A MODULARIZED NEURAL NETWORK WITH LANGUAGE-SPECIFIC OUTPUT LAYERS FOR CROSS-LINGUAL VOICE CONVERSION

Yi Zhou, Xiaohai Tian, Emre Yılmaz, Rohan Kumar Das and Haizhou Li

National University of Singapore, Singapore
yi.zhou@u.nus.edu, {eletia, emre, rohankd, haizhou.li}@nus.edu.sg

ABSTRACT

This paper presents a cross-lingual voice conversion framework that adopts a modularized neural network. The modularized neural network has a common input structure that is shared for both languages, and two separate output modules, one for each language. The idea is motivated by the fact that phonetic systems of languages are similar because humans share a common vocal production system, but acoustic renderings, such as prosody and phonotactic, vary a lot from language to language. The modularized neural network is trained to map Phonetic PosteriorGram (PPG) to acoustic features for multiple speakers. It is conditioned on a speaker i-vector to generate the desired target voice. We validated the idea between English and Mandarin languages in objective and subjective tests. In addition, mixed-lingual PPG derived from a unified English-Mandarin acoustic model is proposed to capture the linguistic information from both languages. It is found that our proposed modularized neural network significantly outperforms the baseline approaches in terms of speech quality and speaker individuality, and mixed-lingual PPG representation further improves the conversion performance.

Index Terms— cross-lingual, voice conversion, mixed-lingual PPG, modularized neural network

1. INTRODUCTION

Voice conversion (VC) aims to modify the speech of one speaker (source) to sound like that of another speaker (target). Cross-lingual VC is a special case where the source and target speakers speak different languages. It is the enabling technology for many real world applications such as speech-to-speech translation [1], foreign language training [2], and movie dubbing, etc.

The early studies of cross-lingual VC rely on parallel data recorded from bilingual speakers [3,4]. These approaches utilize parallel data of the source and target speakers in the target speaker's language during training, and then convert the source speaker's utterances in the other language to the target voice. However, such bilingual source speakers are not always available in practice. Various alignment methods have been proposed to find the closest matching segments between non-parallel speech data across languages, for example, unit selection [5–8], iterative alignment [2,9–11] methods, and the vocal tract length normalization (VTLN) based phone mapping approaches [2, 12, 13]. Yet, the converted speech quality is degraded due to inaccurate alignments [11, 14]. Besides, eigenvoice-based technique [1, 15] is also developed by adapting a pre-trained eigenvoice Gaussian mixture model (EV-GMM) with a few utterances from the target speaker in a different language. Automatic speech recognition (ASR) is also employed to provide the phonetic information. In [16], Kullback-Leibler divergence (KLD) is calculated to map the senones between two languages. In [17], the source utterance is first translated to the target language, which is then used to synthesize the target utterance. Recently, neural network approaches are widely studied. Deep generative models like variational auto-encoder (VAE) [18, 19] and generative adversarial networks (GANs) [20, 21] without requiring parallel data are found effective for VC, though in the same language.

Phonetic PosteriorGram (PPG) has been proposed for VC [22–25] and also successfully applied to cross-lingual conversions [26]. PPG is a frame-level phonetic information representation obtained from the acoustic model in a speaker-independent ASR system. The conversion model is trained to map PPGs to the output acoustic features [22]. To enhance the converted speech quality, a recent study [27] investigates the use of bilingual PPG, which is formed by stacking two monolingual PPGs in source and target languages, as a linguistic representation of both languages for cross-lingual VC. Additionally, an average modeling approach [14, 28] is also proposed to leverage the linguistic and acoustic information from various speakers in different languages [27, 29].

In this paper, we propose a modularized neural network using mixed-lingual PPG for cross-lingual VC. Inspired by the language and speaker factorization acoustic modeling technique in multi-language and multi-speaker text-to-speech (TTS) [30–32], we model the linguistic to acoustic feature transformation in two steps by a shared language-independent module, and two separate language-specific output modules. The language-independent module is shared as a bridge to transfer input linguistic features from multiple speakers and languages into a common space. While the separate language-specific modules are deployed to model the acoustic

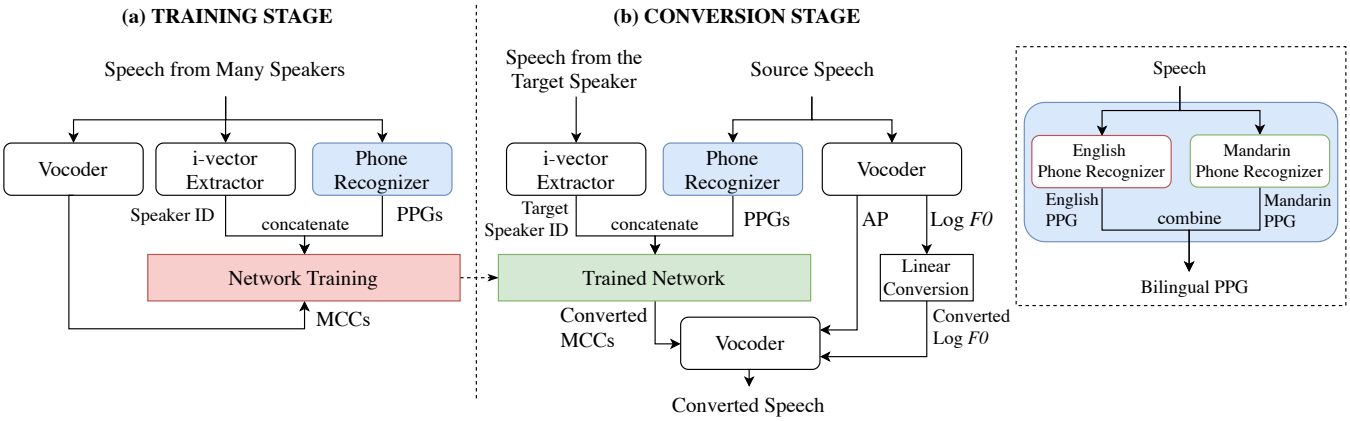

Fig. 1. Block diagram of (a) training stage and (b) conversion stage of the PPG-based cross-lingual VC system. The English and Mandarin phone recognizers convert speech into PPGs that represent linguistic information, as illustrated in the dash box. Note that the PPGs are decoded by two phone recognizers separately, the combined PPG is called Bilingual PPG.

features for each language individually. We hypothesize that the language specific output layers will improve the acoustic renderings of specific languages. Mixed-lingual PPG is extracted by a jointly trained English-Mandarin mixed-lingual acoustic model, which is able to capture the acoustic information of both languages. Hence, the assigned phone posterior probabilities are expected to reflect the shared characteristics of the two phonetic systems.

2. RELATED WORK

This section presents the average modeling cross-lingual VC using bilingual PPG [27] and motivates our proposed work.

2.1. Methodology

Bilingual PPG is an effective characterization used to capture the phonetic information of two languages [27]. Fig. 1 shows the block diagram of the bilingual PPG-based cross-lingual VC framework. To convert between English and Mandarin speakers, an English and a Mandarin phone recognizer are trained individually. Illustrated in the dash box in Fig. 1, monolingual PPGs are first extracted with English and Mandarin phone recognizer. Then, the two monolingual PPGs are combined to form a bilingual PPG to represent the linguistic information of both languages.

As shown in Fig. 1(a), during training, we first extract i-vector [33], linguistic features (PPGs), and acoustic features (mel cepstral coefficients (MCCs)) from bilingual speech data from multiple speakers. Then, we form the input features by augmenting i-vector to PPGs. The output features only contain MCCs. In this way, the mapping network $\mathcal{F}(\cdot)$ is trained to map input features $\mathbf{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_t, ..., \boldsymbol{x}_T\}$ to output features $\mathbf{Y} = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_t, ..., \boldsymbol{y}_T\}$ by the back-propagation through time (BPTT) algorithm. T indicates the number of total frames in the training data. The actual output values $\hat{\mathbf{Y}} = \{\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_t, ..., \hat{\boldsymbol{y}}_T\}$ are obtained according to the mapping network,

$$\hat{\mathbf{Y}} = \mathcal{F}(\mathbf{X}) \tag{1}$$

The conversion stage is shown in Fig. 1(b). We first extract the PPGs and target i-vector from the source speech and the target speech using the same phone recognizer and i-vector extractor, respectively. PPGs and i-vector are then concatenated to form the input features for the mapping network to convert the MCCs. Finally, the converted speech can be synthesized with linearly converted *F0*, source aperiodicity (AP) and converted MCCs.

2.2. Motivations

The successful implementation of the cross-lingual VC framework with bilingual PPG and average modeling motivates us to consider in two directions. First, despite the fact that languages share the common vocal production system [30, 34], their acoustic renderings differ very much from one to another, such as their prosodic and phonotactic renderings [35]. As a result, using only a single output layer may not be able to encode the unique language details. Second, bilingual PPG is obtained by combining English PPG and Mandarin PPG, which are extracted by two individually trained acoustic models. As the acoustic models are trained separately, the resultant bilingual PPG may not provide a unified view on the input speech from two languages. Moreover, it could be biased towards one of the languages.

3. MODULARIZED NEURAL NETWORK AND MIXED-LINGUAL PPG

In this section, we propose a modularized neural network with language-specific output layers to model a two-step acoustic

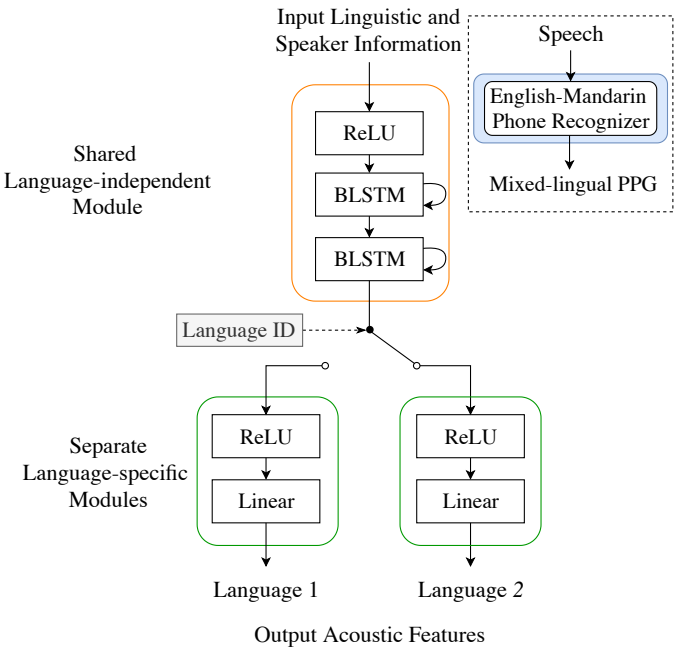

Fig. 2. The architecture of the proposed modularized neural network that employs multi-task learning. language-specific output layers. It consists of a shared language-independent module and two separate language-specific modules. Language ID serves as a switch to select the output module. The English-Mandarin phone recognizer converts speech into mixed-lingual PPGs that represent linguistic information, as illustrated in the dash box.

feature generation process. Instead of using bilingual PPG, we introduce a mixed-lingual PPG estimated by a unified English-Mandarin acoustic model.

3.1. Modularized Neural Network

Multi-task learning is a learning paradigm in machine learning, which can effectively improve the generalization performance on all tasks by sharing representations between multiple related tasks [36, 37]. Motivated by its success in many applications [38, 39], we propose a modularized neural network, as shown in Fig. 2. There are two functional blocks that represent the two steps in speech generation, namely shared language-independent module, and separate language-specific modules.

In the shared language-independent module, both linguistic information (PPGs) and speaker information (i-vector) first pass through shared layers including a rectified linear activation function (ReLU) projection layer and two bidirectional long short-term memory (BLSTM) layers. This shared module can capture the general transformation properties for both languages benefiting from multiple speakers' large database.

In the separate language-specific modules, each module consists of two language-specific output layers including a ReLU layer and a linear layer. Given a language ID, language-independent intermediate features will be first forwarded to its corresponding ReLU layer to be mapped into language-dependent vectors. The subsequent linear layer in the chosen language will then convert the learned complex features to the acoustic features. The mapping network in Equation (1) now can be expressed as

$$\hat{\mathbf{Y}} = \mathcal{F}(\mathbf{X}) = \begin{cases} \mathcal{L}_{en}(\mathcal{S}(\mathbf{X})), & \text{English} \\ \mathcal{L}_{cn}(\mathcal{S}(\mathbf{X})), & \text{Mandarin} \end{cases} \quad (2)$$

where $\mathcal{S}(\cdot)$ indicates the shared language-independent module. $\mathcal{L}_{en}(\cdot)$ and $\mathcal{L}_{cn}(\cdot)$ denote the English and Mandarin language-specific modules, respectively.

During training, the language ID serves as a switch to select the desired language-specific module. The whole network is trained jointly to minimize the mean square errors between the original acoustic features and the predicted ones in each language. In each training step, only the relevant gradients will be computed to update the associated weights, while the loss in the other language will be set to zero.

3.2. Mixed-lingual PPG

Mixed-lingual PPG is obtained by a unified English-Mandarin acoustic model acoustic model for code-switch speech recognition. Specifically, a time-delay neural network-based acoustic model [40,41] is trained with the senones as targets. Being exposed to speech data from both languages during training, this model learns to capture the acoustic similarity while differentiate the specific phones in the two languages. During testing, joint posterior probability distributions are obtained by passing the test frames through the trained model. By doing so, this unified acoustic model takes the union of two phonetic systems into one as if it were for the same phonetic system. The dash box in Fig. 2 shows the phone recognizer for mixed-lingual PPG extraction. Instead of using two phone recognizers in each language as in Fig. 1, now it only consists of a single unified English-Mandarin acoustic model.

4. EXPERIMENTS

4.1. Database and Feature Extraction

We conducted cross-lingual VC experiments between English and Mandarin languages. 10 English speakers (5 female and 5 male) from VCC2016 database [42] and 10 Mandarin speakers (5 female and 5 male) from the Mandarin average model database[1] were chosen for training. Each speaker provided 162 utterances, where 150 utterances were used for training, while the other 12 utterances were used for validation. In total, we had 3,000 utterances from 20 speakers for the average model training. For testing, we selected 4 bilingual speakers (2 female and 2 male) from the EMIME Mandarin

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

Table 1. The composition of the training and test data.

	Database	Selected Speakers
Training	VCC2016 [42] *English*	SF1, SF2, SF3, TF1, TF2 SM1,SM2,TM1,TM2,TM3
	Average Model *Mandarin*	01F, 02F, 03F, 04F, 05F 07M, 08M, 09M, 12M, 13M
Test	EMIME [43] *English/Mandarin*	MF4, MF5 MM1,MM2

Bilingual Database [43]. We converted 20 utterances for each conversion pair. Both intra-gender and inter-gender conversions were performed between the selected speakers, as summarized in Table 1.

All speech signals were sampled at 16kHz with a frame shift of 5ms. The speech data were first analyzed by the WORLD vocoder [44] for feature extraction, generating the spectrum (513-dim), AP (1-dim), and $F0$ (1-dim) features. Then the Speech Signal Processing Toolkit[2] was used to compute the 40-dimensional MCCs.

Kaldi toolkit [45] was used to train the English, Mandarin and the unified English-Mandarin acoustic models. A time-delay neural network with 10 hidden layers and 1280 hidden units per layer was trained with lattice-free maximum mutual information (LF-MMI) [46] criterion using 40-dimensional mel frequency cepstral coefficients (MFCC) features of approximately 200 hours speech data from each language. The hyperparameters were chosen according to the standard recipe provided for the Switchboard database in the Kaldi toolkit [45]. The acoustic data was a random subset of the open-source Librispeech [47] and AI-SHELL2 [48] corpora. The frame rate was adjusted to VC task and set to 5ms. For a similar reason, the frame subsampling rate was set to 1.

The output layers for English and Mandarin had 4,193 and 6,360 context-dependent phone states respectively. To compensate for the imbalance in the size of English and Mandarin phonetic alphabets due to the tonality, we have used 213 separate position-dependent models for English phones and 199 position-independent models for Mandarin phones. With extra silence and spoken noise entries, the English, Mandarin and unified English-Mandarin acoustic model had 215 (213 + 2), 201 (199 + 2) and 414 (213 + 199 + 2) classes, respectively. As a result, a linguistic feature frame of either bilingual PPG or mixed lingual PPG had 414 elements.

For i-vector extractor, the universal background model (UBM) contained 502 speakers (251 female and 251 male) from Switchboard II corpus. In total, we used 1,872 utterances, each was 5 minutes long on average. MFCC features were used to derive the i-vectors. We used gender-independent UBM of 1024 mixture components and total variability matrix with 400 speaker factors. The i-vector dimension were reduced from 400 to 150 after applying LDA.

By combining the PPG frame of 414 elements, and an i-vector of 150 elements, we obtain an input feature frame of 564 dimensions. The output acoustic feature contained 127 elements including a voiced/unvoiced flag (1-dim), the original and their delta and delta-delta coefficients of MCCs (40-dim), log $F0$ (1-dim), AP (1-dim).

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

4.2. Experimental Setup

We have implemented 4 different cross-lingual VC systems for comparison. The details of the baseline and proposed systems were discussed as follows.

- **bPPG-LI:** The cross-lingual VC system [27] using bilingual PPG (bPPG) with a single language-independent (LI) output layer as discussed in Section 2.1. The model consisted of two BLSTM layers with 256 hidden units in each layer. This is benchmarked as our baseline.
- **mPPG-LI:** The cross-lingual VC system using mixed-lingual PPG (mPPG) introduced in Section 3.2 with a single language-independent (LI) output layer. The model used the same network configurations as bPPG-LI.
- **bPPG-LS:** The cross-lingual VC system using bilingual PPG (bPPG) and the proposed modularized neural network with language-specific (LS) output layers as discussed in Section 3.1. All shared layers had 256 nodes including one input ReLU layer and two hidden BLSTM layers. Both language-specific output ReLU layers had 128 nodes.
- **mPPG-LS:** The cross-lingual VC system using mixed-lingual PPG (mPPG) based on our proposed modularized neural network with language-specific (LS) output layers. The model used the same hyper-parameters as bPPG-LS.

Merlin toolkit [49] was used for model training. All models were trained with a learning rate of 0.002, the minibatch size and momentum were set as 25 and 0.9 respectively.

During conversion, $F0$ were converted by a global linear transformation in log-scale [50]. APs were directly copied from source speech, and the MCCs were generated by Maximum Likelihood Parameter Generation algorithm [51]. The post-filtering processing technique in the cepstral domain was used to enhance the conversion quality [52].

4.3. Objective Evaluation

For objective evaluation, mel-cepstrum distortion (MCD) was used to measure the spectral distance between the converted and original speech from bilingual speakers, defined as

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

where D is the MCC feature dimension, $\hat{Y}_d$ and Y_d are the d^{th} coefficients of the converted and original MCCs, respectively. A lower value accounts for a smaller distortion.

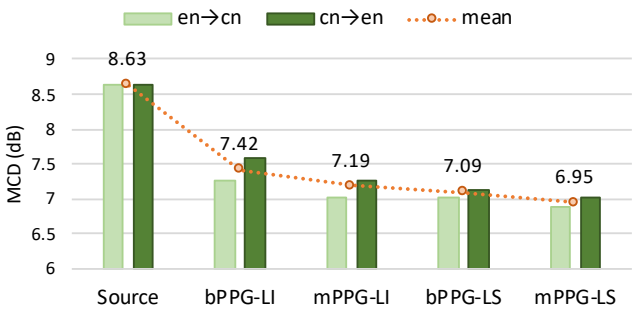

Fig. 3. Average MCD results for cross-lingual VC using bilingual test speakers. Source indicates the source speech without conversion.en → cn means that we convert a source English utterance to a Mandarin target speaker, and vice-versa.

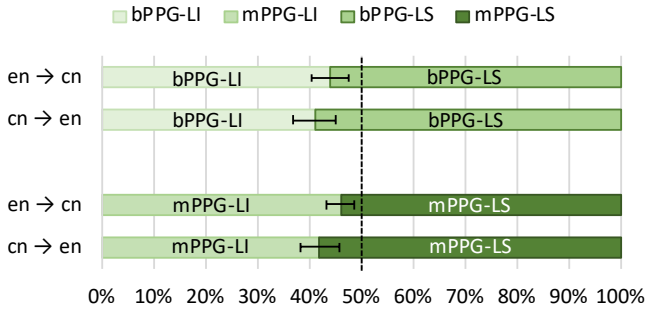

Fig. 4. AB preference test results for speech quality with 95% confidence intervals between systems using proposed language-specific and the baseline language-independent output layers (i.e., bPPG-LI vs. bPPG-LS; and mPPG-LI vs. mPPG-LS). en → cn means that we convert a source English utterance to a Mandarin target speaker, and vice-versa.

Average MCDs of cross-lingual VC are presented in Fig. 3. First, we compare the proposed modularized neural network using the language-specific output layers with the baselines. It is observed that both bPPG-LS and mPPG-LS outperform their counterparts (bPPG-LI and mPPG-LS) with the average MCD dropping from 7.42dB to 7.09dB and 7.19dB to 6.95dB, respectively. It suggests that the proposed modularized neural network yields better conversion performance than the baselines, in terms of MCD. Then we further evaluate the performance of the proposed mixed-lingual PPG. It can be found that systems using mixed-lingual PPG (mPPG-LI and mPPG-LS) consistently outperform those using bilingual PPG (bPPG-LI and bPPG-LS).

4.4. Subjective Evaluation

We further conducted listening tests. AB preference and mean opinion score (MOS) tests were conducted to assess speech quality. Meanwhile, VCC similarity tests, the Same/Different

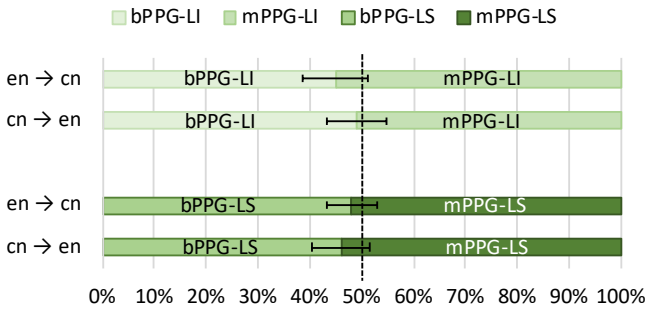

Fig. 5. AB preference test results for speech quality with 95% confidence intervals between systems using proposed mixed-lingual PPG and baseline bilingual PPG (i.e., bPPG-LI vs. mPPG-LI; and bPPG-LS vs. mPPG-LS). en → cn means that we convert a source English utterance to a Mandarin target speaker, and vice-versa.

paradigm from the VCC 2016 [42], were conducted to assess speaker similarity. In addition, to further study listeners' preferences across all VC systems, we also conducted best-worst scaling (BWS) tests [53] on both speech quality and speaker similarity. 12 samples were randomly selected from (4 x 20 = 80) converted samples of each VC system. 12 English and Mandarin bilingual listeners participated all tests.

a) *AB preference tests:* In AB preference tests, A and B were the randomly selected converted speech samples from different systems, and the listeners were asked to compare their voice quality and naturalness.

   First, we examine the effectiveness of the proposed modularized neural network. The speech quality test results are presented in Fig. 4. It is observed that mPPG-LS and bPPG-LS significantly outperform mPPG-LI and bPPG-LI, respectively. And the performance improvement is more remarkable in en → cn than that in cn → en. The test results demonstrate that our proposed modularized neural network is more effective than the baseline network using a single output layer in terms of speech quality.

   Then, we further compare the proposed mixed-lingual PPG with bilingual PPG. Fig. 5 shows the speech quality test results, which suggests that systems using proposed mixed-lingual PPG outperform those using bilingual PPG in both en → cn and cn → en, though the difference is not statistically significant. The speech quality test results confirm that we can consider mixed-lingual PPG as a better representation than bilingual PPG to characterize the linguistic information from two languages.

b) *MOS tests:* In MOS tests, listeners were asked to rate the quality and naturalness of the converted speech on a 5-point scale. The results are presented in Table. 2. It can be firstly observed that mPPG-LS obtained the highest scores in both en → cn and cn → en. We also find that bPPG-LS and

Table 2. MOS test results over all cross-lingual VC systems. en → cn means that we convert a source English utterance to a Mandarin target speaker, and vice-versa.

System	en → cn	cn → en
bPPG-LI	2.87 ± 0.16	2.42 ± 0.72
mPPG-LI	2.98 ± 0.24	2.45 ± 0.58
bPPG-LS	3.55 ± 0.28	3.43 ± 0.38
mPPG-LS	$\mathbf{3.62} \pm 0.28$	$\mathbf{3.53} \pm 0.37$

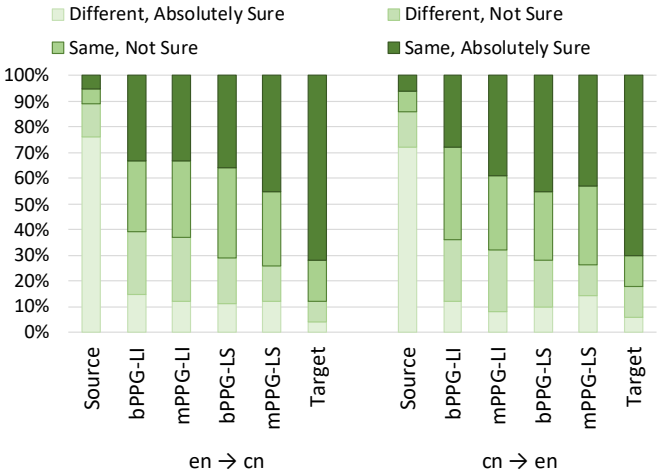

Fig. 6. VCC similarity test result, en → cn means that we convert a source English utterance to a Mandarin target speaker, and vice-versa. Source and target denote the source speaker and target speaker respectively.

mPPG-LS achieve significant improvement over bPPG-LI and mPPG-LI, respectively. The results suggest that our proposed modularized neural network using mixed-lingual PPG is fairly effective in cross-lingual VC.

c) *VCC similarity tests:* In VCC similarity tests, the listeners were asked to compare and select whether the converted samples were uttered by the same target speaker [54]. Fig. 6 presents the similarity test results. The system performances, ranked in ascending order (from the lowest to the highest), are as follows: bPPG-LI, mPPG-LI, bPPG-LS, and mPPG-LS. The results are consistent in both en → cn and cn → en conversions. It therefore demonstrates the capability of our proposed approaches for converting the speaker's identity in cross-lingual VC tasks.

d) *BWS tests:* In BWS tests, we presented 4 converted samples with the same content from each VC system to the listeners. In the quality tests, we asked listeners to compare all the samples and select the best and worst samples in terms of speech quality. In the similarity tests, we provided the target speech as a reference. The listeners were asked to select the most and the least similar samples comparing to the reference in terms of speaker identity. The BWS test results are presented in Table 3. In both quality and similarity tests, mPPG-LS takes the highest percentages (55% and 43%) as the best-converted sample among all VC systems. While, it is chosen as the worst sample at the lowest percentages (16% and 11%). The results also indicate that bPPG-LI performs the worst while bPPG-LS and mPPG-LI achieve the moderate results.

Table 3. BWS test results on (a) speech quality; (b) speaker similarity. N/P means no preference.

Test	System	Best (%)	Worst (%)	N/P (%)
(a)	bPPG-LI	4	42	54
	mPPG-LI	14	20	65
	bPPG-LS	27	22	51
	mPPG-LS	**55**	**16**	29
(b)	bPPG-LI	16	44	40
	mPPG-LI	19	25	56
	bPPG-LS	22	20	58
	mPPG-LS	**43**	**11**	46

According to the subjective tests, the proposed modularized neural network using mixed-lingual PPG (mPPG-LS) consistently outperformed all the rest VC systems in both speech quality and speaker similarity tests, which demonstrates the effectiveness of our proposed approaches for cross-lingual voice conversions. The converted samples can be found from this demo link[3].

5. CONCLUSION

In this paper, we proposed a cross-lingual voice conversion framework based on a modularized neural network using mix-language PPG. By utilizing the shared input module to be language independent while decomposing the output modules to be language specific, the network is robust for output acoustic feature modeling across different languages. Meanwhile, the mixed-lingual PPG extracted from a unified English-Mandarin acoustic model also provides an accurate linguistic representation for conversion quality enhancement. Experimental results successfully demonstrate our proposed approaches outperform the baseline approaches on both speech quality and speaker similarity.

6. ACKNOWLEDGEMENT

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A2b0046), the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-100E-2018-006), and the NUS Start-up Grant FY2016 Non-parametric Approach to Voice Morphing. Yi Zhou is also funded by the NUS research scholarship.

[3] https://vcsamples.github.io/asru_2019/

7. REFERENCES

[1] B Ramani, MP Actlin 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.

[2] 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.

[3] Masanobu Abe, Kiyohiro Shikano, and Hisao Kuwabara, "Cross-language voice conversion," in *IEEE ICASSP*, 1990, pp. 345–348.

[4] 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.

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

[6] Helenca Duxans, Daniel Erro, Javier Pérez, Ferran Diego, Antonio Bonafonte, and Asunción Moreno, "Voice conversion of non-aligned data using unit selection," *TC-STAR WSST*, 2006.

[7] David Sündermann, Harald Höge, Antonio Bonafonte, Hermann Ney, and Julia Hirschberg, "Text-independent cross-language voice conversion," in *INTERSPEECH*, 2009.

[8] 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.

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

[10] Alejandro José Uriz, Pablo Daniel Agüero, Antonio Bonafonte, and Juan Carlos Tulli, "Voice conversion using k-histograms and frame selection," in *INTERSPEECH*, 2009.

[11] 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.

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

[13] Srinivas Desai, B Yegnanarayana, and Kishore Prahallad, "A framework for cross-lingual voice conversion using artificial neural networks," *7th ICON*, 2009.

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

[15] Malorie Charlier, Yamato Ohtani, Tomoki Toda, Alexis Moinet, and Thierry Dutoit, "Cross-language voice conversion based on eigenvoices," in *INTERSPEECH*, 2009, pp. 1635–1638.

[16] Feng-Long Xie, Frank K Soong, and Haifeng Li, "A KL Divergence and DNN-Based Approach to Voice Conversion without Parallel Training Sentences," in *INTERSPEECH*, 2016, pp. 287–291.

[17] Zhenye Gan, Xiaotian Xing, Hongwu Yang, and Guangying Zhao, "Mandarin-tibetan cross-lingual voice conversion system based on deep neural network," in *ACM CSAI*, 2018, pp. 67–71.

[18] 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.

[19] Yuki Saito, Yusuke Ijima, Kyosuke Nishida, and Shinnosuke Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *IEEE ICASSP*, 2018, pp. 5274–5278.

[20] 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, pp. 282–289.

[21] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *IEEE ICASSP*, 2019, pp. 6820–6824.

[22] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *IEEE ICME*, 2016, pp. 1–6.

[23] 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.

[24] Songxiang Liu, Jinghua Zhong, Lifa Sun, Xixin Wu, Xunying Liu, and Helen Meng, "Voice conversion across arbitrary speakers based on a single target-speaker utterance," *INTERSPEECH*, pp. 496–500, 2018.

[25] Xiaohai Tian, Eng Siong Chng, and Haizhou Li, "A speaker-dependent wavenet for voice conversion with non-parallel data," in *INTERSPEECH*, 2019, pp. 201–205.

[26] 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.

[27] Yi Zhou, Xiaohai Tian, Haihua Xu, Rohan Kumar Das, and Haizhou Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *IEEE ICASSP*, 2019, pp. 6790–6794.

[28] Xiaoxue Gao, Xiaohai Tian, Rohan Kumar Das, Yi Zhou, and Haizhou Li, "Speaker-independent spectral mapping for speech-to-singing conversion," in *IEEE APSIPA ASC (Accepted)*, 2019.

[29] Yi Zhou, Xiaohai Tian, Rohan Kumar Das, and Haizhou Li, "Many-to-many cross-lingual voice conversion with a jointly trained speaker embedding network," in *IEEE APSIPA ASC (Accepted)*, 2019.

[30] Heiga Zen, Norbert Braunschweiler, Sabine Buchholz, Mark JF Gales, Kate Knill, Sacha Krstulovic, and Javier Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 6, pp. 1713–1724, 2012.

[31] Yuchen Fan, Yao Qian, Frank K Soong, and Lei He, "Speaker and language factorization in DNN-based TTS synthesis," in *IEEE ICASSP*, 2016, pp. 5540–5544.

[32] Bo Li and Heiga Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in *INTERSPEECH*, 2016.

[33] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[34] Yishuang Ning, Zhiyong Wu, Runnan Li, Jia Jia, Mingxing Xu, Helen Meng, and Lianhong Cai, "Learning cross-lingual knowledge with multilingual BLSTM for emphasis detection with limited training data," in *IEEE ICASSP*, 2017, pp. 5615–5619.

[35] Haizhou Li, Bin Ma, and Kong Aik Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[36] Rich Caruna, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[37] Yu Zhang and Qiang Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

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

[39] Mingyang Zhang, Xin Wang, Fuming Fang, Haizhou Li, and Junichi Yamagishi, "Joint Training Framework for Text-to-Speech and Voice Conversion Using Multi-Source Tacotron and WaveNet," in *INTERSPEECH*, 2019, pp. 1298–1302.

[40] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang, "Phoneme recognition using time-delay neural networks," *Backpropagation: Theory, Architectures and Applications*, pp. 35–61, 1995.

[41] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015, pp. 3214–3218.

[42] 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.

[43] Mirjam Wester and Hui Liang, "The emime mandarin bilingual database," Tech. Rep. EDI-INF-RR-1396, The University of Edinburgh, 2011.

[44] 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.

[45] 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.

[46] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *INTERSPEECH*, 2016, pp. 2751–2755.

[47] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE ICASSP*, 2015, pp. 5206–5210.

[48] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu, "AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale," *arXiv preprint arXiv:1808.10583*, 2018.

[49] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An open source neural network speech synthesis system," *SSW*, 2016.

[50] 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.

[51] 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.

[52] Yoshimura Takayoshi, Tokuda Keiichi, Masuko Takashi, Kobayashi Takao, and Kitamura Tadashi, "Incorporating a Mixed Excitation Model and Postfilter into HMM-Based Text-to-Speech Synthesis," *Systems and Computers in Japan*, vol. 36, no. 12, pp. 43–50, 2005.

[53] Berrak Sisman, Mingyang Zhang, and Haizhou Li, "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1085–1097, 2019.

[54] 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.