We used 593 sentences from subset A for training and 100 sentences from subset B for evaluation. 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 [21, 22] were extracted as excitation parameters. The STRAIGHT analysis-synthesis system [23] was employed for parameter extraction and waveform generation. The feature vector consisted of spectral and excitation parameters and their delta and delta-delta features. 5-state left-to-right HSMMs were used. The DFT length to calculate the MS was set to 8192. Diagonal covariance matrices were used in the HSMM, the GV probability density function, and the MS probability density function.

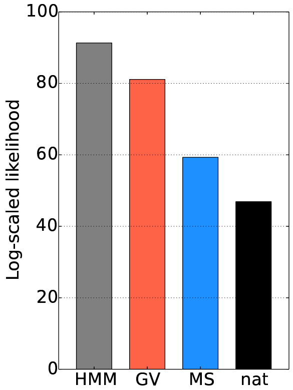

Fig. 3. HMM likelihood for the generated spectral parameters.

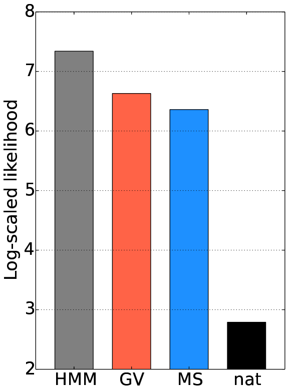

Fig. 4. HMM likelihood for the generated F_0 contours.

We compared the following parameter trajectories, **HMM**: parameters generated by the traditional generation [3], **GV**: parameters generated by the conventional algorithm considering the GV [11], **MS**: parameters generated by the proposed algorithm considering the MS, and **nat**: natural speech parameters. To confirm the effect of the generation algorithms, we did not compare them to parameters filtered by the MS-based post-filter [15]. In order to avoid slightly artificial sounds caused by enhancing the high modulation frequency components of "MS" speech, we applied 50 Hz-cutoff low pass filter to the generated parameter trajectories used in the subjective evaluation[1]. We first conducted an objective evaluation by the likelihoods used in the conventional and proposed generation algorithms. Then, we conducted a subjective evaluation on speech quality. The "GV" was used in the aperiodic component to generate synthetic speech.

4.2. Objective Evaluation Using Likelihood

The generation algorithms were evaluated by the HMM, GV, and MS likelihoods for the generated trajectories. Additionally, we estimated the log-MS $\log s_d(m)$ probability density function and also calculated its likelihood to deeply discuss the results. Figures 3–10 show 4 types of likelihoods for spectral and F_0 components, respectively. Note that these results were averaged between two speakers.

HMM and GV likelihoods: We can find in Fig. 3 and Fig. 4 that the HMM likelihoods for both spectral and F_0 components in the proposed algorithm ("MS") are lower than the traditional algorithm ("HMM") and the conventional algorithm ("GV"), but they are still reasonably higher than those of natural speech parameter trajectories ("nat"). For the GV likelihoods shown in Fig. 5 and Fig. 6, "MS" can effectively recover the GV likelihood as in the

[1] As a result of our preliminary subjective test, we have found that there was no significant quality difference between analysis-synthesized speech samples with/without the MS components over 50 Hz.

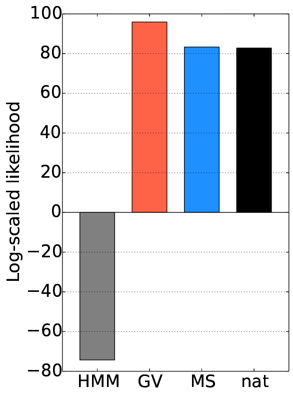

Fig. 5. GV likelihood for the generated spectral parameters.

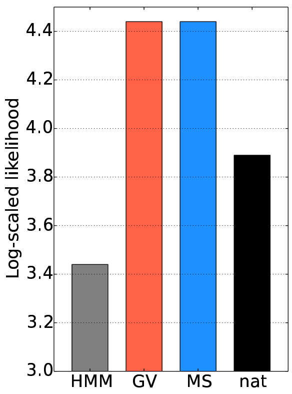

Fig. 6. GV likelihood for the generated F_0 contours.

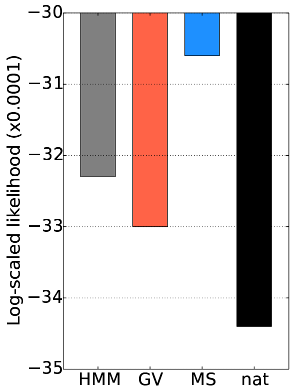

Fig. 7. MS likelihood for the generated spectral parameters.

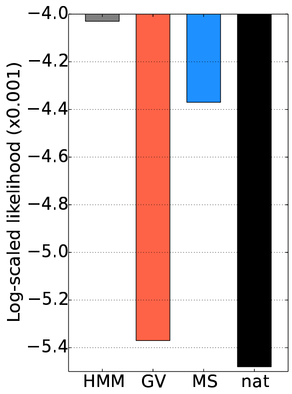

Fig. 8. MS likelihood for the generated F_0 contours.

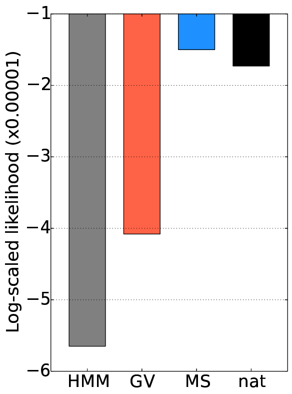

Fig. 9. Log-MS likelihood for the generated spectral parameters.

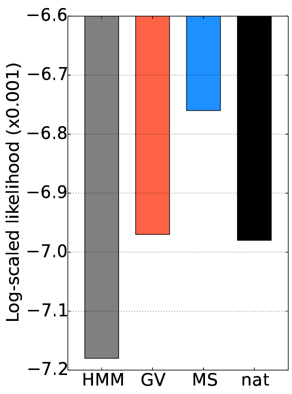

Fig. 10. Log-MS likelihood for the generated F_0 contours.

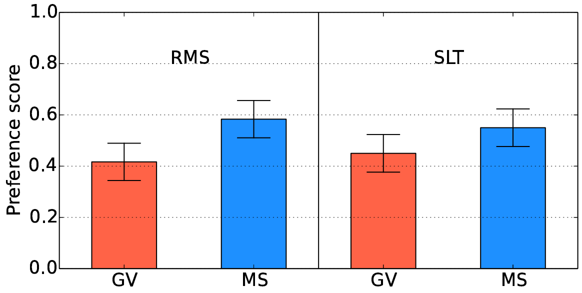

Fig. 11. Preference scores on speech quality with 95% confidence interval.

"GV." These results demonstrate that the proposed algorithm well preserves the conventional criteria.

MS and log-MS likelihoods: In the spectral component, it is observed in Fig. 7 that the MS likelihood of "MS" is larger than that of "HMM" and "GV." In the F_0 component shown in Fig. 8, the MS likelihood of "MS" is larger than that of the "GV." From the comparisons between "MS" and "GV," which use the same initial parameter trajectories but different objective functions, it is shown that the proposed objective function is effective for recovering the MS likelihood. However, the MS likelihood of "HMM" is higher than that of "MS." Moreover, in both spectral and F_0 components, the MS likelihoods of "HMM" are higher than those of "nat." These results are hard to interpret. To analyze them, the Fig. 9 and Fig. 10 illustrate the log-MS likelihoods. We can see more reasonable results in these figures. These results imply that a probability density of the MS is well modeled by the Gaussian distribution on the logarithm domain. Nevertheless, we confirmed there was no perceptual quality difference between the MS modeling and the log-MS modeling in the proposed parameter generation algorithm.

4.3. Subjective Evaluation

We conducted a preference test (AB test) on speech quality by 8 listeners. The pair of synthetic speech of "GV" and "MS" was presented to listeners in a random order. Listeners were asked which sample sounded better in terms of speech quality.

The result of the preference test for two speakers is illustrated in Fig. 11. We can see that the score of "MS" increases over the "GV." This result shows that the proposed algorithm can generate better quality of synthetic speech than the conventional algorithm considering the GV.

5. CONCLUSION

In this paper, we have integrated a Modulation Spectrum (MS) into the parameter generation algorithm in HMM-based speech synthesis. The parameter trajectories are generated by jointly maximizing the HMM and MS likelihoods. The experimental results have demonstrated that the proposed algorithm can recover the MS while preserving the conventional HMM and Global Variance (GV) likelihoods. Moreover, the result of perceptual evaluation on speech quality has shown that the proposed algorithm yields significantly better quality than the conventional algorithm considering the GV. As future work, we will conduct more investigations, such as a comparison to the MS-based post-filter [15] and application to continuous F_0 modeling [24]. Moreover, we will incorporate the MS into the training phase like the GV-constrained trajectory training [16].

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.”

6. 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[13] 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, Apr. 2014.

[14] 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, Apr. 2014.

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

[16] T. Toda and S. Young. Trajectory training considering global variance for HMM-based speech synthesis. In *Proc. ICASSP*, pp. 4025–4028, Taipei, Taiwan, Aug. 2009.

[17] K. Tokuda, T. Masuko, B. Miyazaki, and T. Kobayashi. Multi-space probability distribution HMM. *IEICE Trans., Inf. and Syst.*, Vol. E85-D, No. 3, pp. 455–464, 2002.

[18] 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.

[19] 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.

[20] J. Kominek and A. W Black. The CMU ARCTIC speech databases for speech synthesis research. In *Tech. Rep. CMU-LTI-03-177*, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, U.S.A, 2003.

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

[22] 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., Sept. 2006.

[23] 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.

[24] 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.