Non-native Speech Synthesis Preserving Speaker Individuality based on Partial Correction of Prosodic and Phonetic Characteristics

Yuji Oshima, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology, Japan

{shinnosuke-t, tomoki, neubig, ssakti, s-nakamura}@is.naist.jp

Abstract

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 HMM-based speech synthesis, which synthesizes foreign language speech of a specific non-native speaker reflecting the speaker-dependent acoustic characteristics extracted from the speaker's natural speech in his/her mother tongue, tends to cause a degradation of speaker individuality in synthetic speech compared to intra-lingual speech synthesis. This paper proposes a new approach to cross-lingual speech synthesis that preserves speaker individuality by explicitly 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 speech 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 prosodic correction method based on model adaptation, and (2) a phonetic correction method based on spectrum replacement for unvoiced consonants. The experimental results demonstrate that these proposed methods are capable of significantly improving naturalness while preserving the speaker individuality in synthetic speech.

Index Terms: cross-lingual speech synthesis, English-Read-by-Japanese (ERJ), speaker individuality, HMM-based speech synthesis, prosodic correction, phonetic correction

1. Introduction

Cross-lingual speech synthesis that synthesizes foreign language speech with a non-native speaker's own voice characteristics hold promise to improve a wide variety of applications. For example, it could allow Computer-Assisted Language Learning (CALL) applications to let learners listen to reference speech with their own voices [1], generate speech-to-speech translation output with the input speaker's voice [2], and allow for movie dubbing [3] preserving the original speakers' voice characteristics.

There have been many attempts at developing cross-lingual speech synthesis based on statistical voice conversion [4] or Hidden Markov Model (HMM)-based speech synthesis [5]. For example, one-to-many Gaussian Mixture Model (GMM)-based voice conversion can be applied to unsupervised speaker adaptation in cross-lingual speech synthesis [6, 7]. In addition, cross-lingual adaptation parameter mapping [8, 9] and cross-lingual frame mapping [10] 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 tends to be inferior to that of intra-lingual speech synthesis.

In this paper, we focus on an alternative paradigm where we explicitly collect utterances from the speaker in his/her *non-native tongue*. There is a small amount of previous work on this paradigm, for example [11], in which an HMM-based speech synthesizer was developed 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 [12, 13].

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 [14]) is highly different from its native English counterpart due to Japanese-accented prosody or pronunciation [15, 16]. 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 ERJ speech synthesis preserving speaker individuality based on the partial correction of prosodic and phonetic characteristics. The *prosody correction* method partly adapts the native English speaker's HMM parameters by using the target speaker's ERJ speech. The *phonetic correction* method partly replaces the generated spectral parameters of the ERJ speaker with those of the native English speaker, applying replacement to only unvoiced consonants, the acoustic characteristics of which are less affected by speaker differences. The experimental results demonstrate that the proposed methods are capable of improving naturalness and intelligibility of ERJ 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 [17]. The output probability distribution function of the c-th HMM state is given by:

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

where $\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta\Delta\boldsymbol{c}_t^\top\right]^\top$ is a feature vector including a static feature vector $\boldsymbol{c}_t$ and its dynamic feature vectors $\Delta\boldsymbol{c}_t$ and $\Delta\Delta\boldsymbol{c}_t$. $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the mean vector and the covariance matrix of Gaussian distribution $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right)$ of the c-th HMM-state, respectively.

Model adaptation for HMM-based speech synthesis [18]

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:

$$\hat{\boldsymbol{\mu}}_c = \boldsymbol{A}\boldsymbol{\mu}_c + \boldsymbol{b}, \quad (2)$$

$$\hat{\boldsymbol{\Sigma}}_c = \boldsymbol{A}\boldsymbol{\Sigma}_c\boldsymbol{A}^\top, \quad (3)$$

where the transformation matrix $\boldsymbol{A}$ and the bias vector $\boldsymbol{b}$ are adaptation parameters. 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 [19].

3. Proposed Partial Correction of Prosodic and Phonetic Characteristics

This section describes our proposed method for synthesizing more naturally sounding ERJ speech while preserving speaker individuality. A subset of the native English speaker's HMM parameters are used to improve the naturalness of synthetic speech from the ERJ speaker's HMMs.

3.1. Prosody Correction based on Model Adaptation

The ERJ speaker's HMMs are created by adapting the native speaker's pre-trained HMMs to the ERJ speech data. However, the standard adaptation process transforming all HMM parameters makes synthetic speech from the adapted HMMs sound as unnatural as the original ERJ speech. It is well known that large differences between ERJ speech and native English speech are often observed in duration and power [20, 21]. 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.

Figure 1 illustrates an overview of the proposed prosody correction method. 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 ERJ speech data in the standard manner [18], and duration and power are kept unchanged. Consequently, the adapted HMMs model the spectral envelope and excitation parameters of the target ERJ speech and duration and power patterns of the native English 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 ERJ speaker with those of the native English speaker. Although there are many studies in speech perception [22, 23] 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 ERJ speech. Therefore, we

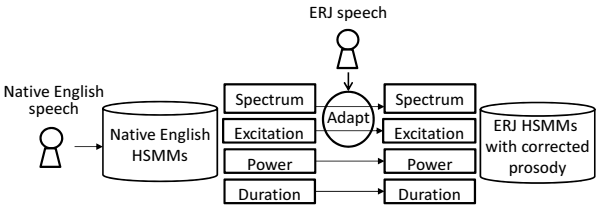

Figure 1: An overview of the proposed prosody correction method based on model adaptation techniques.

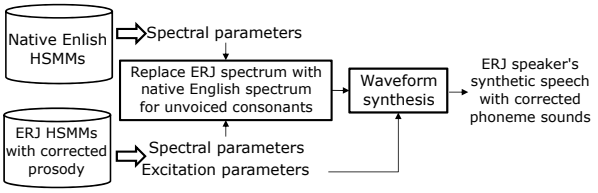

Figure 2: An overview of the proposed phonetic correction method based on the spectrum replacement for unvoiced consonants.

can expect that replacing the spectrum of unvoiced consonants with their native counterparts may improve naturalness without causing adverse effects on speaker individuality.

Figure 2 illustrates an overview of the proposed phonetic correction method. First, we generate two kinds of synthetic speech parameters from the native speaker's HMMs and the ERJ 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 ERJ 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 [10, 24, 25], we found that spectral replacement caused no significant degradation, and thus for simplicity we use it 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 [26] 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 [27] were extracted as excitation parameters by STRAIGHT [28, 29]. The feature vector consists of spectral and excitation parameters and their delta and delta-delta features. 5-state left-to-right HSMMs [30] were used. The log-scaled power and the mel-cepstral coefficients were trained in the same stream. CSMAPLR + MAP [31] were used for model adaptation, and the block diagonal matrix corresponding to static parameters and their delta and delta-delta features was used as the adaptation matrix. Intra-gender adaptation was performed in adaptation from the native speakers to several non-native speakers.

We evaluate synthetic speech of the following systems:

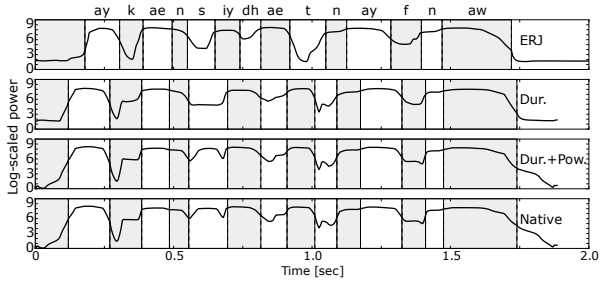

Figure 3: An example of the power trajectories of synthesized English speech samples for a sentence "I can see that knife now."

ERJ: speaker-dependent HSMMs trained using ERJ speech

HMM+VC: A GMM that converted the parameters generated from "Native" to the ERJ speech parameters [7][1]

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

Native: speaker-dependent HSMMs trained using native 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.

4.2. Evaluation of Prosodic Correction

4.2.1. Naturalness and Speaker Individuality

In this section, we evaluate the effect of the proposed prosody correction 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 tet 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.

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." Figure 4 shows the results of subjective evaluation. Although the conventional VC-based system ("HMM+VC") improves the naturalness compared to "ERJ" and the fully adapted HMMs ("Adapt"), its scores on speaker individuality decrease significantly. On the other hand, the proposed methods "Dur." and "Dur.+Pow." yield better scores on naturalness than "ERJ" and "Adapt" while maintaining the scores on speaker individuality.

[1]We adopt the one-to-one GMM-based conversion framework instead of the one-to-many framework [7].

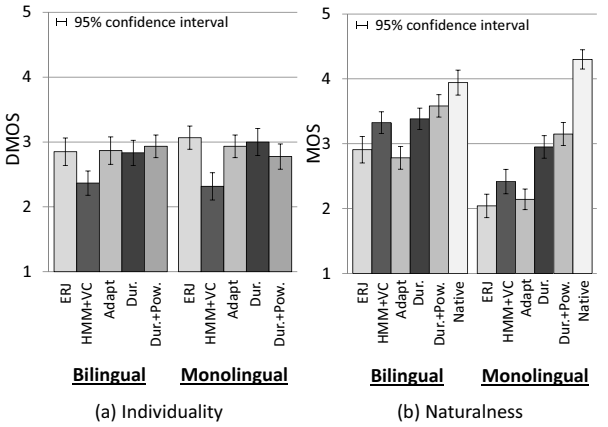

Figure 4: Results of subjective evaluation using the proposed prosody correction method.

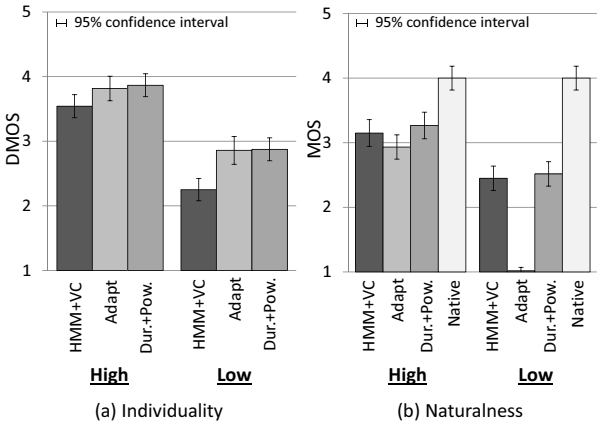

Figure 5: Results of subjective evaluation using the proposed prosody correction method with ERJ speakers that have various English proficiency levels.

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 [32] sentences from the ERJ database [14] uttered by 2 male and 2 female speakers who had the best ("High") or the worst ("Low") English proficiecy level, based on evaluation from various perspectives (i.e., rhythm and accent) in [14][2] . 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 5 shows the result of the subjective evaluations. The results are calculated for each proficiency level. In term 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 effec-

[2]Multiple scores assigned to each speaker [14] were averaged to determine the best and worst English proficiency levels.

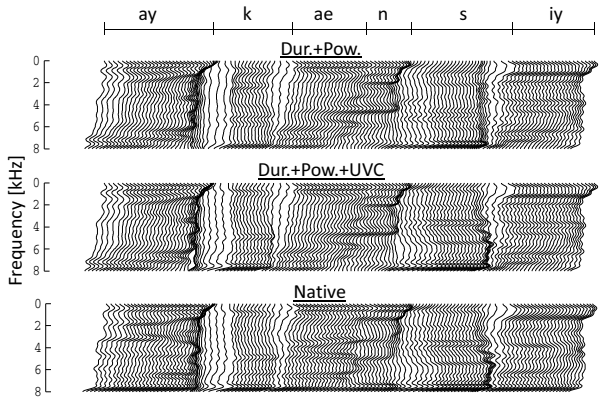

Figure 6: An example of the spectrograms of the synthesized English speech samples.

tiveness of the proposed prosody correction method over various proficiency levels.

4.3. Evaluation of Phonetic Correction Method

Next, we evaluate 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 use 60 CMU ARCTIC sentences uttered by "Monolingual" and "Bilingual" from Section 4.2.1, and 60 TIMIT sentences uttered by 4 speakers from Section 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 6 illustrates 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 7 shows the results 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.[3] 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 Section 4.3. Synthetic speech samples of "HMM+VC," "Dur.+Pow.+UVC," and "Native" were presented to listeners in random order. The word correct rate and word accuracy were calculated for each proficiency level.

Figure 8 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

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

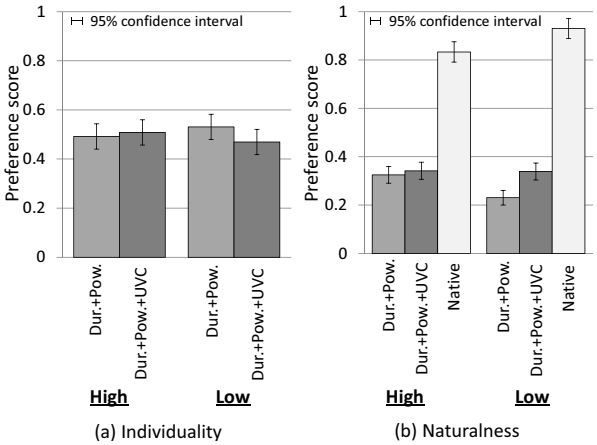

Figure 7: Results of the subjective evaluations using the proposed phonetic correction method.

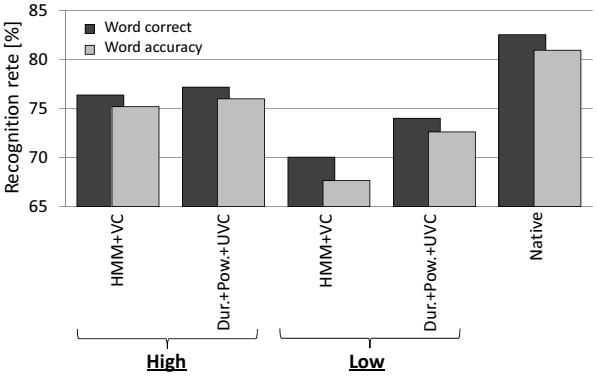

Figure 8: Results of the dictation test for intelligibility.

(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. Summary

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 the proposed methods are capable of improving naturalness while preserving speaker individuality in synthetic speech.

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

6. References

[1] A. W. Black, "Speech synthesis for educational technology," in *Proc. SLaTE*, Farmington, PA, USA, Oct. 2007, pp. 104–107.

[2] J. Dines, H. Liang, L. Saheer, M. Gibson, W. Byrne, K. Oura, K. Tokuda, J. Yamagishi, S. King, M. Wester *et al.*, "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.

[3] O. Turk, O. Buyuk, A. Haznedaroglu, and L. M. Arslan, "Application of voice conversion for cross-language rap singing transformation," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3597–3600.

[4] 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.

[5] 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, Apr. 2013.

[6] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, and T. Dutoit, "Cross-language voice conversion based on eigenvoices," in *Proc. INTERSPEECH*, Brighton, UK, Sep. 2009, pp. 1635–1638.

[7] 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," in *Proc. INTERSPEECH*, Aug. 2011, pp. 2769–2772.

[8] H. Liang, Y. Qian, F. K. Soong, and G. Liu, "A cross-language state mapping approach to bilingual (Mandarin-English) TTS," in *Proc. ICASSP*, Mar. 2008, pp. 4641–4644.

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

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

[11] 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.

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

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

[14] 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," in *Proc. ICA*, vol. 1, 2004, pp. 557–560.

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

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

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

[18] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. and Syst*, vol. 90, no. 2, pp. 533–543, Feb. 2007.

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

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

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

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

[23] ——, "Speaker individualities in speech spectral envelopes and fundamental frequency contours," in *Speaker Classification II*. Springer, 2007, pp. 157–176.

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

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

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

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

[28] 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, Apr. 1999.

[29] 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*, Sep. 2001, pp. 59–64.

[30] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis system," *IEICE Trans., Inf. and Syst., E90-D*, vol. 90, no. 5, pp. 825–834, 2007.

[31] 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, Jun. 2009.

[32] 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.