Many-to-many Cross-lingual Voice Conversion using Bilingual Phonetic PosteriorGram by an Average Modeling Approach

Yi Zhou

Department of Electrical and Computer Engineering
National University of Singapore, Singapore
yi.zhou@u.nus.edu

Abstract

In cross-lingual voice conversion (VC), the source and target speakers speak in different languages making it impossible to obtain the same utterance as paired training data. Phonetic PosteriorGram (PPG) is an effective solution to address this problem, which can bridge between multiple speakers and languages. PPG is the posterior probabilities of the phonetic classes in an utterance, which can be obtained by a speaker independent automatic speech recognition (ASR) system. The model can be trained to map PPG to the acoustic features. In our work, we proposed bilingual PPG as a more effective phonetic characterization of two languages. To further enhance the conversion performance of bilingual PPG-based cross-lingual VC framework, we also propose an average modeling approach to leverage the linguistic and acoustic information from other speakers in different languages. However, the conversion performance is highly dependent on the quality of ASR systems. There still exist many research problems to be addressed to achieve a better cross-lingual VC result.

Index Terms: cross-lingual, voice conversion, Phonetic PosteriorGram (PPG), AMA

1. Introduction

Voice conversion (VC) is a technique to modify the speech of one speaker (source) to make it sound like that of another speaker (target) while preserving the linguistic information [1]. According to whether the source speaker and target speaker speak the same language, VC can be broadly divided into intralingual VC and cross-lingual VC. In intralingual VC, most existing methods rely on the parallel data during training, where the source and target speakers need to record the same utterances [2]. Figure 1 shows a typical framework for intralingual VC using parallel data. The source and target utterances are first aligned to form source-target pairs for model training. Then the trained model can generate the converted target speech given an utterance from the source speaker [3]. However, to achieve a reasonable performance, it usually needs a relatively large database for training, which is not practical in real-life applications [4].

In cross-lingual VC, the source and target speakers speak different languages [5, 6], hence, parallel data is not available. To achieve cross-lingual conversions, we need to address the non-parallel training data problem. First, various alignment methods have been proposed to find the optimal source-target frame pairs from speech of different languages in the training database. For example, unit selection [7, 8], iterative frame alignment methods [9, 10, 11], and VTLN-based mapping approaches [12] are widely developed. However, the quality of converted speech is highly dependent on the alignment performance. Second, we may consider non-parallel VC techiqnues. Several recent approaches like Variational Autoencoder [13, 14] and Adversarial Generative Networks (GAN) [15, 16, 17] can be considered as possible solutions for cross-lingual VC. While, there is still a big gap between the converted speech and the natural one even in intralingual VC [15]. Last, we can create parallel data from non-parallel data. Phonetic PosteriorGram (PPG) is one possible solution, which represents the linguistic information of speech data. Since PPG and acoustic features are extracted from the same utterance, they are born parallel [18]. PPG-based cross-lingual VC between English and Mandarin speakers has been reported in [19], which utilizes an English ASR system to obtain monolingual PPG in English. During training, English PPG is trained to be mapped to the mel cepstral coefficients (MCCs). During testing, given a source Mandarin utterance, we use the same English ASR system to extract the English PPG. Then we pass PPG to the trained model to obtain the converted MCCs.

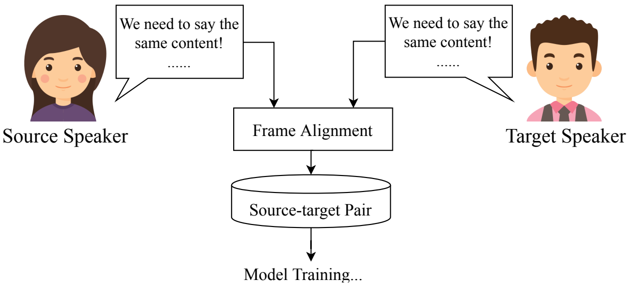

We need to say the same content! ......
We need to say the same content! ......
Source Speaker
Target Speaker
Frame Alignment
Source-target Pair
Model Training...

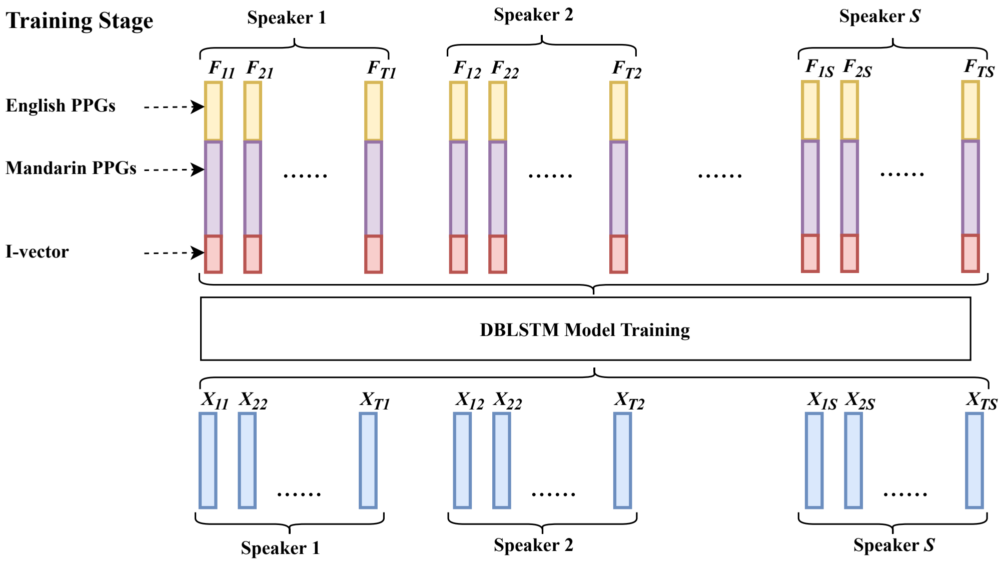
Figure 1: *The intralingual VC framework using parallel data.*

2. Limitations

In the above PPG-based cross-lingual VC framework, we can easily find that there is a language mismatch in the PPG representations during conversion. Since different language usually have distinct phonetic classes, using English PPG to represent a Mandarin utterance may result in performance degradation in the converted voice. At the same time, using speech data in one language for model training is not able to fully describe the linguistic and acoustic information of another language.

3. Contributions

We have proposed a cross-lingual with bilingual PPG by an average modeling approach (AMA) [20]. First, bilingual PPG is formed by concatenating English and Mandarin PPG by using two ASR systems in each language. Second, speech data from multiple speakers in both languages are used during training by an AMA. Speaker i-vector is used as a condition to generate



Training Stage
Speaker 1
Speaker 2
Speaker S
F_{11} F_{21} F_{T1}
F_{12} F_{22} F_{T2}
F_{1S} F_{2S} F_{TS}
English PPGs
Mandarin PPGs
I-vector
……
DBLSTM Model Training
X_{11} X_{22} X_{T1}
X_{12} X_{22} X_{T2}
X_{1S} X_{2S} X_{TS}
Speaker 1
Speaker 2
Speaker S

Figure 2: *The proposed bilingual PPG-based cross-lingual VC system by an average modeling approach.*

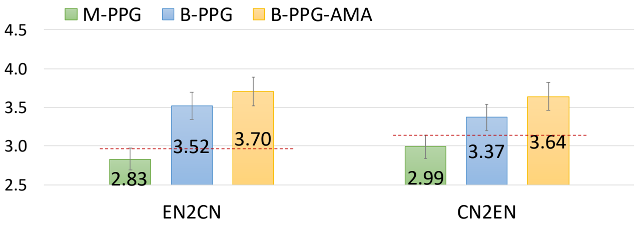


Figure 3: *Speech quality test result, B-PPG-AMA is our proposed framework, a higher score accounts for a better quality.*

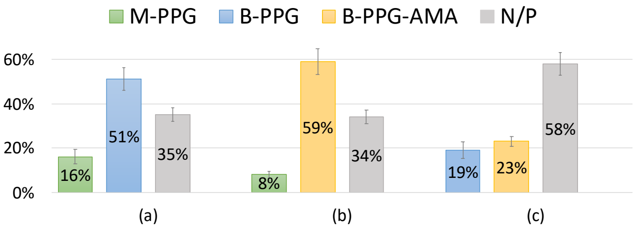


Figure 4: *Speaker similarity test result, B-PPG-AMA is the proposed framework, a higher value indicates a higher similarity.*

the desired target voice. The proposed approach is shown in Figure 2. During training, instead of using monolingual PPG, we extract English PPG, Mandarin PPG and combine them to be bilingual PPG. Then i-vector is also augmented to bilingual PPG to form the input features for model training. During conversion, we first get the bilingual PPG from a source utterance. Then we extract the speaker i-vector from the target speech. Similarly, we can augment i-vector to bilingual PPG and feed them into the trained model for MCC generation.

By doing so, our proposed framework can realize many-to-many cross-lingual VC benefiting from other speaker's large database in both languages. The speech quality and speaker similarity test results are presented in Figure 3 and Figure 4, respectively. We only give a brief discussion here, B-PPG-AMA is our proposed bilingual PPG based system using the average modeling approach. Both results show our proposed approach achieves the best performance result, which confirms our contributions in cross-lingual VC.

4. Future Works

There are four main aspects we can consider for future improvement in cross-lingual VC.

1. Currently, we are working on mixed-lingual PPG to replace bilingual PPG, which can be obtained by a unified English-Mandarin acoustic model. We aim to improve the linguistic representation by customizing the acoustic model in the ASR system for VC task.

2. We also try to improve the speaker embedding for average model adaptation in another submitted work. Instead of using speaker i-vector, we augment an auxiliary speaker embedding network to the primary VC network for joint training. As the speaker embedding is optimized for VC task, the overall performance is optimal.

3. Deep neural network based approaches have been widely studied in speech synthesis area [21, 22] in the last decade. Several machine learning techniques can be considered to improve the network of our current system such as attention [23] and multi-task learning [24].

4. We will study the effect of integrating PPG into the recent advanced speech synthesis tools like WaveNet [25] and Tacotron [26] for performance improvement. We have obtained some preliminary results.

5. Acknowledgements

First, I would like to sincerely thank my supervisor, Professor Haizhou Li from National University of Singapore. Second, this research is supported by NUS Start-up Grant FY2016 'Non-parametric approach to voice morphing'. Last, Yi Zhou is also funded by NUS research scholarship.

6. References

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

[2] B. Ramani, M. A. Jeeva, P. Vijayalakshmi, and T. Nagarajan, "A multi-level gmm-based cross-lingual voice conversion using language-specific mixture weights for polyglot synthesis," *Circuits, Systems, and Signal Processing*, vol. 35, no. 4, pp. 1283–1311, 2016.

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

[4] X. Tian, Z. Wu, S. W. Lee, Q. H. Nguyen, M. Dong, and E. S. Chng, "System fusion for high-performance voice conversion," in *INTERSPEECH*, 2015, pp. 2759–2763.

[5] M. Abe, K. Shikano, and H. 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.

[6] M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and N. Campbell, "Cross-language voice conversion evaluation using bilingual databases," *IPSJ Journal*, vol. 43, no. 7, pp. 2177–2185, 2002.

[7] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *IEEE Proc. ICASSP*, vol. 1, 1996, pp. 373–376.

[8] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, "Text-independent cross-language voice conversion," in *Proc. INTERSPEECH*, 2009.

[9] D. Erro, A. Moreno, and A. 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.

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

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

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

[13] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *IEEE APSIPA ASC*, 2016, pp. 1–6.

[14] ——, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. INTERSPEECH*, 2017, pp. 3364–3368.

[15] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv:1711.11293*, 2017.

[16] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks," in *IEEE SLT*, 2018, pp. 266–273.

[17] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *IEEE Proc. ICASSP*, 2019, pp. 6820–6824.

[18] X. Tian, J. Wang, H. Xu, E.-S. Chng, and H. Li, "Average modeling approach to voice conversion with non-parallel data," in *Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 227–232.

[19] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, "Personalized, cross-lingual tts using phonetic posteriorgrams," in *Proc. INTERSPEECH*, 2016, pp. 322–326.

[20] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *IEEE Proc. ICASSP*, 2019, pp. 6790–6794.

[21] F.-L. Xie, F. K. Soong, and H. Li, "A kl divergence and dnn-based approach to voice conversion without parallel training sentences." in *INTERSPEECH*, 2016, pp. 287–291.

[22] Y. Ning, Z. Wu, R. Li, J. Jia, M. Xu, H. Meng, and L. Cai, "Learning cross-lingual knowledge with multilingual blstm for emphasis detection with limited training data," in *IEEE Proc. ICASSP*, 2017, pp. 5615–5619.

[23] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *IEEE Proc. ICASSP*, 2019, pp. 6905–6909.

[24] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[25] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[26] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.