PROSODY-CONTROLLABLE HMM-BASED SPEECH SYNTHESIS USING SPEECH INPUT

Yuri Nishigaki, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan
Email: {shinnosuke-t, tomoki}@is.naist.jp

ABSTRACT

This paper proposes an HMM-based speech synthesis system that makes it possible to control the prosody of the synthesized speech through speech input. As creative activities using speech synthesis technologies have been rapidly growing in popularity, there is great demand for interfaces to synthesize speech of a specific target speaker as the users want. The proposed system allows the users to guide prosody of synthetic speech of the target speaker by using their own speech while preserving the original functionality of the HMM-based speech synthesis as a text-to-speech synthesis system. The proposed system consists of 3 main modules: a duration determination module, a F_0 modification module, and a speech parameter generation module. The first 2 modules ensure that the duration and F_0 of the input speech are reflected in the synthetic speech, and the last module generates synthetic speech parameters according to the determined duration. We examine properties of each module on speech quality and prosodic mimicking ability of synthetic speech, with experimental resulting demonstrate the effectiveness of the proposed system.

***Index Terms*—** HMM-based speech synthesis, prosody modification, duration unit, model adaptation, unvoiced/voiced information,

1. INTRODUCTION

Various techniques of corpus-based speech synthesis have achieved remarkable improvements in terms of speech quality and reproduction accuracy of speaker individuality in synthetic speech [1, 2, 3]. Due to this innovation, there is now a high demand for systems to synthesize speech or singing voices of a specific target speaker, e.g., text-to-speech synthesis system CeVIO [4] and singing voice synthesis system VOCALOID [5]. Such systems are being more and more widely utilized for speech-based creative activities.

Text-to-speech is a fundamental technique for speech-based creative activity. Especially, HMM-based speech synthesis [6, 7] is used because this makes it possible to flexibly control the voice characteristics and speaking style of synthetic speech. Compared to systems that manually control speech features [8], HMM-based speech synthesis allows for higher controllability through controlling the model parameters [9]. On the other hand, this sort of flexible control is not easy for novice users because it is still difficult to control the voice characteristics as the users want.

In this paper, we consider the use of speech information uttered by the users as supportive information for manual control. For example, [10] proposed a method for automatic parameter optimization of VOCALOID referring the user's input singing voice. This system allows the users to reflect the voice characteristics of their input singing voice on the synthesized singing voice.

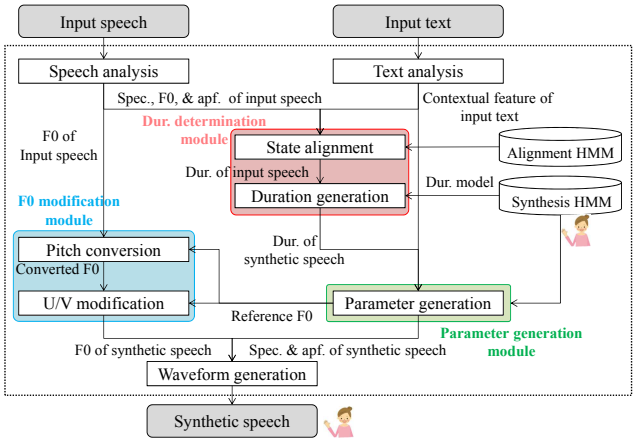

Fig. 1. An overview of the proposed system. Spec., apf., and dur. indicate spectral parameters, aperiodicity, and duration, respectively.

As a related work in speech synthesis, voice conversion techniques [11] can be utilized to convert users' input voices into the target speaker's voice. Because voice conversion can preserve the input prosody, such a system can synthesize the target speaker's voice with any variety of prosody. However, speech quality and speaker individuality in speech created by voice conversion tend to be worse than speech created by HMM-based speech synthesis. To address this problem, [12] proposed a system to reflect input prosody utilizing the HMM-based speech synthesis framework. Although this system makes it possible to control the prosody by input speech while preserving the quality of HMM-based speech synthesis, it is not guaranteed that this system works similarly to standard HMM-based speech synthesis systems because this system needs to use contextual features including the input prosody information. We assume that we can create an easier-to-control system if both prosody control by input speech and the original functions of HMM-based speech synthesis are available, which means that input speech-based and manual control are both available.

In this paper, we propose an HMM-based speech synthesis system with prosody modification based on speech input that preserves the original functions of HMM-based speech synthesis. The proposed system shown in Fig. 1 consists of 3 main modules: (1) a **duration determination module** to determine duration of synthetic speech using HMM-state alignments, (2) an F_0 **modification module** to generate the F_0 contour by modifying that of the input speech, and (3) a **speech parameter generation module** to generate spectral parameters and aperiodicity of synthetic speech. The first 2 modules ensure that the duration and F_0 of the input speech to are reflected on the synthetic speech, and standard HMM-based speech synthe-

sis is also available by using only the last module. This paper investigates the influence of each module on the speech quality and prosody mimicking ability of synthetic speech. The experimental results demonstrate the effectiveness of the proposed system.

2. HMM-BASED SPEECH SYNTHESIS

In the training stage, natural speech parameters (spectral parameters, the F_0 contour, and aperiodicity) sequences are modeled with context-dependent HMMs. Introducing HMM-state duration models, these speech parameters and the HMM-state duration are modeled in a unified training framework [13]. In the synthesis stage, HMMs corresponding to input text are constructed from the trained context-dependent HMMs, and the HMM-state duration $\boldsymbol{q} = [q_1, \cdots, q_t, \cdots, q_T]$ is determined by maximizing the duration likelihood as follows:

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

where T is a total number of frames, q_t is a HMM-state index at frame t, and $\boldsymbol{\lambda}$ indicates the parameter sets of the HMMs. The speech parameter sequence of synthetic speech is generated by maximizing HMM likelihood under the explicit constraint between static and dynamic features [14] as follows:

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

where $\boldsymbol{y} = \left[\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top\right]^\top$ is a speech parameter sequence. $\boldsymbol{y}_t = \left[y_1, \cdots, y_d, \cdots, y_D\right]^\top$ is a D-dimensional speech feature vector at frame t, $\boldsymbol{W}$ is the weighting matrix for calculating the dynamic features [7]. The synthetic speech waveform is synthesized by a vocoding process using the generated speech parameters. Although the speech quality in synthetic speech tends to be worse than to natural speech, the speech quality is improved by considering global variance in the parameter generation [15].

3. PROSODY-CONTROLLABLE HMM-BASED SPEECH SYNTHESIS USING SPEECH INPUT

This section describes how the proposed system controls the prosody of synthetic speech. Here, we call a person who utters input speech as the *input speaker*, and also call the HMMs used for state alignment and synthesizing waveform (i.e., the target speaker's HMMs) as the *alignment HMMs* and the *synthesis HMMs*, respectively.

3.1. Duration Determination Module

In the duration determination module, we first construct sentence HMMs corresponding to the input text using the alignment HMM, then perform HMM-state alignment to extract the state duration of the input speech. The state duration of synthetic speech is determined given the duration of input speech. In this paper, we address the 3 following problems of this module.

3.1.1. Duration Generation Using HMM-state Alignments of Input Speech

This module requires alignment HMMs that accurately model input speech features for robust alignment. We suppose two cases using the input speaker's pre-recorded speech: (1) input speaker-dependent HMMs can be trained from a large amount of the input speaker's pre-recorded speech, and (2) the target speaker's HMMs are adapted [16] using a small amount of input speaker's speech. Note that a speaker-independent or an average-voice model [16] could be adopted.

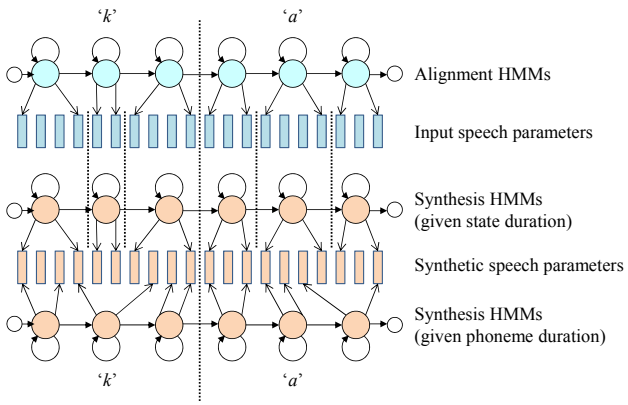

Fig. 2. An example of state or phoneme-level duration mapped from input speech to synthetic speech. HMM-state duration is redistributed by the duration models of the synthesis HMMs when the phoneme duration is given.

3.1.2. Better Duration Units

Because the alignment and synthesis HMMs are trained using other speakers' speech features, it is not guaranteed that their HMM-states model the corresponding speech segment[1]. Therefore, state-level duration mapping is expected to deteriorate the speech quality of synthetic speech. In this paper, we consider 3 types of the duration unit mapped from the input speech to the synthetic speech: state, phoneme, and mora duration. When the phoneme or mora duration is used, the state duration of synthetic speech is determined by maximizing the duration likelihood given the duration of each phoneme as shown in Fig. 2.

3.1.3. Robustness to Various Speaking Style of Input Speech

As we described in Section 1, this system is expected to be used for speech-based creative activities. In such a case, voice characteristics of input speech vary widely, and are not always the same as those of the target speaker's speech used for training the synthesis HMMs. The mismatch between voice characteristics causes a degradation in accuracy. In order to address this problem, we adapt the alignment HMMs using input speech features. The alignment process is performed utterance by utterance.

3.2. F_0 Modification Module

This module generates an F_0 contour of synthetic speech using the input speaker's F_0 contour with a reference F_0 contour generated from the target speaker's HMMs (synthesis HMMs) in the parameter generation module.

3.2.1. Pitch Conversion

To correct the F_0 range difference between input speaker's and target speaker's F_0 contours, the input speaker's F_0 contour is linearly converted as follows:

$$\hat{x}_t = \frac{\sigma_y}{\sigma_x}\left(x_t - \mu_x\right) + \mu_y, \tag{3}$$

[1] For example, a central sub-segment of a speech segment of one phoneme is assumed to be modeled with the 2nd HMM-state of the alignment HMMs, but there is no guarantee that it is modeled with the same HMM-state of the synthesis HMMs.

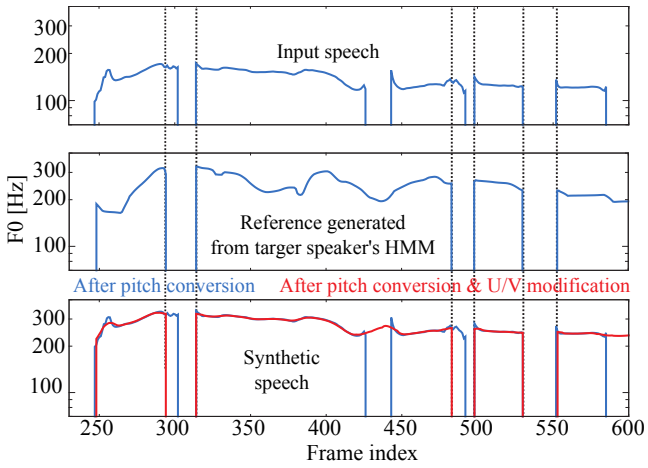

Fig. 3. An example of F_0 contours of input speech, reference generated from target speaker's HMMs, and synthetic speech.

where x_t is the log-scaled F_0 value of input speech at frame t, μ_x and σ_x are mean and standard deviation of x_t, μ_y and σ_y are those of the target speaker, respectively. These statistics are calculated utterance by utterance.

3.2.2. Unvoiced/Voiced Region Modification

Because the input speaker's and target speaker's reference F_0 contour have the unvoiced/voiced regions related to their own spectral parameters, we expect speech quality of synthetic speech to suffer from mismatch between the input speaker's U/V region and the target speaker's spectral parameters[2]. Therefore, we perform U/V region modification after the pitch conversion process. An example is shown in Fig. 3. A continuous F_0 contour [17, 18] is first estimated by spline-based interpolation using the F_0 contour of input speech. Micro prosody [19] is removed during this process. The finally determined F_0 contour is generated by restoring the U/V region of the reference F_0 contour.

3.3. Speech Parameter Generation Module

This module generates the spectral parameters, F_0 contour, and aperiodicity given the determined duration. The generated F_0 is used for the F_0 modification module as described. The synthetic speech is synthesized using the spectral parameters and aperiodicity of this module and the F_0 contour of F_0 modification module.

Power is an essential component to reflect emphasis of input speech on synthetic speech. The straightforward approach is to replace the generated power with that of the input speech, but there is a trade-off between the degree to which is reflected and unnatural emphasis caused by U/V modification described in the F_0 modification module. Therefore, we perform Maximum APosterior (MAP) adaptation [20] of synthesis HMMs using input speech parameters. The degree of power reflection can be controlled by the hyper parameter of the MAP adaptation. For example, setting the hyper parameter to 0 effectively reflects the power of input speech but causes unnatural emphasis.

[2]Especially, speech quality is strongly degraded when spectral parameters at voiced frame is excited by unvoicing noise signals.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Setup

We trained a synthesis Hidden Semi-Markov Model (HSMM) [21] for an Japanese female speaker as the target speaker. We used 450 sentences from subset A-through-I of ATR phoneme-balanced sentences [22] for training. 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 [23, 24] were extracted as excitation parameters. The STRAIGHT analysis-synthesis system [25] 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. 2 male and 2 female speakers different from the target speaker were used for input speakers. The input-speaker dependent HMMs for alignments were trained using 450 sentences from subset A-through-I of ATR phoneme-balanced sentences uttered by each input speaker.

4.2. Evaluation of Duration Determination Module

4.2.1. Evaluation for Alignment Accuracy and Duration Unit

We investigate effects of the duration unit and adaptation of alignment HMMs. We prepared HMM-state ("state"), phoneme ("phone"), and mora ("mora") duration as the duration unit, and also prepared input-speaker-dependent HMMs ("Ref."), HMMs adapted using 1 utterance ("1 utt.") or 56 utterances ("56 utts.") uttered in advance by input speaker. Additionally, the target speaker's HMMs were used to test accuracy when HMMs were not adapted ("Target"). We conducted a MOS test on speech quality of synthetic speech and a DMOS test on prosodic mimicking ability, using combinations of 3 duration units and 4 alignment HMMs (a total of 12 combinations). In the MOS test, the synthetic speech of each combination was presented to listeners in random order. In the DMOS test, input speech was first presented as a reference, then synthetic speech was presented. 10 listeners rated the speech quality using a 5-point scale in the MOS test. Similarly, 14 listeners rated to what extent the prosody of input speech is reflected on synthetic speech using a 5-point scale in the DMOS test. The 53 sentences uttered by each input speaker was used for the evaluation data. U/V modification was performed in the F_0 modification module, but power adaptation was not performed in the parameter generation module.

The results are shown in Figures 4 and 5. Comparing scores between the adapted HMMs ("1 utt." and "56 utts.") and the non-adapted HMM ("Target"), the scores of the adapted HMMs achieved better score, and "56 utts." has a similar score to "Ref." on speech quality and prosodic mimicking ability, excepting when using mora duration ("mora") in the DMOS test. These results demonstrate that adaptation of the alignment HMMs is effective on both speech quality and prosodic mimicking ability.

Mora duration tends to deteriorate the prosodic mimicking ability compared to other duration units. This is because the HMM-state duration of synthetic speech is strongly affected by the duration models of synthesis HMMs when the mora duration is used.

Additionally, we conducted a t-test between the best score ("Ref." HMMs with phoneme duration) and "1 utt." with each duration unit. We confirmed that there is a significant difference at the 1% level when state duration and mora duration were used on mimicking ability evaluation, and state duration was used on speech quality, suggesting that phoneme duration is the most effective.

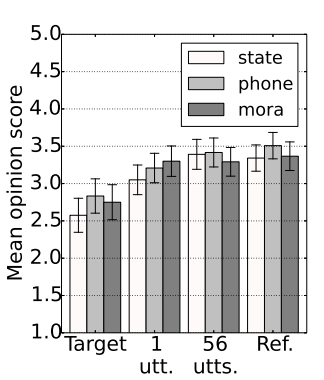

Fig. 4. Mean opinion scores on speech quality of synthetic speech to investigate the effect of the duration unit and alignment HMM adaptation. (Error bars indicate 95% confidence intervals.)

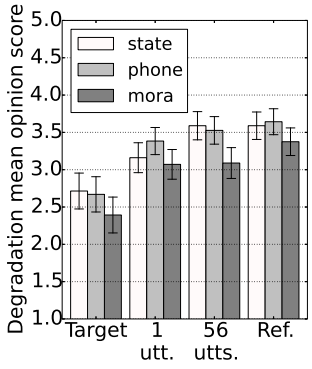

Fig. 5. Degradation mean scores on prosodic mimicking ability of synthetic speech to investigate the effect of the duration unit and alignment HMM adaptation. (Error bar indicate 95% confidence intervals.)

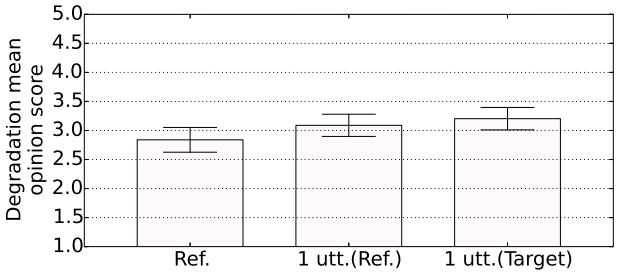

Fig. 6. Degradation mean opinion scores with 95% confidence with various speaking style of input speech.

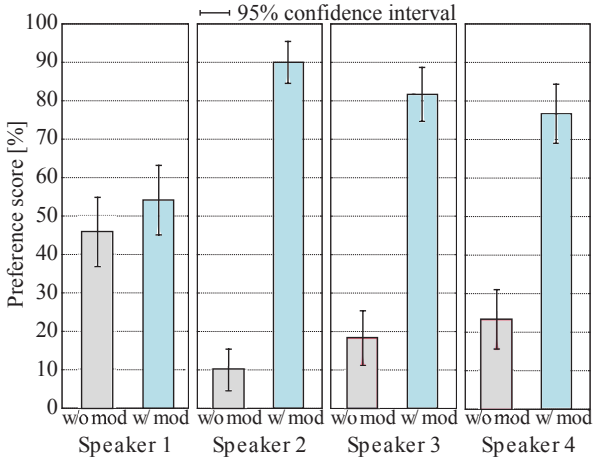

Fig. 7. Preference scores on speech quality between F_0 contours before and after modifying unvoiced/voiced region.

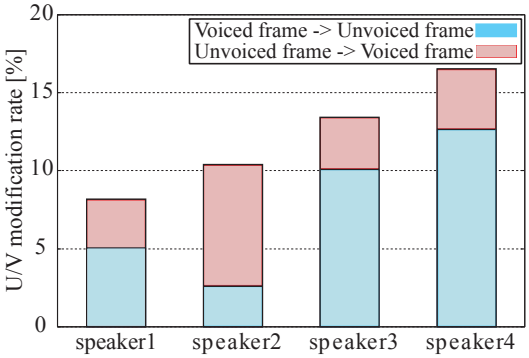

Fig. 8. U/V modification rates between F_0 contours before and after modifying the U/V region.

4.2.2. Evaluation of Robustness to Various Speaking Styles of Input Speech

We prepared another evaluation data to investigate effectiveness when the input speaker uses various, possibly extreme, speaking styles. We first extracted 10 famous phrases of Japanese television dramas and animation, and then recorded input speaker's voices, asking them to imitate the extracted phrases. As the alignment HMMs, input speaker-dependent HMMs ("Ref.") trained using speaking speech and "Ref." HMMs and "Target" HMMs adapted using the input speech were used. 8 listeners participated in a DMOS test on prosodic mimicking ability. U/V modification was performed in the F_0 modification module, but power adaptation was not performed in the parameter generation module.

Figure 6 shows the result. The adapted HMMs have a better score than the speaker-dependent HMMs trained using regularly-spoken speech. This result demonstrates that adaptation using input speech is effective to improve robustness to various speaking style of input speech. "1 utt. (Ref.)" and "1 utt. (Target)" use the different HMMs to be adapted, but there is no significant difference on mimicking ability. Therefore, we can find this adaptation process works robustly with the HMMs to be adapted.

4.3. Evaluation of F_0 Modification Module

Next, we investigated effectiveness of the U/V modification in the F_0 modification module. We conducted a preference AB test on speech quality, using F_0 contours with and without U/V modification ("w/ mod" and "w/o mod"). The alignment HMMs and the duration unit were input-speaker-dependent HMMs and phoneme duration, respectively. Moreover, we calculated the U/V modification ratio before and after U/V modification.

Figures 7 and 8 show the result for speech quality and U/V modification rate, respectively. We can see that the U/V modification achieves quality improvements for 3 input speakers. Moreover, we can find the largest improvements for speaker 2, likely due to the fact that speaker 2 had the largest ratio modifying unvoiced frames to voiced frames. Because exciting spectral parameters at the voiced frame by unvoicing noise signals causes significant quality degradation as described in Section 3, the significant improvements for speaker 2 are natural.

4.4. Evaluation of Speech Parameter Generation Module

Finally, we investigated the effectiveness of power adaptation of the synthesis HMMs. The 10 phrases in Section 4.2.2 were used for evaluation. The hyper parameter was set to 1.0 from the preliminary evaluation. We conducted a preference XAB test on prosodic mimicking ability, using the adapted HMMs ("w/ power") and non-adapted HMMs ("w/o power"). Before the adaptation, we replaced the 0th mel-cepstral coefficient with the log-scaled power and re-trained the synthesis HMMs. Input speech was first presented, then

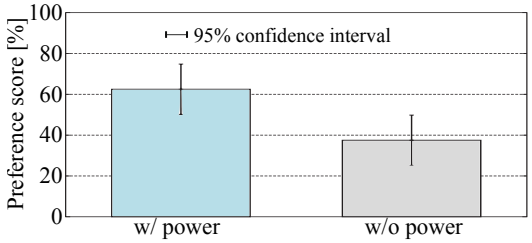

Fig. 9. Preference scores on prosodic mimicking ability to confirm the effectiveness of the power component adaptation.

pairs of synthetic speech sample was presented. The listener selected a better sample in term of prosody mimicking ability. As shown in Fig. 9, we can confirm the effectiveness of power adaptation, with the adapted HMMs achieving a better score than the non-adapted HMMs.

5. CONCLUSION

This paper has proposed HMM-based speech synthesis with prosody modification based on speech input. The proposed system consists of the duration determination, F_0 modification, and speech parameter generation module. The experimental results have demonstrated the properties of each module on naturalness and prosody mimicking ability of synthetic speech. As future work, we will try HMM selection considering speech input.

Acknowledgements: Part of this work was supported by JSPS KAKENHI Grant Number 26280060 and 2430073.

6. REFERENCES

[1] 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.

[2] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis. In *Proc. ICASSP*, Brisbane, Australia, Apr. 2015.

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

[4] Cevio creative studio http://cevio.jp/.

[5] H. Kenmochi and H. Ohshita. Vocaloid - commercial singing synthesizer based on sample concatenation. In *Proc. INTERSPEECH*, pp. 4011–4012, Antwerp, Belgium, Aug. 2007.

[6] 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.

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

[8] M. Abe, H. Mizuno, O. Mizuno, Y. Noda, S. Takahashi, and S. Nakajima. Speech design tool: Sesign. *IEICE Trans., Inf. and Syst.*, Vol. J84-D-II, No. 6, pp. 927–935, Jun. 2001.

[9] 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.

[10] T. Nakano and M. Goto. VOCALISTENER: A singing-to-singing synthesis system based on iterative parameter estimation. In *Proc. SMC*, pp. 343–348, Porto, Portugal, Jul. 2009.

[11] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[12] T. Nose, Y. Ota, and T. Kobayashi. HMM-based voice conversion using quantized F0 context. *IEICE Trans., Inf. and Syst.*, Vol. E93-D, No. 9, pp. 2483–2490, Sep. 2010.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A hybrid approach to electrolaryngeal speech enhansement based on spectral subtraction and statistical voice conversion. In *Proc. INTERSPEECH*, pp. 3067–3071, Lyon, France, Sep. 2013.

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

[20] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Transactions on audio, speech, and language processing*, Vol. 17, No. 1, pp. 66–83, 2009.

[21] 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.

[22] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara. ATR technical report. No. TR-I-0166M, 1990.

[23] 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.

[24] 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.

[25] 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.