IEICE TRANS. INF. & SYST., VOL.E99–D, NO.12 DECEMBER 2016

PAPER

Non-Native Text-to-Speech Preserving Speaker Individuality Based on Partial Correction of Prosodic and Phonetic Characteristics

Yuji OSHIMA[†], **Shinnosuke TAKAMICHI**[††a)], ***Nonmembers***, **Tomoki TODA**[†,†††b)], ***Member***, **Graham NEUBIG**[†c)], ***Nonmember***, **Sakriani SAKTI**[†d)], ***and*** **Satoshi NAKAMURA**[†e)], ***Members***

SUMMARY This paper presents a novel non-native speech synthesis technique that preserves the individuality of a non-native speaker. Cross-lingual speech synthesis based on voice conversion or Hidden Markov Model (HMM)-based speech synthesis is a technique to synthesize foreign language speech using a target speaker's natural speech uttered in his/her mother tongue. Although the technique holds promise to improve a wide variety of applications, it tends to cause degradation of target speaker's individuality in synthetic speech compared to intra-lingual speech synthesis. This paper proposes a new approach to speech synthesis that preserves speaker individuality by using non-native speech spoken by the target speaker. Although the use of non-native speech makes it possible to preserve the speaker individuality in the synthesized target speech, naturalness is significantly degraded as the synthesized speech waveform is directly affected by unnatural prosody and pronunciation often caused by differences in the linguistic systems of the source and target languages. To improve naturalness while preserving speaker individuality, we propose (1) a prosody correction method based on model adaptation, and (2) a phonetic correction method based on spectrum replacement for unvoiced consonants. The experimental results using English speech uttered by native Japanese speakers demonstrate that (1) the proposed methods are capable of significantly improving naturalness while preserving the speaker individuality in synthetic speech, and (2) the proposed methods also improve intelligibility as confirmed by a dictation test.

key words: *cross-lingual speech synthesis, English-Read-by-Japanese, speaker individuality, HMM-based speech synthesis, prosody correction, phonetic correction*

1. Introduction

According to recent improvements in synthetic speech quality [1]–[3] and robust building of speech synthesis systems [4], [5], statistical parametric speech synthesis [6] has been a promising technique to develop speech-based systems. Cross-lingual speech synthesis, which synthesizes foreign language speech with a non-native speaker's own voice characteristics, holds promise to improve a wide variety of applications. For example, it makes it possible to build Computer-Assisted Language Learning (CALL) systems that let learners listen to reference speech with their own voices [7], and speech-to-speech translation systems that output with the input speaker's voice [8].

There have been many attempts at developing cross-lingual speech synthesis based on statistical voice conversion [9] or Hidden Markov Model (HMM)-based speech synthesis [10]. For example, one-to-many Gaussian Mixture Model (GMM)-based voice conversion can be applied to unsupervised speaker adaptation in cross-lingual speech synthesis [11], [12]. In addition, cross-lingual adaptation parameter mapping [13]–[15] and cross-lingual frame mapping [16] have also been proposed for HMM-based speech synthesis. These approaches use a non-native speaker's natural voice in his/her *mother tongue* to extract speaker-dependent acoustic characteristics and make it possible to synthesize naturally sounding target language voices. However, speaker individuality in cross-lingually adapted speech

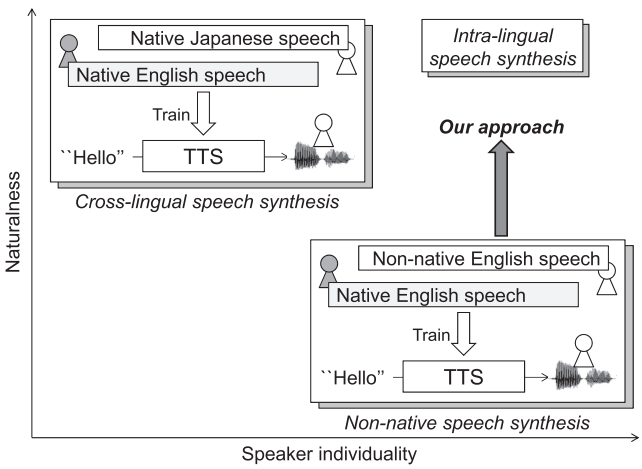

Fig. 1 Comparison of cross-lingual speech synthesis and non-native speech synthesis. The target language and non-native speaker's mother tongue are English and Japanese, respectively. Cross-lingual speech synthesis generates naturally sounding speech but the speaker individuality of the target speaker tends to be inferior to that of the intra-lingual speech synthesis. On the other hand, non-native speech synthesis can well reproduce the speaker individuality on synthetic speech but its naturalness tends to be degraded. Our approach is based on improvements of non-native speech synthesis to synthesize more naturally sounding speech than the non-native speech synthesis while preserving the speaker individuality close to that of the intra-lingual speech synthesis.

Manuscript received May 31, 2016.
Manuscript revised July 16, 2016.
Manuscript publicized August 30, 2016.

[†]The authors are with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
[††]The author is with the Department of Information Physics and Computing, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, 113–8656 Japan.
[†††]The author is with Information Technology Center, Nagoya University, Nagoya-shi, 464–8601 Japan.
a) E-mail: shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
b) E-mail: tomoki@icts.nagoya-u.ac.jp
c) E-mail: neubig@is.naist.jp
d) E-mail: ssakti@is.naist.jp
e) E-mail: s-nakamura@is.naist.jp
DOI: 10.1587/transinf.2016EDP7231

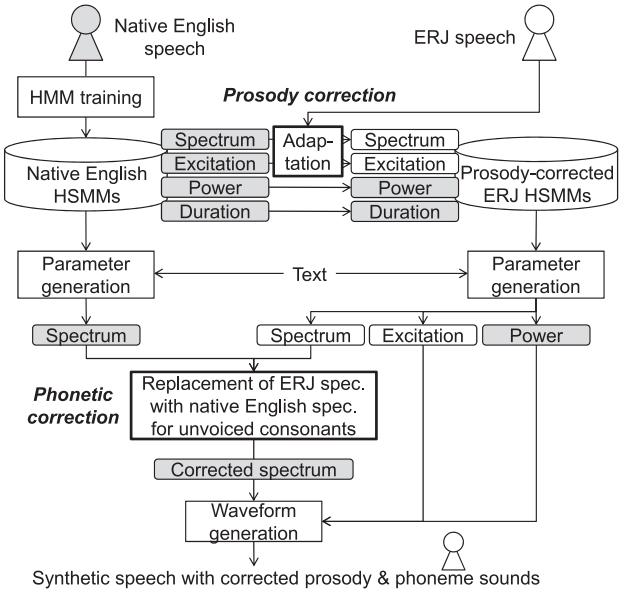

Fig. 2 An overview of the proposed non-native speech synthesis framework consisting of a prosody correction module and a phonetic correction module. In the prosody correction module, power and duration components of the native (English) speaker are copied to the non-native (ERJ) HSMMs, and spectrum of non-native speech is replaced with that of native speech.

tends to be inferior to that of intra-lingual speech synthesis.

The alternative paradigm is to explicitly collect utterances from the speaker in his/her *non-native tongue*, as shown in Fig. 1. There is a small amount of previous work in this paradigm. For example, Kajima *et al.* [17] developed an HMM-based speech synthesizer using speech features converted from a native speaker into the target non-native speaker with statistical voice conversion. However, it is known that the use of non-native speech deteriorates the naturalness in synthetic speech [18], [19].

Specifically, we focus on a particularly difficult cross-lingual case: English speech synthesis preserving a Japanese speaker's voice characteristics. Due to the large disconnect between these two languages, it has been noted that English speech read by a native Japanese speaker (English-Read-by-Japanese; ERJ [20]) is highly different from its native English counterpart due to Japanese-accented prosody or pronunciation [21], [22]. On the other hand, there is a large demand in Japan for CALL and speech translation technology, and thus overcoming these obstacles is of considerable merit.

This paper proposes a method to improve naturalness of non-native speech synthesis preserving speaker individuality based on the partial correction of prosodic and phonetic characteristics[†], inspired by the previous work on improvements of naturalness of disordered speech for creating a personalized speech synthesis system [24]. The overview of the proposed method is shown in Fig. 2.

[†]In this paper, we present results of additional experimental evaluations, more discussions, and more evaluations than those in our previous work [23].

The *prosody correction* method partly[††] adapts the native speaker's HMM parameters by using the target speaker's non-native speech. The *phonetic correction* method partly replaces the generated spectral parameters of the non-native speaker with those of the native speaker, applying replacement to only unvoiced consonants, the acoustic characteristics of which are less affected by speaker differences. The experimental results using ERJ speech demonstrate that the proposed methods are capable of improving naturalness and intelligibility of non-native speech while preserving speaker individuality.

2. HMM-Based Speech Synthesis

We adopt an HMM-based speech synthesis approach, modeling spectrum, excitation, and state duration parameters in a unified framework [25]. The output probability distribution function of the c-th HMM state is given by:

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

where $\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta\Delta\boldsymbol{c}_t^\top]^\top$ is a feature vector including a static feature vector $\boldsymbol{c}_t$ and its dynamic feature vectors $\Delta\boldsymbol{c}_t$ and $\Delta\Delta\boldsymbol{c}_t$. The vector $\boldsymbol{\mu}_c$ and the matrix $\boldsymbol{\Sigma}_c$ are the mean vector and the covariance matrix of Gaussian distribution $\mathcal{N}(\cdot; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ of the c-th HMM-state, respectively. Note that HMM state duration is also modeled by the Gaussian distribution as an explicit duration model.

Model adaptation for HMM-based speech synthesis [26] enables us to build the target speaker's HMMs by transforming the pre-trained HMM parameters using the target speaker's adaptation speech data. The transformed mean vector $\hat{\boldsymbol{\mu}}_c$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_c$ are calculated as follows:

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

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

where the transformation matrix $\boldsymbol{A}$ and the bias vector $\boldsymbol{b}$ are adaptation parameters. Usually the probability density functions are clustered into multiple classes and the corresponding adaptation parameters are applied to them. Because the spectrum, excitation, and state duration parameters are all adaptable, not only segmental features but also prosodic features can be adapted simultaneously.

In synthesis, a sentence HMM is first created based on context obtained from an input text. Then, given the HMM-state duration determined by maximizing the duration likelihood, the synthetic speech parameter sequence is generated by maximizing the HMM likelihood under the constraint on the relationship between static and dynamic features [27].

3. Proposed Partial Correction of Prosodic and Phonetic Characteristics

This section describes our proposed method for synthesizing

[††]Whereas the previous work [24] adapts spectral parameters for improvements of disordered speech, our work adapts power and duration for improvements of non-native speech as described in Sect. 3.

more naturally sounding non-native speech while preserving speaker individuality. A subset of the native speaker's HMM parameters are used to improve the naturalness of synthetic speech from the non-native speaker's HMMs.

3.1 Prosody Correction Based on Model Adaptation

The non-native speaker's HMMs are created by adapting the native speaker's pre-trained HMMs to the non-native speech data. However, the standard adaptation process transforming all HMM parameters makes synthetic speech from the adapted HMMs sound as unnatural as the original non-native speech. It is well known that large differences between ERJ speech and native English speech are often observed in duration and power [28], [29]. Therefore, we propose an adaptation process to make it possible to use the native speaker's patterns of duration and power for synthesizing more naturally sounding ERJ speech.

As the observed speech features modeled by the native speaker's pre-trained HMMs, we use log-scaled power, spectral envelope, and excitation parameters. In adaptation, the output probability density functions of only the spectral envelope and excitation parameters are adapted to the target non-native speech data in the standard manner [26], and duration and power are kept unchanged. Consequently, the adapted HMMs model the spectral envelope and excitation parameters of the target non-native speech and duration and power patterns of the native speaker[†].

3.2 Phonetic Correction Based on Spectrum Replacement for Unvoiced Consonants

The proposed phonetic correction method partly replaces generated spectral envelope parameters of the non-native speaker with those of the native speaker. Although there are many studies in speech perception [30], [31] showing the effect of the speaker differences on pitch and vowels, such studies focusing on unvoiced consonants are limited. Considering these previous studies, we expect that unvoiced consonants are less affected by speaker differences. On the other hand, pronunciation significantly affects the naturalness of non-native speech. Therefore, we can expect that replacing the spectrum of unvoiced consonants with their native counterparts may improve naturalness without causing adverse effects on speaker individuality.

First, we generate two kinds of synthetic speech parameters from the native speaker's HMMs and the non-native speaker's HMMs with corrected prosody, respectively. Note that these parameters are temporally aligned because the two HMMs share the same HMM-state duration models. Then, the non-native speaker's spectral envelope parameters corresponding to unvoiced consonants are replaced with those of the native speaker. For voiced frames aligned to HMM states for unvoiced consonants, spectral replacement is not performed, as it has the potential to reduce both naturalness and individuality. Note that it is also possible to replace not spectral features but the state output probability distributions. Although such an implementation is expected to avoid generating discontinuities caused by directly concatenating spectral parameters [16], [32], [33], we found that spectral replacement caused no significant degradation, and thus for simplicity we use it in this paper.

[†]We may choose another combination of speech parameters not to be adapted, e.g., not only duration and power patterns but also the excitation parameters. In our preliminary experiment, we found that the effect of the excitation parameters on naturalness was much smaller than that on the speaker individuality. Therefore, we decided to adapt the excitation parameters in this paper.

4. Experimental Evaluations

4.1 Experimental Conditions

We used 593 sentences spoken by a male and a female native English speaker for training and 50 sentences for evaluation from the CMU ARCTIC [34] speech database. Speech signals were sampled at 16 kHz. The log-scaled power and the 1st-through-24th mel-cepstral coefficients were extracted as spectral parameters, and log-scaled F_0 and 5 band-aperiodicity [35] were extracted as excitation parameters by STRAIGHT [36], [37]. The feature vector consists of spectral and excitation parameters and their delta and delta-delta features. 5-state left-to-right HSMMs [38] were used. The log-scaled power and the mel-cepstral coefficients were trained in the same stream. CSMAPLR + MAP [39] were used for model adaptation, and the block diagonal matrix corresponding to static parameters and their delta and delta-delta features and the bias vector were used as the linear transform for adaptation. Intra-gender adaptation was performed in adaptation from the native speakers to several non-native speakers. For comparison, we constructed a traditional GMM-based voice conversion system, which is labeled as "HMM+VC" below. We built a 64-mixture GMM for spectral parameter conversion and a 16-mixture GMM for band-aperiodicity conversion. The log-scaled F_0 was linearly converted.

We evaluate synthetic speech of the following systems:

- **ERJ**: speaker-dependent HSMMs trained using ERJ speech.
- **HMM+VC**: a GMM that converted the parameters generated from "Native" to the ERJ speech parameters [12][††]
- **Adapt**: HSMMs for which all parameters were adapted
- **Dur.**: HSMMs for which all parameters except duration were adapted
- **Dur.+Pow.**: HSMMs for which all parameters except duration and the log-scaled power were adapted
- **Dur.+Pow.+UVC**: HSMMs for which all parameters except duration and the log-scaled power were adapted and the unvoiced consonants are further corrected.
- **Native**: speaker-dependent HSMMs trained using native

[††]We adopt the one-to-one GMM-based conversion framework instead of the one-to-many framework [12].

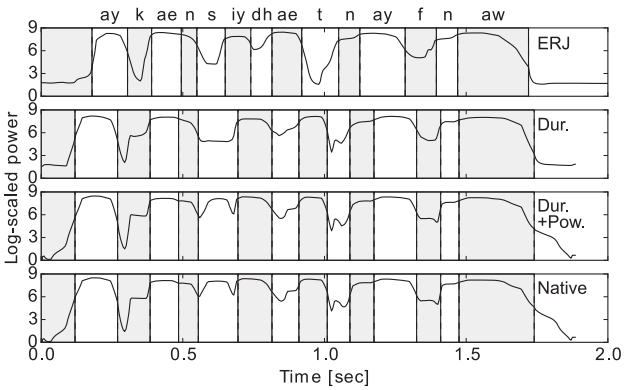

Fig. 3 An example of the power trajectories of synthesized English speech samples for a sentence "I can see that knife now." We can find the power trajectory modified by the proposed prosody modification.

English speech.

We separately investigated the effect of the proposed prosody correction method and phonetic correction method on naturalness and speaker individuality. We also investigated intelligibility of synthetic speech by the proposed methods. These evaluations were conducted using various ERJ speakers, such as male and female speakers, and speakers with high and low English proficiency levels. Six native English listeners participated in each evaluation except Sect. 4.2.1.

4.2 Evaluation of Prosody Correction

4.2.1 Effectiveness on Naturalness and Speaker Individuality and Effect of Listener's Mother Tongue

We conducted listening tests to evaluate effectiveness of the proposed prosody correction method, and to investigate influence of listener's mother tongue on the proposed method. As ERJ speech data, we used 593 CMU ARCTIC sentences uttered by 2 male Japanese students in their 20s, "Bilingual" and "Monolingual." The speaker "Bilingual" was a relatively skilled speaker who experienced a 1-year stay in Australia, and "Monolingual" was a less skilled speaker. We conducted a DMOS test on speaker individuality using all systems except "Native," and a MOS test on naturalness using all systems. Analysis-synthesized speech of the ERJ speakers was used as reference speech in the DMOS test. Also, we prepared 2 types of the listener, the 6 native Japanese and 6 native English speakers in order to investigate the effect of the listeners' mother tongue.

Figure 3 shows an example of the log-scaled power trajectory. We can see that the proposed duration correction method ("Dur.") makes duration of the ERJ speech ("ERJ") equivalent to that of the native English speech ("Native"), and the proposed duration and power correction method ("Dur.+Pow.") further makes the power trajectory of "ERJ" equivalent to that of "Native."

Figures 4 and 5 show the results of the subjective evaluation on speaker individuality and naturalness evaluated by

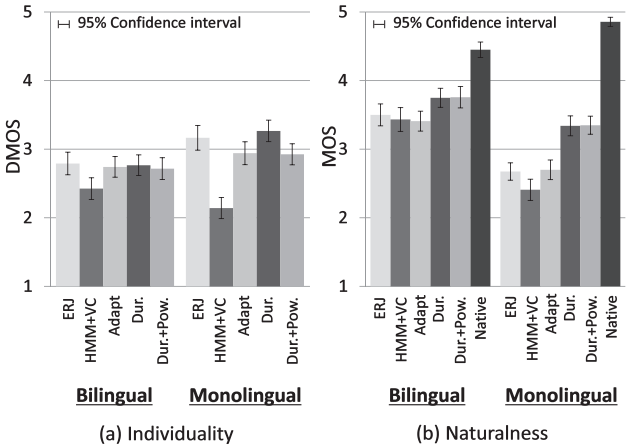

Fig. 4 Results of subjective evaluation on speaker individuality (left) and naturalness (right) using the proposed prosody correction method (evaluated by native Japanese speakers).

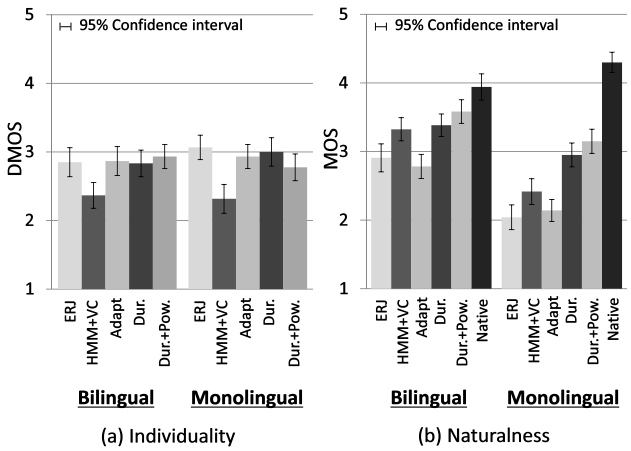

Fig. 5 Results of subjective evaluation on speaker individuality (left) and naturalness (right) using the proposed prosody correction method (evaluated by native English speakers).

native Japanese speakers and native English speakers, respectively. Compared between Fig. 4 (a) and Fig. 5 (a), we can see that the tendency of the individuality score is almost the same between listeners who have different mother tongues. On the other hand, the naturalness scores evaluated by the English speakers (Fig. 5) tend to be worse than those evaluated by the Japanese speakers (Fig. 4). Next, we focus the effect of the power correction. We can see that the differences of naturalness scores between power-corrected and non-corrected methods evaluated by the English speakers are larger than those evaluated by the Japanese speakers. We expect that this is because the English speakers are more sensitive to the stress of the synthetic speech than the Japanese speakers.

Finally, we discuss the effectiveness of the proposed prosody correction evaluated by the English speakers shown in Fig. 5. Although "HMM+VC" improves the naturalness compared to "ERJ" and the fully adapted HMMs

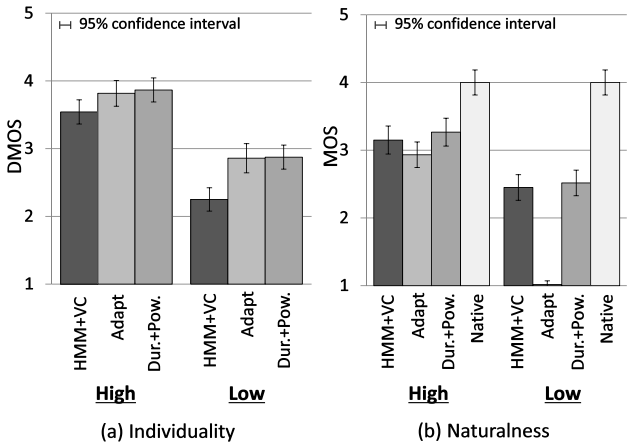

Fig. 6 Results of subjective evaluation on speaker individuality (left) and naturalness (right) using the proposed prosody correction method with ERJ speakers that have various English proficiency levels.

("Adapt"), their scores on speaker individuality decrease significantly. On the other hand, the proposed methods "Dur." and "Dur.+Pow." achieve better scores on naturalness than "ERJ" and "Adapt"[†] while maintaining the scores on speaker individuality.

4.2.2 Effects of the English Proficiency Level of ERJ Speakers

In order to investigate whether or not the proposed prosody correction method is effective for various ERJ speakers, we further conducted the MOS and DMOS tests using other ERJ speakers who have various English proficiency levels. We used TIMIT [40] sentences from the ERJ database [20] uttered by 2 male and 2 female speakers who had the best ("High") or the worst ("Low") English proficiency level, based on evaluation from various perspectives (i.e., rhythm and accent)[††]. The systems used in the DMOS test were "HMM+VC," "Adapt," and "Dur.+Pow.". Those in the MOS test were "HMM+VC," "Adapt," "Dur.+Pow.," and "Native." The system "ERJ" was not evaluated because it was similar to "Adapt" as shown in the previous evaluation.

Figure 6 shows the result of the subjective evaluations. The results are calculated for each proficiency level. In terms of speaker individuality, "Dur.+Pow." keeps scores as high as those of "Adapt." On the other hand, we can observe that "Adapt" causes a significant degradation in naturalness for the low proficiency level. The proposed method "Dur.+Pow." causes no degradation in naturalness and maintains scores as high as those of "HMM+VC." These results indicate the effectiveness of the proposed prosody correction method over various proficiency levels.

[†]There are significant differences between "Dur." and "Dur.+Pow.," "Dur." and "ERJ," and "Dur." and "Adapt" at the 1% confidence level.

[††]Multiple scores assigned to each speaker [20] were averaged to determine the best and worst English proficiency levels. We compared the scores in each gender, and chose speakers of "High" and "Low" from each gender.

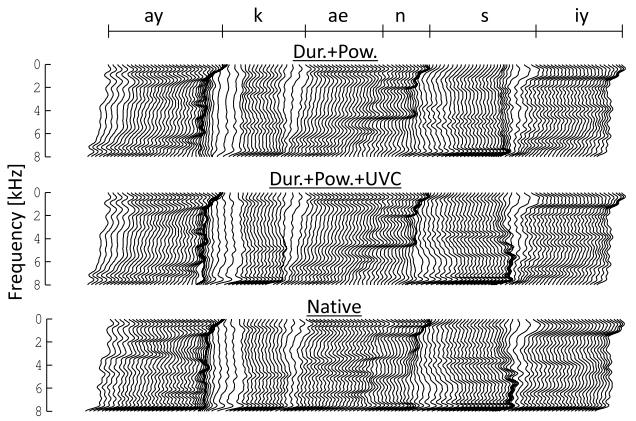

Fig. 7 An example of the spectrograms of the synthesized English speech samples. Compared the spectra of "Dur.+Pow." and "Dur.+Pow.+UVC," we can see that the spectra of /k/ and /s/ of "Dur.+Pow." are replaced with those of "Native."

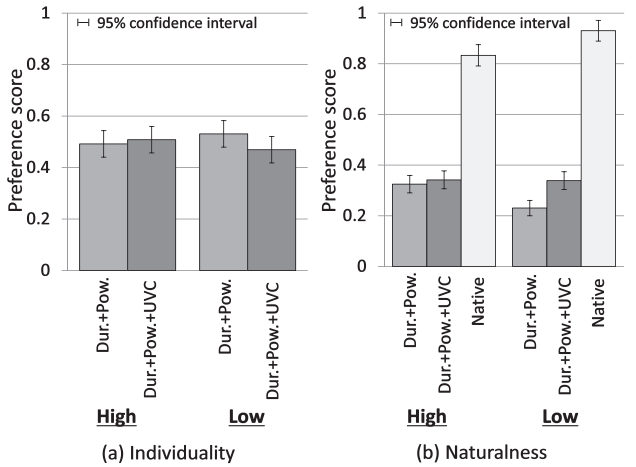

Fig. 8 Results of the subjective evaluations using the proposed phonetic correction method.

4.3 Evaluation of Phonetic Correction Method

Next, we evaluated the effectiveness of the proposed phoneme correction and its dependency on the English proficiency level of each ERJ speaker. As the ERJ speech data, we used 60 CMU ARCTIC sentences uttered by "Monolingual" and "Bilingual" from Sect. 4.2.1, and 60 TIMIT sentences uttered by 4 speakers from Sect. 4.2.2. "Bilingual" and "Monolingual" speakers were regarded as belonging to "High" and "Low" proficiency levels, respectively. We compared "Dur.+Pow." to the proposed method further correcting the phonetic characteristics ("Dur.+Pow.+UVC"). We conducted a preference XAB test on speaker individuality using "Dur.+Pow." and "Dur.+Pow.+UVC" and a preference AB test on naturalness using "Dur.+Pow.," "Dur.+Pow.+UVC," and "Native."

Figure 7 shows an example of the spectrogram. We can see that the spectral segments corresponding the unvoiced consonants (i.e., /k/ and /s/) are replaced and are the same as those of "Native." Figure 8 shows the results

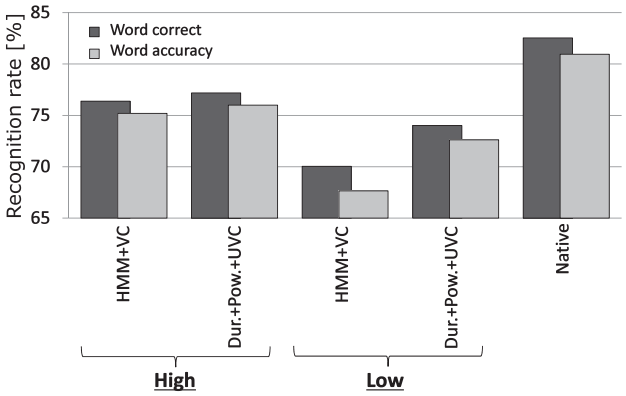

Fig. 9 Results of the dictation test for intelligibility.

of the subjective evaluation. The results of the subjective evaluation are calculated in each proficiency level. We can observe that "Dur.+Pow.+UVC" yields a better naturalness score for the low proficiency level, although there is no significant improvement for the high proficiency level. We can also observe that "Dur.+Pow.+UVC" maintains speaker individuality scores almost equal to those of "Dur.+Pow." for both high and low proficiency levels.† These results demonstrate that the proposed phonetic correction method is effective for the ERJ speakers whose English proficiency levels are low, and does not cause any adverse effects.

4.4 Evaluation of Intelligibility

To evaluate intelligibility of synthetic speech, we conducted a manual dictation test. We used the same ERJ data as used in Sect. 4.3 for training. 50 Semantically Unpredictable Sentences (SUS) [41] were used for evaluation††. Each listener evaluated 50 samples, 10 samples per system. Synthetic speech samples of "HMM+VC," "Dur.+Pow.+UVC," and "Native" were presented to the listeners in random order. The word correct rate and word accuracy were calculated for each proficiency level.

Figure 9 shows the result of the dictation test. It can be observed that "Dur.+Pow.+UVC" yields intelligibility improvements compared to "HMM+VC" for the low proficiency level (4% and 5% improvements for the word correct rate and the word accuracy, respectively). On the other hand, their scores are similar to each other for the high proficiency level. These results show that the proposed method is more effective than the conventional VC-based method in terms of intelligibility as well.

5. Conclusion

This paper has proposed a novel non-native speech synthesis technique preserving speaker individuality based on partial correction of prosodic and phonetic characteristics. The proposed prosody correction method adopted a native English speaker's acoustic models for power and duration. The proposed phonetic correction method replaced the non-native speaker's spectra with the native English speaker's spectra for unvoiced consonants. The experimental results have demonstrated that (1) the proposed methods are capable of significantly improving naturalness while preserving the speaker individuality in synthetic speech, and (2) the improvement by the proposed methods in intelligibility is also confirmed by the dictation test.

†We have found there is no significant difference between "Dur.+Pow.+UVC" and "Dur.+Pow." at the 1% confidence level.

††The SUS sentences are semantically acceptable but anomalous. Therefore, the listeners will expect the part-of-speech, but will not be able to predict more than that. Such sentences are more suitable for the dictation test than the CMU ARCTIC sentences.

Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 26280060, 26·10354, and was executed under the Commissioned Research for "Research and Development on Medical Communication Support System for Asian Languages based on Knowledge and Language Grid" of National Institute of Information and Communications Technology (NICT), Japan.

References

[1] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," Proc. ICASSP, pp.3872–3876, Florence, Italy, May 2014.
[2] 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 J. Sel. Topics Signal Process., vol.8, no.2, pp.239–250, 2014.
[3] S. Takamichi, T. Toda, A.W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," IEEE Trans. Audio, Speech, Language Process., vol.24, no.4, pp.755–767, 2016.
[4] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," IEEE Trans. ASLP, vol.17, no.6, pp.1208–1230, June 2009.
[5] P. Lanchantin, M.J.F. Gales, S. King, and J. Yamagishi, "Multiple-average-voice-based speech synthesis," Proc. ICASSP, pp.285–289, Florence, Italy, May 2014.
[6] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," Speech Commun., vol.51, no.11, pp.1039–1064, 2009.
[7] A.W. Black, "Speech synthesis for educational technology," Proc. SLaTE, pp.104–107, Farmington, PA, USA, Oct. 2007.
[8] J. Dines, H. Liang, L. Saheer, M. Gibson, W. Byrne, K. Oura, K. Tokuda, J. Yamagishi, S. King, M. Wester, T. Hirsimäki, R. Karhila, and M. Kurimo, "Personalising speech-to-speech translation: unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis," Computer Speech & Language, vol.27, no.2, pp.420–437, Feb. 2013.
[9] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. ASLP, vol.15, no.8, pp.2222–2235, Nov. 2007.
[10] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," Proc. IEEE, vol.101, no.5, pp.1234–1252, April 2013.
[11] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, and T. Dutoit, "Cross-language voice conversion based on eigenvoices," Proc. INTERSPEECH, pp.1635–1638, Brighton, UK, Sept. 2009.
[12] N. Hattori, T. Toda, H. Kawai, H. Saruwatari, and K. Shikano,

"Speaker-adaptive speech synthesis based on eigenvoice conversion and language-dependent prosodic conversion in speech-to-speech translation," Proc. INTERSPEECH, pp.2769–2772, Aug. 2011.

[13] Y. Qian, H. Liang, and F.K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS," IEEE Trans. Audio, Speech, Language Process., vol.17, no.6, pp.1231–1239, Aug. 2009.

[14] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," Proc. INTERSPEECH, pp.528–531, Brighton, UK, Sept. 2009.

[15] K. Oura, J. Yamagishi, M. Wester, S. King, and K. Tokuda, "Analysis of unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using KLD-based transform mapping," Speech Commun., vol.54, no.6, pp.703–714, 2012.

[16] Y. Qian, J. Xu, and F.K. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation," Proc. ICASSP, pp.5120–5123, May 2011.

[17] S. Kajima, A. Iida, K. Yasu, Y. Aikawa, T. Arai, and T. Sugawara, "Development of a Japanese and English speech synthesis system based on HMM using voice conversion for the people with speech communication disorder," SIG-SLP (in Japanese), vol.2008, no.12, pp.121–126, Feb. 2008.

[18] A.C. Janska and R.A. Clark, "Native and non-native speaker judgements on the quality of synthesized speech," Proc. INTERSPEECH, pp.1121–1124, Makuhari, Japan, Sept. 2010.

[19] S. Aryal and R. Gutierrez-Osuna, "Can voice conversion be used to reduce non-native accents?," Proc. ICASSP, pp.7879–7883, Florence, Italy, May 2014.

[20] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese to support CALL research," Proc. ICA, vol.1, pp.557–560, 2004.

[21] P. Ladefoged, A Course in Phonetics, Third Edition, Harcourt Brace Jovanovich College Publishers, 1993.

[22] S. Kohmoto, Applied English phonology: teaching of English pronunciation to the native Japanese speaker, Tanaka Press, Tokyo, Japan, 1965.

[23] Y. Oshima, S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Non-native speech synthesis preserving speaker individuality based on partial correction of prosodic and phonetic characteristics," Proc. INTERSPEECH, pp.299–303, Dresden, Germany, Sept. 2015.

[24] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," Acoust. Sci. technol., vol.33, no.1, pp.1–5, 2012.

[25] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. EUROSPEECH, pp.2347–2350, Budapest, Hungary, April 1999.

[26] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," IEICE Trans. Inf. & Syst., vol.E90-D, no.2, pp.533–543, Feb. 2007.

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

[28] N. Minematsu, C. Guo, and K. Hirose, "CART-based factor analysis of intelligibility reduction in Japanese English," Proc. INTERSPEECH, pp.2069–2072, Geneva, Switzerland, Sept. 2003.

[29] H. Suzuki, G. Ohyama, and S. Kiritani, "In search of a method to improve the prosodic features of English spoken by Japanese," Proc. ICSLP, pp.965–968, Kobe, Japan, Nov. 1990.

[30] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," Proc. ICSLP, vol.3, pp.1183–1186, 1994.

[31] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes and fundamental frequency contours," In Speaker Classification II, pp.157–176, Springer, 2007.

[32] D. Erro and A. Moreno, "Frame alignment method for cross-lingual voice conversion," Proc. INTERSPEECH, pp.1969–1972, Antwerp, Belgium, Aug. 2007.

[33] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," Proc. ICASSP, pp.81–84, Toulose, France, May 2006.

[34] J. Kominek and A.W. Black, "CMU ARCTIC databases for speech synthesis CMU Language Technologies Institute," CMU-LTI-03-177, Tech. Rep., 2003.

[35] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," Proc. INTERSPEECH, pp.2266–2269, Pittsburgh, PA, USA, Sept. 2006.

[36] H. Kawahara, I. Masuda-Katsuse, and A. de 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 Commun., vol.27, no.3-4, pp.187–207, April 1999.

[37] 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, pp.59–64, Sept. 2001.

[38] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A Hidden semi-Markov model-based speech synthesis system," IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.825–834, 2007.

[39] 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.

[40] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, and D.S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus," NISTIR 4930, NIST, Gaithersburg, MD, Tech. Rep., 1993.

[41] C. Benoît, M. Grice, and V. Hazan, "The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," Speech Communication, vol.18, no.4, pp.381–392, 1996.

Yuji Oshima graduated from the Department of Arts and Science, Faculty of Education, Osaka Kyoiku University in Japan, in 2013, and received his M.E. degree from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan, in 2015. His research interests include speech synthesis.

Shinnosuke Takamichi received his B.E. from Nagaoka University of Technology, Japan, in 2011 and his M.E. and D.E. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan, in 2013 and 2016, respectively. He was a short-time researcher at the NICT, Kyoto, Japan in 2013, a visiting researcher of Carnegie Mellon University (CMU) in United States, from 2014 to 2015, and Research Fellow (DC2) of Japan Society for the Promotion of Science, from 2014 to 2016. He is currently a Project Research Associate of the University of Tokyo. He received the 7th Student Presentation Award from ASJ, the 35th Awaya Prize Young Researcher Award from ASJ, the 8th Outstanding Student Paper Award from IEEE Japan Chapter SPS, the Best Paper Award from APSIPA ASC 2014, the Student Paper Award from IEEE Kansai Section, the 30th TELECOM System Technology Award from TAF, the 2014 ISS Young Researcher's Award in Speech Field from the IEICE, the NAIST Best Student Award (Ph.D course), and the Best Student Award of Graduate School of Information Science (Ph.D course). His research interests include electroacoustics, signal processing, and speech synthesis. He is a student member of ASJ and IEEE SPS, and a member of ISCA.

Tomoki Toda earned his B.E. degree from Nagoya University, Aichi, Japan, in 1999 and his M.E. and D.E. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003, respectively. He is a Professor at the Information Technology Center, Nagoya University. He was a Research Fellow of JSPS from 2003 to 2005. He was then an Assistant Professor (2005-2011) and an Associate Professor (2011-2015) at the Graduate School of Information Science, NAIST. His research interests include statistical approaches to speech processing. He received more than 10 paper/achievement awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal).

Graham Neubig received his B.E. from University of Illinois, Urbana-Champaign in 2005, and his M.S. and Ph.D. in informatics from Kyoto University in 2010 and 2012 respectively. From 2012, he has been an assistant professor at the Nara Institute of Science and Technology, where he is pursuing research in machine translation and spoken language processing.

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR and U-STAR (2006-2011). In 2009-2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. From 2011, she has been an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015-2016, under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation". She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Satoshi Nakamura is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampoli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.