CROSS-LINGUAL VOICE CONVERSION WITH BILINGUAL PHONETIC POSTERIORGRAM AND AVERAGE MODELING

Yi Zhou[1], *Xiaohai Tian*[1], *Haihua Xu*[2], *Rohan Kumar Das*[1] *and Haizhou Li*[1]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Temasek Laboratories, Nanyang Technological University, Singapore

ABSTRACT

This paper presents a cross-lingual voice conversion approach using bilingual Phonetic PosteriorGram (PPG) and average modeling. The proposed approach makes use of bilingual PPGs to represent speaker-independent features of speech signals from different languages in the same feature space. In particular, a bilingual PPG is formed by stacking two monolingual PPG vectors, which are extracted from two monolingual speech recognition systems. The conversion model is trained to learn the relationship between bilingual PPGs and the corresponding acoustic features. To leverage the linguistic and acoustic information from other speakers in different languages, an average model is trained with multiple speakers in both source and target languages. I-vector is utilized as an additional input feature of the average model for network adaptation. Experiments are performed for intralingual and cross-lingual voice conversion between English and Mandarin speakers. Both objective and subjective evaluations demonstrate the effectiveness of our proposed approach.

Index Terms— cross-lingual, voice conversion, Phonetic PosteriorGram (PPG), average modeling approach (AMA)

1. INTRODUCTION

Voice conversion (VC) aims to modify the speech of one speaker (source) to make it sound as if it were spoken by another speaker (target). Most of existing VC techniques are designed for intralingual conversions given parallel training data, where the source and target speakers speak the same text in the same language. A number of models have been established to realize the spectral feature mapping from the source to target, such as Gaussian mixture models [1, 2], neural network based methods [3–8], frequency warping methods [9–12], exemplar based methods [13–16] and so on.

In cross-lingual voice conversion, the phonetic systems of source and target languages are different, and parallel training data is not possible. Therefore, cross-lingual voice conversion is a more challenging task than intralingual conversion. In [17, 18], a bilingual conversion model is trained on parallel data of the target language, which requires the source speaker to speak both source and target languages. However, in practice, a bilingual source speaker is not always available. Non-parallel alignment techniques are hence developed to find source-target frame pairs from non-parallel utterances, for instance, unit selection [19, 20] and the iterative frame alignment methods [21,22]. But the conversion performances are moderate due to their inaccurate alignments [22]. Alternatively, vocal tract length normalization based phone mapping approaches are developed [23, 24], where the warping functions are estimated between the closest phone or acoustic classes between the source and target speech. Recently, a Phonetic PosteriorGram (PPG) based cross-lingual VC technique [25] has been reported, which makes use of monolingual PPGs as speaker-independent features to bridge across speakers and language boundaries. Nevertheless, PPGs of one language cannot effectively characterize the phonetic contents of another language owing to the fact that different languages have distinct phone sets.

In this paper, we propose to use bilingual PPGs for cross-lingual VC with an average modeling approach (AMA). Bilingual PPGs are formed by stacking two monolingual PPG vectors, which are extracted from two automatic speech recognition (ASR) systems trained in source and target languages, respectively. Then the conversion model is trained to map bilingual PPGs to the acoustic features. As the target speech only contains monolingual information, it is not able to fully describe the linguistic and acoustic information of another language. To address this problem, a cross-lingual AMA is employed to capture both linguistic and acoustic information from different languages. In this way, the proposed method is expected to perform cross-lingual VCs in both directions between the source and target languages.

2. CROSS-LINGUAL VOICE CONVERSION WITH MONOLINGUAL PPG

Phonetic PosteriorGram (PPG) is a time-versus-class vector that represents the posterior probabilities of phonetic classes for a specific time frame [25]. PPG is estimated from an ASR system, which is trained by a large multi-speaker database. As the output of an ASR system is designed to be invariant with different speakers, the extracted PPGs are considered as speaker independent [26]. Small speech seg-

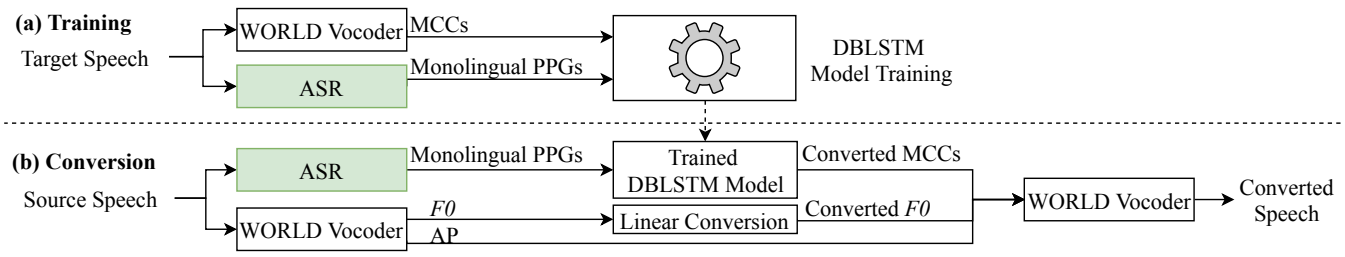

Fig. 1. Block diagram of (a) training and (b) conversion workflows of the cross-lingual VC system with monolingual PPGs.

ments like frames can be shared in different languages [24], hence frame-level PPGs are also believed to be language independent. Due to their speaker-independent and language-independent properties, PPGs have been successfully applied for cross-lingual [25] voice conversion.

2.1. Methodology

Fig. 1(a) presents the framework of monolingual PPG based cross-lingual VC. During training, monolingual PPGs $\mathbf{X} \in \mathbb{R}^{D_m \times N}$ and their corresponding Mel Cepstral Coefficients (MCCs) $\mathbf{Y} \in \mathbb{R}^{D_a \times N}$ are first extracted for the target speaker. N denotes the number of frame, D_a and D_m denote the dimensions of acoustic features and monolingual PPGs, respectively. Then, the conversion model $F(\cdot)$ is trained between monolingual PPGs and MCCs as

$$\mathbf{Y} = F(\mathbf{X}) + \mathbf{e} \tag{1}$$

where $\mathbf{e}$ is the error between predicted MCCs and the ground truth. In particular, the ASR adopted for PPG extraction is trained in the same language with the target speech.

Fig. 1(b) shows the conversion process. Given a source speech in a different language, we first use the same ASR to extract monolingual PPGs, denoted as $\mathbf{X}' \in \mathbb{R}^{D_m \times N}$. Then, the extracted PPGs are used as input to the conversion model to predict the converted MCCs $\hat{\mathbf{Y}} \in \mathbb{R}^{D_a \times N}$ as

$$\hat{\mathbf{Y}} = F(\mathbf{X}'). \tag{2}$$

The converted speech is then reconstructed with converted fundamental frequency ($F0$), source aperiodic component (AP) and the generated MCCs.

2.2. Limitation

Although monolingual PPG works properly for cross-lingual VC, the phonetic information represented by it of one language is inaccurate for another language. As a result, the converted voice is unnatural and biased to one language. Moreover, it is noted that the monolingual PPG based VC system is mainly designed for one-direction conversions from the source language to the target language [25], but not suitable for conversions in both directions.

3. CROSS-LINGUAL VC WITH BILINGUAL PPG AND AVERAGE MODELING

In this section, we introduce bilingual PPG and average modeling as a solution to cross-lingual VC.

3.1. Cross-lingual VC with bilingual PPG

To capture accurate phonetic information of both source and target languages, bilingual PPGs are introduced for cross-lingual VC. During training, as shown in Fig. 2(a), monolingual PPGs, $\mathbf{X}_{en} \in \mathbb{R}^{D_{en} \times N}$, $\mathbf{X}_{cn} \in \mathbb{R}^{D_{cn} \times N}$, are extracted by two ASR systems (English and Mandarin). The bilingual PPG is then formed by stacking the monolingual PPGs, denoted as $\mathbf{X} = [\mathbf{X}_{en}{}^{\mathrm{T}}, \mathbf{X}_{cn}{}^{\mathrm{T}}]^{\mathrm{T}}$. We have D_{en} and D_{cn} to indicate the dimensions of English and Mandarin PPGs. Then, a conversion model is trained to learn the feature mapping between bilingual PPGs and MCCs as described in Section 2.

At run-time, as shown in Fig. 2(b), we first extract bilingual PPGs from the source speech and then transform the extracted bilingual PPGs to MCCs according to Eq. (2).

3.2. AMA with bilingual PPG

To leverage both linguistic and acoustic information of different languages (e.g. English and Mandarin), an average model trained with multiple speakers from both languages is employed. We present the i-vector as part of the input features by augmenting it to the PPG features for average model adaptation. For speaker k, the input features are represented as $[\mathbf{X}_{en,k}{}^{\mathrm{T}}, \mathbf{X}_{cn,k}{}^{\mathrm{T}}, \mathbf{I}_k]^{\mathrm{T}}$ with $\mathbf{I}_k$ denoting the i-vector. The training and conversion details can be found in [27].

4. EXPERIMENT

4.1. Database and Feature Extraction

We chose 10 English speakers (5 male and 5 female) from VCC2016 database [28] and 10 Mandarin speakers (5 male and 5 female) from a Mandarin average model database [1] for average model training. Each speaker consisted of 162

[1] http://www.data-baker.com/hc_pm_en.html

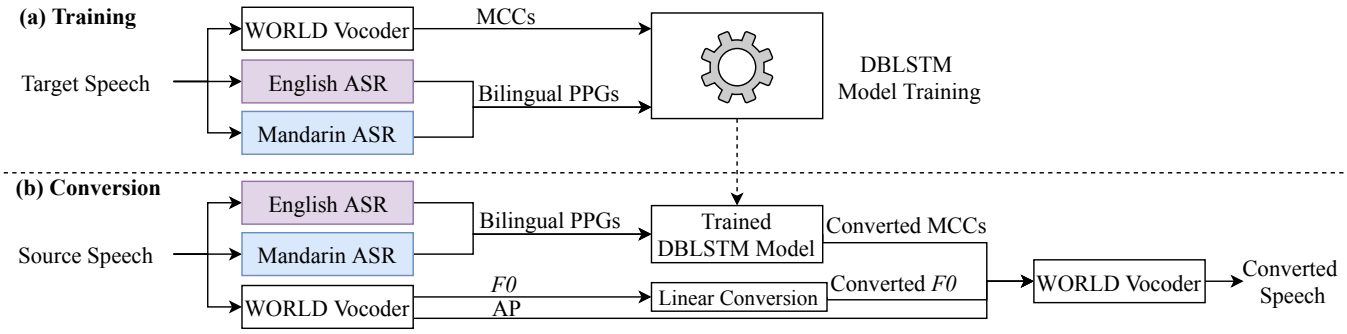

Fig. 2. Block diagram of (a) training workflow and (b) conversion workflow of the proposed system with bilingual PPGs.

utterances. For testing, 2 English speakers *TF1, TM1* from VCC2018 database [29] and 2 Mandarin speakers *14M, 16F* from the same Mandarin database were used, with 20 sentences from each speaker. All of the selected speakers are native and monolingual.

Both DNN-HMM ASR models were trained using the Kaldi toolkit [30]. The English and Mandarin ASR systems were trained on Wall Street Journal (WSJ) [31] and Aishell [32] corpus, respectively. The English ASR consisted of 5 hidden layers of 1024 units in each layer, and a soft-max output layer of 132 units, while the Mandarin ASR consisted of 6 hidden layers with 2048 units in each layer, and a soft-max output layer of 209 units.

With all speech signals sampled at 16kHz, the WORLD vocoder [33] was used to extract the spectrum (513-dim), AP (1-dim), and $F0$ (1-dim), after which we used the Speech Signal Processing Toolkit [2] to compute the 40-dimensional MCCs. The i-vector dimension was fixed as 150 after applying linear discriminant analysis.

4.2. Experiment Setup

Three different systems were implemented for comparison:

- **M-PPG:** the cross-lingual VC system with monolingual PPG as described in Section 2.1, and we benchmark it as our baseline. Two English and two Mandarin conversion models were trained for the target speakers *SF1*, *SM1*, *01F* and *07M*. For each system, 150 and 12 utterances were used for training and validation. The input features of English and Mandarin systems had 132 and 209 dimensions, respectively.

- **B-PPG:** the proposed bilingual PPG VC system as in Section 3.1. Similarly, two English models and two Mandarin models were trained with bilingual PPG on the same data. The input features had 341 dimensions.

- **B-PPG-AMA:** the proposed VC system with bilingual PPG and average modeling. Two gender-dependent average models were trained, and each model had 10 speakers (5 English and 5 Mandarin). In total, one average model used 1500 sentences for training and 120 sentences for validation. Each speaker's i-vector was obtained from his own 150 training speech samples as described in Section 3.2. Both average models were trained with 491-dimensional input feature vectors.

All models were trained with Merlin toolkit [34] using the same settings: two DBLSTM layers of 256 hidden units in each layer, 25 minibatch size, 0.9 momentum and 0.002 learning rate. The network had a common output of 127-dimensional features including MCCs (40-dim), log $F0$ (1-dim), AP (1-dim) with their delta and delta-delta coefficients, and the voiced/unvoiced flag (1-dim).

During conversion, APs were directly copied from source speech, while $F0$ was converted by a global linear transformation in log-scale [35]. The MCCs were generated by Maximum Likelihood Parameter Generation algorithm [36]. A post-filtering in the cepstral domain was employed to further enhance the speech quality.

4.3. Evaluations

We conducted both objective and subjective evaluations. Since the reference speeches were not available for cross-lingual VC, we only report the intralingual VC results for objective evaluation. The subjective tests were conducted to evaluate both intralingual and cross-lingual VCs.

4.3.1. Objective Evaluation

The Mel-Cepstral Distortion (MCD) was used as an objective measure of the spectral distance between the converted and target speeches. It can be calculated by the equation:

$$MCD[dB] = 10/\ln 10 \sqrt{2 \sum_{d=1}^{D_a} (\hat{Y}_d - Y_d)^2}, \qquad (3)$$

where D_a is the dimension of MCCs, $\hat{Y}_d$ and Y_d are the d^{th} coefficients of the corresponding converted and target MCCs, and the lower value accounts for a smaller distortion.

[2] https://sourceforge.net/projects/sp-tk/

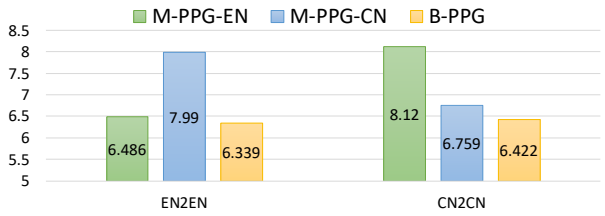

Fig. 3. Average MCD results on intralingual VCs. M-PPG-EN and M-PPG-CN indicate the conversion models are trained using English and Mandarin monolingual PPGs, respectively. EN2EN and CN2CN refer to the intralingual conversions of English and Mandarin, respectively.

The objective results are presented in Fig. 3. It is observed that monolingual PPG is only effective on the conversion of its corresponding language. For example, in EN2EN conversion, M-PPG-EN outperforms M-PPG-CN with the MCDs of 6.486 and 7.99, respectively. The results also suggest that our proposed B-PPG outperforms M-PPG-EN and M-PPG-CN in both EN2EN and CN2CN conversions, which demonstrate the proposed B-PPG can capture more detailed phonetic characterization information for intralingual VC.

4.3.2. Subjective Evaluation

The Mean Opinion Score (MOS) and ABX preference test were conducted for subjective evaluations on both intralingual and cross-lingual VCs. 24 listeners, including 12 Mandarin and 12 English, participated all the tests. For each test, 12 sentences were randomly selected from each system.

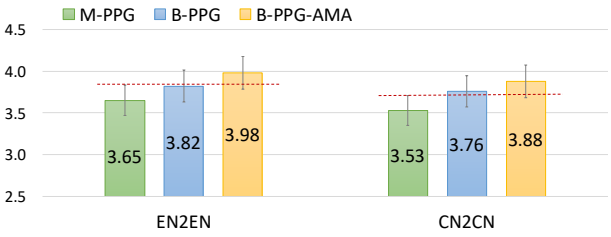

Fig. 4. Intralingual MOS test results with 95% confidence intervals. EN2EN and CN2CN refer to intralingual VC of English and Mandarin, respectively.

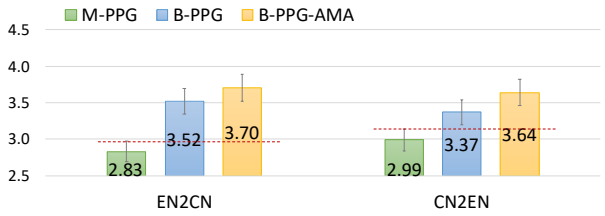

Fig. 5. Cross-lingual MOS test results with 95% confidence intervals. EN2CN and CN2EN denote English to Mandarin conversion and Mandarin to English conversion.

In the MOS test, listeners were asked to rate the quality and naturalness of the converted speech on a 5-point scale.

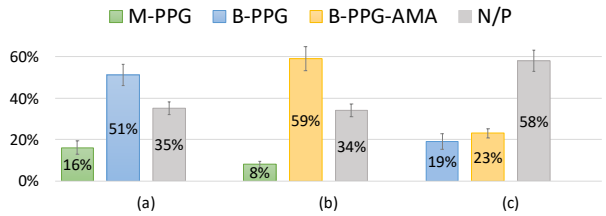

Fig. 6. ABX preference test results for speaker similarity with 95% confidence intervals, N/P stands for no preference. (a) M-PPG vs. B-PPG; (b) M-PPG vs. B-PPG-AMA; (c) B-PPG vs. B-PPG-AMA.

Fig. 4 shows the MOS results for intralingual VC. We observe that the proposed B-PPG and B-PPG-AMA outperform the M-PPG baseline, but the difference is not statistically significant. While in Fig. 5, our proposed B-PPG and B-PPG-AMA significantly outperform M-PPG for cross-lingual conversions. The results indicate VCs, especially cross-lingual VCs, clearly benefit from bilingual PPG, as they can effectively capture the phonetic classes in two languages.

In ABX preference test, X was the reference target speech. Listeners were asked to choose which one was more similar to X given converted samples A and B from different systems.

Shown in Fig. 6(a) and Fig. 6(b), both B-PPG and B-PPG-AMA significantly outperform M-PPG in terms of speaker similarity. Then we compare our proposed B-PPG with B-PPG-AMA. The results are shown in Fig. 6(c), which suggest that the converted speech of B-PPG-AMA achieves better performances than that of B-PPG regarding to the speaker similarity, though the difference is not statically significant.

Both quality and similarity tests demonstrate the effectiveness of proposed bilingual PPG and average modeling for intralingual and cross-lingual voice conversions. The converted samples can be found on this website [3].

5. CONCLUSION

This paper presents the cross-lingual voice conversion techniques based on bilingual PPG and average modeling. By using bilingual PPGs to represent the phonetic contents, input linguistic information performs more robust across different languages. Additionally, the average modeling approach further enhances the cross-lingual conversion performances. Experimental results confirm that our proposed methods outperform the baseline system in terms of both speech quality and speaker similarity. Last, our proposed methods are suitable for conversions in either direction of two languages.

6. ACKNOWLEDGEMENT

This research is supported by the NUS Start-up Grant FY2016 'Non-parametric approach to voice morphing'. Yi Zhou is also funded by NUS research scholarship. The Mandarin average model database is provided by Data-baker.

[3] http://yizhou012.github.io/vc

7. REFERENCES

[1] Yannis Stylianou, Olivier Cappé, and Eric Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[2] Tomoki Toda, Alan W Black, and Keiichi 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.

[3] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[4] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *IEEE ICASSP*, 2015, pp. 4869–4873.

[5] Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura, "Adaptive wavenet vocoder for residual compensation in gan-based voice conversion," in *IEEE SLT*, 2018.

[6] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *INTERSPEECH*, 2017, pp. 3364–3368.

[7] Jaime Lorenzo-Trueba, Fuming Fang, Xin Wang, Isao Echizen, Junichi Yamagishi, and Tomi Kinnunen, "Can we steal your vocal identity from the internet?: Initial investigation of cloning obama's voice using gan, wavenet and low-quality found data," *arXiv:1803.00860*, 2018.

[8] Mingyang Zhang, Berrak Sisman, Sai Sirisha Rallabandi, Haizhou Li, and Li Zhao, "Error reduction network for dblstm-based voice conversion," in *arXiv:1809.09841*, 2018.

[9] Daniel Erro, Asunción Moreno, and Antonio Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[10] Elizabeth Godoy, Olivier Rosec, and Thierry Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[11] Xiaohai Tian, Zhizheng Wu, Siu Wa Lee, and Eng Siong Chng, "Correlation-based frequency warping for voice conversion," in *ISCSLP*, 2014, pp. 211–215.

[12] Xiaohai Tian, Zhizheng Wu, Siu Wa Lee, Quy Hy Nguyen, Minghui Dong, and Eng Siong Chng, "System fusion for high-performance voice conversion," in *INTERSPEECH*, 2015, pp. 2759–2763.

[13] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based voice conversion in noisy environment," in *IEEE SLT*, 2012, pp. 313–317.

[14] Xiaohai Tian, Siu Wa Lee, Zhizheng Wu, Eng Siong Chng, Haizhou Li, Xiaohai Tian, Siu Wa Lee, Zhizheng Wu, Eng Siong Chng, and Haizhou Li, "An exemplar-based approach to frequency warping for voice conversion," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 25, no. 10, pp. 1863–1876, 2017.

[15] Berrak Sisman, Haizhou Li, and Kay Chen Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," in *IEEE ASRU*, 2017, pp. 677–684.

[16] Berrak Sisman, Mingyang Zhang, and Haizhou Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder," in *INTERSPEECH*, 2018, pp. 1978–1982.

[17] Masanobu Abe, Kiyohiro Shikano, and Hisao Kuwabara, "Statistical analysis of bilingual speaker's speech for cross-language voice conversion," *The Journal of the Acoustical Society of America*, vol. 90, no. 1, pp. 76–82, 1991.

[18] Mikiko Mashimo, Tomoki Toda, Hiromichi Kawanami, Kiyohiro Shikano, and Nick Campbell, "Cross-language voice conversion evaluation using bilingual databases," *IPSJ Journal*, vol. 43, no. 7, pp. 2177–2185, 2002.

[19] David Sundermann, Harald Hoge, Antonio Bonafonte, Hermann Ney, Alan Black, and Shri Narayanan, "Text-independent voice conversion based on unit selection," in *IEEE ICASSP*, 2006, vol. 1, pp. I–81–I–84.

[20] Hao Wang, Frank Soong, and Helen Meng, "A spectral space warping approach to cross-lingual voice transformation in hmm-based tts," in *IEEE ICASSP*, 2015, pp. 4874–4878.

[21] Daniel Erro and Asunción Moreno, "Frame alignment method for cross-lingual voice conversion," in *INTERSPEECH*, 2007, pp. 1969–1972.

[22] Daniel Erro, Asunción Moreno, and Antonio Bonafonte, "Inca algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[23] David Sundermann, Hermann Ney, and H Hoge, "Vtln-based cross-language voice conversion," in *IEEE ASRU*, 2003, pp. 676–681.

[24] Yao Qian, Ji Xu, and Frank K Soong, "A frame mapping based hmm approach to cross-lingual voice transformation," in *IEEE ICASSP*, 2011, pp. 5120–5123.

[25] Lifa Sun, Hao Wang, Shiyin Kang, Kun Li, and Helen M Meng, "Personalized, cross-lingual tts using phonetic posteriorgrams," in *INTERSPEECH*, 2016, pp. 322–326.

[26] Zhao Guanlong, Sonsaat Sinem, Levis John, Chukharev-Hudilainen Evgeny, and Gutierrez-Osuna Ricardo, "Accent conversion using phonetic posteriorgrams," in *IEEE ICASSP*, 2018, pp. 5314–5318.

[27] Xiaohai Tian, Junchao Wang, Haihua Xu, Eng-Siong Chng, and Haizhou Li, "Average modeling approach to voice conversion with non-parallel data," in *Proceedings of Odyssey The Speaker and Language Recognition Workshop*, 2018, pp. 227–232.

[28] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "The voice conversion challenge 2016," in *INTERSPEECH*, 2016, pp. 1632–1636.

[29] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel method," *arXiv:1804.04262*, 2018.

[30] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE ASRU*, 2011, number EPFL-CONF-192584.

[31] Douglas B Paul and Janet M Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[32] Bu Hui, Du Jiayu, Na Xingyu, Wu Bengu, and Zheng Hao, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Oriental COCOSDA*, 2017, pp. 1–5.

[33] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[34] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An open source neural network speech synthesis system," *Proceedings of Speech Synthesis Workshop (SSW), Sunnyvale, USA*, 2016.

[35] Berrak Şişman, Haizhou Li, and Kay Chen Tan, "Transformation of prosody in voice conversion," in *APSIPA ASC*, 2017, pp. 1537–1546.

[36] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *IEEE ICASSP*, 2000, vol. 3, pp. 1315–1318.