IEICE TRANS. INF. & SYST., VOL.E102–D, NO.6 JUNE 2019

LETTER

Prosody Correction Preserving Speaker Individuality for Chinese-Accented Japanese HMM-Based Text-to-Speech Synthesis

Daiki SEKIZAWA[†a)], **Shinnosuke TAKAMICHI**[†b)], ***Nonmembers***, ***and*** **Hiroshi SARUWATARI**[†c)], ***Member***

SUMMARY This article proposes a prosody correction method based on partial model adaptation for Chinese-accented Japanese hidden Markov model (HMM)-based text-to-speech synthesis. Although text-to-speech synthesis built from non-native speech accurately reproduces the speaker's individuality in synthetic speech, the naturalness of the synthetic speech is strongly degraded. In the proposed model, to improve the naturalness while preserving the speaker individuality of Chinese-accented Japanese text-to-speech synthesis, we partially utilize HMM parameters of native Japanese speech to synthesize prosody-corrected synthetic speech. Results of an experimental evaluation demonstrate that duration and F_0 correction are significantly effective for improving naturalness.

key words: *HMM-based text-to-speech synthesis, non-native speech, Chinese-accented Japanese, prosody*

1. Introduction

Text-to-speech synthesis is a method to artificially synthesize speech from text. Hidden Markov model (HMM) [1], deep neural network [2]-based, and end-to-end [3] ones are often used for synthesizing natural speech of the desired text and speaker. Synthesizing non-native speech is a challenging but important task to establish computational theories of a variety of languages and speech. Although acoustic models built from non-native speech can accurately reproduce the speaker's individuality in synthetic speech, the naturalness of the synthetic speech is strongly degraded due to the language system differences between the spoken language and the speaker's mother tongue.

To improve the naturalness of Japanese-accented English (i.e., English spoken by Japanese) text-to-speech synthesis, Oshima et al. proposed prosody correction method that preserve speaker individuality [4]. Using frameworks of HMM-based text-to-speech synthesis and model adaptation [5], the HMM parameters are *partially* updated to fit the non-native speaker's speech parameters while fixing the remaining HMM parameters of the native speaker's speech. This method has successfully improved the naturalness in synthetic speech thanks to focusing on the prosody system difference between Japanese and English. Since the differences are dependent on the language pairs, investigating whether this framework can be applied to other non-native speech is an intriguing task.

In this paper, we apply Oshima et al.'s method to Chinese-accented Japanese (i.e., Japanese spoken by Chinese) text-to-speech synthesis. More and more native Chinese-speakers are speaking Japanese every year, so correcting and synthesizing Chinese-accented speech is natural to target. Considering prosody system differences between Japanese and Chinese, we empirically investigate a prosody correction method that preserves speaker individuality. Furthermore, we also investigate the use of other correction methods that are not investigated by Oshima et al. [4]. The experimental result demonstrates that duration and F_0 correction significantly improve the naturalness while preserving speaker individuality, regardless of the non-native speakers' level of Japanese proficiency.

In Sect. 2 of this paper, we briefly describe HMM-based text-to-speech frameworks and conventional prosody correction methods for Japanese-accented English text-to-speech synthesis. Section 3 reviews prosody system differences between Japanese and Chinese and proposes the prosody correction methods for Chinese-accented Japanese text-to-speech synthesis. Section 4 empirically investigates which correction method is most effective. We conclude in Sect. 5 with a brief summary and mention of future work.

2. Prosody Correction for Japanese-Accented English Text-to-Speech Synthesis [4]

2.1 HMM-Based Text-to-Speech Synthesis and Model Adaptation

HMM-based text-to-speech synthesis [1] is a framework to simultaneously model spectrum, excitation, and HMM state duration. The output probability distribution of the c-th HMM state is

$$b_c(\boldsymbol{Y}_t) = \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \tag{1}$$

where $\boldsymbol{Y}_t$ is a feature vector consisting of static and dynamic speech features at frame t. $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the mean vector and covariance matrix of the Gaussian distribution $\mathcal{N}(\cdot;\cdot,\cdot)$ of the c-th HMM state. The HMM state duration is also modeled with the Gaussian distribution in a type of HMM called a *hidden semi-Markov model (HSMM)*.

Model adaptation for HMM-based text-to-speech [5] is a technique that builds the target speaker's HSMMs by transforming pre-trained HSMM parameters using the target speaker's speech data. In this work, we adopt the CMLLR

Manuscript received December 18, 2018.
Manuscript revised January 15, 2019.
Manuscript publicized March 11, 2019.

[†]The authors are with the University of Tokyo, Tokyo, 113–8656 Japan.
a) E-mail: sekizawa-daiki963@g.ecc.u-tokyo.ac.jp
b) E-mail: shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
c) E-mail: hiroshi_saruwtari@ipc.i.u-tokyo.ac.jp

DOI: 10.1587/transinf.2018EDL8264

adaptation method [5], which transforms $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ as

$$\boldsymbol{\mu}'_c = \boldsymbol{A}\boldsymbol{\mu}_c + \boldsymbol{b}, \tag{2}$$

$$\boldsymbol{\Sigma}'_c = \boldsymbol{A}\boldsymbol{\Sigma}_c\boldsymbol{A}^\top, \tag{3}$$

where $\boldsymbol{A}$ and $\boldsymbol{b}$ are the transformation matrix and bias vector estimated using the target speaker's feature vectors. Note that, model parameters for spectrum, excitation, and duration can be adapted in the same manner.

2.2 Prosody Correction for Japanese-Accented English

Though Japanese has mora (sub-syllable)-timed isochrony and is a pitch-accented language, English has stress-timed isochrony and is a stress-accented language. Therefore, the stress and duration of Japanese-accented English speech are significantly different from those of native English speech. Correction of such features for Japanese-accented English text-to-speech synthesis can be done by partial adaptation of HSMMs [4]. First, a native English speaker's HSMMs are trained using the speaker's speech data. Then, the model parameters are adapted using the non-native English (i.e., Japanese-accented English) speech of the non-native speaker. In adaptation, model parameters of the HMM state duration and power are not updated, and those of native speaker's speech are fixed. The synthesis procedure is done in the standard manner. Since the duration and power of the synthesized speech are equal to those of the native English speaker, we can accurately synthesize speech reflecting a native speaker's rhythm and stress. Also, since other model parameters (such as spectrum and F_0) are adapted, the synthetic speech retains the non-native speaker's individuality.

3. Prosody Correction for Chinese-Accented Japanese Text-to-Speech Synthesis

We apply the prosody correction discussed in Sect. 2 to Chinese-accented Japanese text-to-speech synthesis. First, a native Japanese speaker's HSMMs are trained using the speaker's speech, and then the model parameters are partially adapted using the non-native (i.e., Chinese-accented) Japanese speech.

Since Chinese has syllable-timed isochrony and is a tonal language, we expect that correction of pitch (i.e., F_0) and rhythm (i.e., duration) will improve the naturalness of the Chinese-accented Japanese text-to-speech synthesis. Furthermore, inspired by [6], this paper investigates temporal delta features of spectrum and prosody. The below is a list of model parameters to be fixed.

1. Delta feature of F_0
2. Delta feature of mel-cepstral coefficients
3. Power [4]
4. HMM state duration
5. F_0

Correction of parameters 1-through-3 is shown in Fig. 1. Duration correction (4) is not included in the figure but is done in the same manner. For F_0 correction (5), the native Japanese speaker's F_0 is generated in synthesis first, and then the log-scaled F_0 is linearly transformed [7] to retain non-native speaker's F_0 ranges.

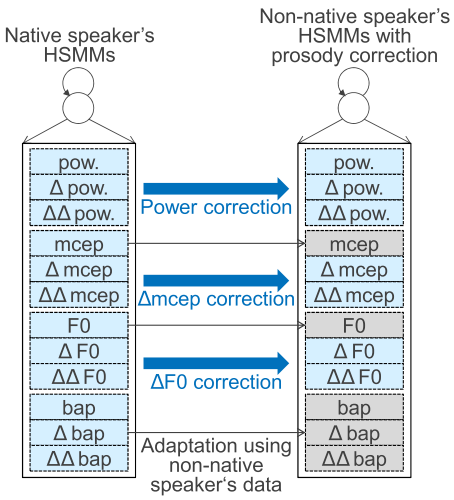

Fig. 1 Proposed correction. "pow," "mcep," and "bap" indicate power, mel-cepstral coefficient, and band-aperiodicity, respectively. Δ represents delta features.

4. Experimental Evaluation

4.1 Experimental Conditions

We used 5,000 sentences from the JSUT corpus (a speech corpus uttered by a single native Japanese speaker) [8] as the native Japanese speaker's speech data. Non-native speakers were four female speakers (labeled F1, F2, F3, F4) selected from the UME-JRF corpus [9]. The amount of adaptation data for each non-native speaker was approximately 220 sentences (the number of sentences varied from speaker to speaker), and the test data comprised 30 sentences not included in the training and adaptation data. To evaluate the performance of the proposed method for a variety of non-native Japanese proficiency levels, we selected non-native speakers who ranked at low, middle, and high proficiency levels. The UME-JRF corpus includes the Japanese proficiency scores in several terms. We averaged the scores for each non-native speaker and define a scalar value for each. The averaged scores (1–5) were 1.50 (F1), 2.6 (F2), 3.2 (F3), and 4.05 (F4). Speech signals were sampled at 16 kHz. The log-scaled power and the 1st-through-39th mel-cepstral coefficients were extracted as spectral parameters, and log-scaled F_0 and five band-aperiodicity [10] were extracted as excitation parameters by STRAIGHT [11], [12]. The feature vector consists of spectral and excitation parameters and their delta and delta-delta features. Five-state left-to-right HSMMs were used. The log-scaled power and the mel-cepstral coefficients were trained in the same stream. The block diagonal matrix corresponding to static, delta, and delta-delta parameters was used as the linear transform for adaptation. Before training and adaptation, 50 Hz-cutoff

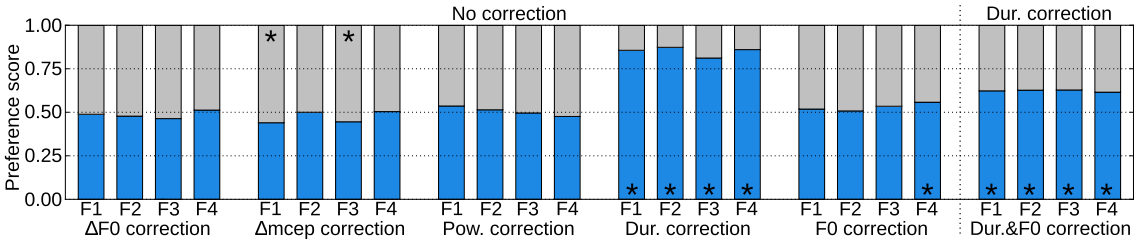

Fig. 2 Preference scores on naturalness. "∗" indicates a preferred method with the p-value smaller than 0.05. "Pow." and "Dur." indicate power and duration, respectively.

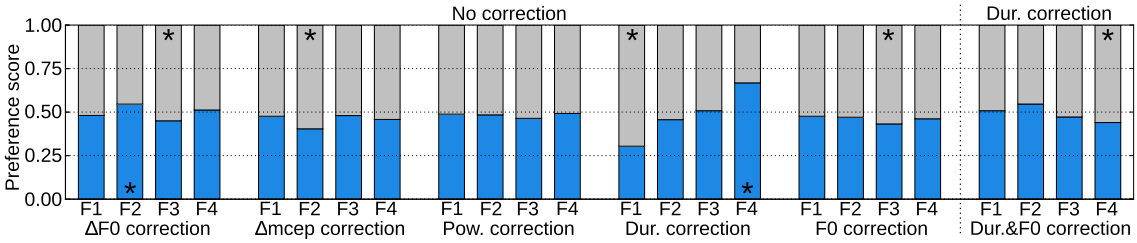

Fig. 3 Preference scores on speaker similarity. "∗" indicates a preferred method with the p-value smaller than 0.05. "Pow." and "Dur." indicate power and duration, respectively.

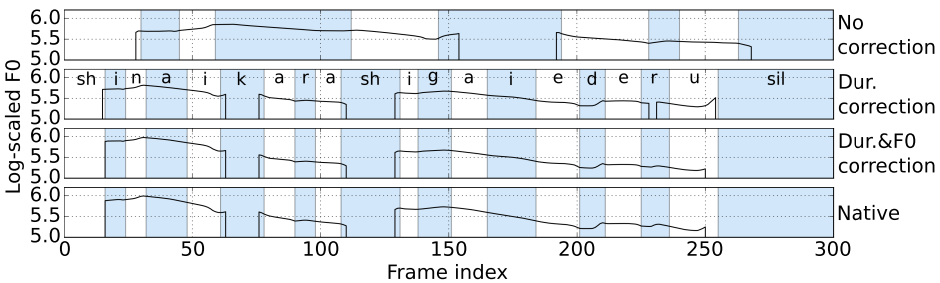

Fig. 4 Example of generated F_0 patterns. The sentence is "shinai kara shigai e deru" (tokenized by words). "Dur." indicates duration.

speech parameter trajectory smoothing [13] were applied to the mel-cepstral coefficients. In synthesis, speech parameter generation considering global variance [14] was used.

We evaluated the following systems.

- No correction: All model parameters were adapted using the target non-native speaker.
- Correction: The model parameters were partially adapted/fixed (five patterns as listed in Sect. 3).

We evaluated the naturalness and speaker individuality of the speech synthesized by these systems. The evaluation was performed for each of the system-pairs and non-native speakers. Preference AB or XAB tests were conducted to evaluate the naturalness and speaker individuality, respectively. The reference speech of the XAB test was the non-native speaker's natural speech. The evaluation was conducted with our crowdsourcing evaluations system. We used *Lancers* [15], which is one of the largest crowdsourcing services in Japan. The number of listeners was varied in each evaluation but at least 25 Japanese listeners participated in each. In total, 48 tests were conducted and more than 1,200 listeners participated.

4.2 Experimental Results

The results of naturalness and speaker individuality are shown in Fig. 2 and Fig. 3, respectively. As we can see in Fig. 2, the proposed duration correction significantly improved naturalness regardless of the non-native speaker's Japanese proficiency level. Similarly, F_0 correction improved the naturalness in one non-native speaker (F4). We then conducted the same evaluation to compare duration correction and duration and F_0 correction. The result is shown on the right in Fig. 2. By combining duration correction and F_0 correction, we can further improve the naturalness. On the other hand, delta feature correction brought no significant improvements and sometimes caused significant degradation of naturalness. Also, power correction, which was effective in Japanese-accented English [4], also brought no improvements. This is because power is not dominant in Chinese and Japanese speech.

From Fig. 3, we can see that duration correction and F_0 correction did not degrade speaker similarity, excluding some cases (duration correction for speaker F1 and F_0 cor-

rection for speaker F3). Also, even when combining methods ("Dur.&F0 correction"), there was no significant degradation (excluding speaker F4). These results demonstrate that duration and F_0 correction are significantly effective for improving naturalness while preserving speaker similarity.

An example of the corrected duration and F_0 is shown in Fig. 4. We can see that a pitch contour without correction ("No correction") are significantly different from that of a native Japanese speaker ("Native"). Our duration and F_0 correction can dramatically refine the pitch contour and make it close to the native speaker's pitch contour.

5. Conclusion

We have proposed a prosody correction method for improving speech quality while preserving speaker individuality for Chinese-accented Japanese HMM-based text-to-speech synthesis. On the bases of a partial adaptation of a native speaker's HSMM, we corrected HMM state duration and F_0 models. The experimental results demonstrated that duration and F_0 correction significantly improve the naturalness while preserving speaker individuality, regardless of non-native speakers' Japanese proficiency level. As future work, we will investigate the effectiveness of this framework for other mother tongues and target languages.

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] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in Proc. ICASSP, May 2013.
[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R.A. Saurous, "Tacotron: Towards end-to-end speech synthesis," Interspeech 2017, pp.4006–4010, 2017.
[4] Y. Oshima, S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Non-native text-to-speech preserving speaker individuality based on partial correction of prosodic and phonetic characteristics," IEICE Trans. Inf. & Syst., vol.E99-D, no.12, pp.3132–3139, 2016.
[5] 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 Trans. Audio, Speech, Language Process., vol.17, no.1, pp.66–83, 2009.
[6] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," Acoustical Science and Technology, vol.33, no.1, pp.1–5, 2012.
[7] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio, Speech, Language Process., vol.15, no.8, pp.2222–2235, 2007.
[8] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis," vol.abs/1711.00354, 2017.
[9] "Japanese speech database read by foreign students (UME-JRF)," http://research.nii.ac.jp/src/UME-JRF.html.
[10] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," Proc. INTERSPEECH, Pittsburgh, U.S.A., pp.2266–2269, Sep. 2006.
[11] H. Kawahara, I. Masuda-Katsuse, and A.D. Cheveigné, "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 Communication, vol.27, no.3–4, pp.187–207, 1999.
[12] 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," MAVEBA, Firentze, Italy, pp.1–6, Sept. 2001.
[13] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura, "The NAIST text-to-speech system for the Blizzard Challenge 2015," Proc. Blizzard Challenge Workshop, Berlin, Germany, Sept. 2015.
[14] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.816–824, 2007.
[15] "Lancers," https://www.lancers.jp/.