Spectrum and Prosody Conversion for Cross-lingual Voice Conversion with CycleGAN

Zongyang Du *, Kun Zhou *, Berrak Sisman [†]* and Haizhou Li *

* Department of Electrical and Computer Engineering, National University of Singapore, Singapore

[†] Information Systems Technology and Design (ISTD) Pillar, Singapore University of Technology and Design, Singapore

Abstract—Cross-lingual voice conversion aims to change source speaker's voice to sound like that of target speaker, when source and target speakers speak different languages. It relies on non-parallel training data from two different languages, hence, is more challenging than mono-lingual voice conversion. Previous studies on cross-lingual voice conversion mainly focus on spectral conversion with a linear transformation for F0 transfer. However, as an important prosodic factor, F0 is inherently hierarchical, thus it is insufficient to just use a linear method for conversion. We propose the use of continuous wavelet transform (CWT) decomposition for F0 modeling. CWT provides a way to decompose a signal into different temporal scales that explain prosody in different time resolutions. We also propose to train two Cycle-GAN pipelines for spectrum and prosody mapping respectively. In this way, we eliminate the need for parallel data of any two languages and any alignment techniques. Experimental results show that our proposed *Spectrum-Prosody-CycleGAN* framework outperforms the *Spectrum-CycleGAN* baseline in subjective evaluation. To our best knowledge, this is the first study of prosody in cross-lingual voice conversion.
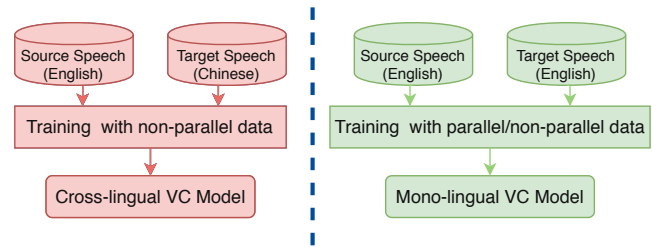
Index Terms: Cross-lingual voice conversion, prosody, CycleGAN, continuous wavelet transform

I. Introduction

Voice conversion (VC) aims to convert speaker characteristics from source speaker to target speaker [1]. It is an enabling technique for many applications, such as text-to-speech synthesis [2], [3] and emotional voice conversion [4], [5].

In this paper, we focus on cross-lingual voice conversion [6], where the source and target speakers speak different languages. Cross-lingual voice conversion is more challenging than mono-lingual voice conversion [7] because source and target speakers speak in two different phonetic systems and prosodic styles. Furthermore, parallel training data is not easily available [8]–[10]. The comparison between cross-lingual and mono-lingual voice conversion is illustrated in Fig. 1(a). and Fig. 1(b). In the training phase, cross-lingual voice conversion is trained with non-parallel speech from the source and target speaker speaking with different languages, whereas mono-lingual voice conversion can be trained with parallel/non-parallel speech from source and target speaker speaking with the same language [11]–[14].

Previous studies on mono-lingual voice conversion mainly focus on finding a mapping function between the source and target through parallel training data, which includes Gaussian

(a) Training phases of mono-lingual and cross-lingual voice conversion.
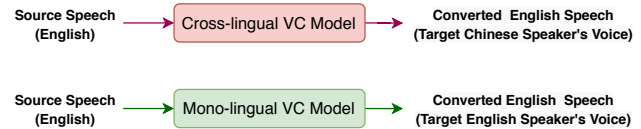
(b) Mono-lingual voice conversion converts source to target speech in the same language at run-time. However, cross-lingual voice conversion converts an English utterance from an English speaker to a Chinese speaker, or a Chinese utterance from a Chinese speaker to an English speaker. Taking the former as an example, we would like the converted voice to sound like a native English speaker with voice characteristics similar to the Chinese speaker.

Fig. 1. A comparison between cross-lingual and mono-lingual voice conversion in the training and conversion phase. Red boxes represent cross-lingual voice conversion, while green boxes represent mono-lingual voice conversion.

mixture model (GMM) [15], non-negative matrix factorization (NMF) based sparse representation [16], [17], and group sparse representation [7]. Recent deep learning approaches, such as deep neural network (DNN) [18] and generative adversarial network (GAN) [19], have greatly improved conversion quality.

One way of cross-lingual voice conversion is to rely on multilingual speakers to provide same-speaker, cross-lingual training data. Some statistical approaches for spectral mapping include codebook mapping [20] and Gaussian mixture model (GMM) [9], that show comparable performance with that of mono-lingual voice conversion. However, collecting such data from a bilingual speaker can be expensive and time-consuming. Besides, system performance also depends largely on the speaker's proficiency in the language pair. To circumvent the need for bilingual data, hidden Markov model (HMM) [21], unit selection [22], [23] and the iterative frame alignment [24], [25] were proposed to find source-target frame pairs from non-parallel utterances.

More recently, GAN-based methods, including cycle-consistent adversarial network (CycleGAN) [26]–[28] and variational autoencoding wasserstein generative adversarial networks (VAW-GAN) [29], [30] have achieved high-quality mono-lingual voice conversion with non-parallel data. GAN-based methods for cross-lingual voice conversion [10] have shown similar results with monolingual voice conversion without the need for parallel training data, nor external modules (such as ASR).

Unfortunately, prosody conversion for cross-lingual voice conversion has not been well studied. The differences between two languages lie not only in phonetic systems, but also in linguistic prosody and speaking style, that are characterized by the F0 contour of speech. Motivated by the success of CycleGAN in spectral conversion, we would like to study the conversion of both spectrum and prosody in this paper. In the context of cross-lingual voice conversion, the source speaker is a native speaker of the language, therefore, the source linguistic prosody, such as sentence-level pitch trajectory, should be carried over to the target as much as possible, while the conversion is expected to handle speaker-dependent prosody elements such as pitch level, phoneme-level pitch patterns etc.

It is well known that speaker characteristics are usually related to: 1) prosodic factors which are concerned with syllables and larger units of speech instead of individual phonetic segments (vowels and consonants); and 2) segmental factors that involve short time spectrum. As an essential prosodic factor, fundamental frequency (F0) represents the variation of the vocal pitch over the whole time domains [31]. Previous voice conversion studies propose to convert F0 with a simple linear transformation [8] or some statistical methods such as GMM [32]. Since F0 is hierarchical in nature and affected by short-term as well as long-term dependencies, we believe that these methods are insufficient to model the variations of F0 in different temporal scales [4], [33]. Therefore, we propose to use CWT decomposition to analyse F0 in different time scales, and find a mapping for each time scale. The CWT decomposition describes a frame-based F0 value with a set of CWT coefficients, that represent prosodic features. We note that this is the first study of F0 modeling for cross-lingual voice conversion.

In this paper, we propose a cross-lingual voice conversion framework based on CycleGAN, that can learn the mappings of spectrum and prosody between the speakers speaking with different languages. We also use CWT to decompose F0 into 10 different temporal scales to describe the prosody from the phone level to the utterance level. It is noted that our proposed framework does not rely on any training data from bilingual speakers or any other external modules such as speech recognition or time alignment procedures.

The main contributions of this paper include: 1) we propose a cross-lingual voice conversion framework based on Cycle-GAN to convert the spectrum and prosody; 2) we propose to analyze F0 in different time resolutions with CWT; and 3) we explored the effect of prosody conversion in cross-lingual

voice conversion. To our best knowledge, this paper reports the first attempt to incorporate generative models and CWT-based prosody analysis for cross-lingual voice conversion.

The rest of this paper is organized as follows: In Section II, we provide the motivation and related work to set the stage for this study. In Section III, we propose a spectro-prosodic cross-lingual voice conversion framework. In Section VI, the experiment results are presented. Section V concludes the discussion.
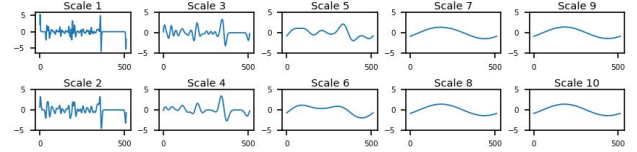
(a) 10 scales of CWT coefficients of an English utterance by one female speaker: "And to wonder what was going to happen next".
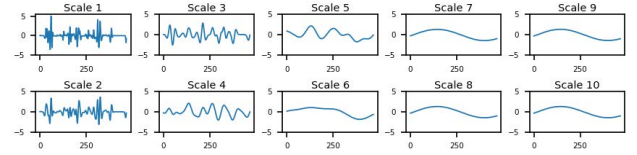
(b) 10 scales of CWT coefficients of an English utterance by another female speaker: "And to wonder what was going to happen next".

Fig. 2. Illustration of 10-scale CWT coefficients of the same utterance but spoken by two different female speakers. Lower scales capture the short-term variations (such as phoneme) and higher scales capture the long-term variations (such as sentence).

II. RELATED WORK

A. Prosody Modelling with CWT

Generally speaking, speech can be characterized by spectral and prosodic features [34], [35]. Prosody is supra-segmental and hierarchical in nature, hence prosody modelling is not as straghtforward as frame-based spectral modelling in voice conversion.

Continuous wavelet transform (CWT) has shown to be effective in simultaneous analysis and visualization of various time scales of a signal [36]. It provides a way to describe F0 in different time scales. We also note that CWT has been successfully applied in analysis and modelling of F0 in speech synthesis [36]–[38] and mono-lingual voice conversion [33], [39]. In this paper, we further this idea and decompose F0 into 10 temporal scales.

Given a continuous input signal f, its continuous wavelet transform $W(\tau, t)$ can be written as:

$$W(\tau, t) = \tau^{-1/2} \int_{-\infty}^{+\infty} f(x)\psi(\frac{x-t}{\tau})dx \qquad (1)$$

where ψ is the Mexican hat mother wavelet. The original signal f can be recovered from the wavelet representation $W(\tau, t)$ by inverse transform using the following formula:

$$f(t) = \int_{-\infty}^{+\infty} \int_{0}^{+\infty} W(\tau, t) \tau^{-5/2} \psi\left(\frac{x-t}{\tau}\right) dx d\tau \qquad (2)$$

(a) The original F0 contour

(b) F0 contour after pre-processing
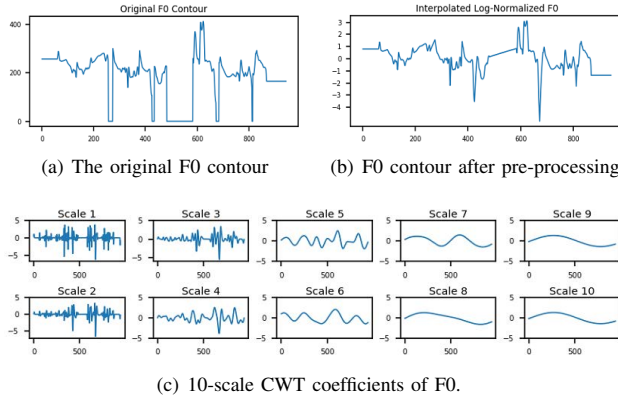

(c) 10-scale CWT coefficients of F0.

Fig. 3. 10-scale CWT coefficients of an English utterance spoken by the source speaker: "Thousands of men were rushing into the Northland".

Suppose that we decompose the input signal f into 10 scales [40], [41], f can be represented by 10 separate components given by:

$$W_i(f,t) = W(f)(2^{i+1}\tau_0, t)(i + 2.5)^{-5/2} \quad (3)$$

where $i = 1, ..., 10$ and $\tau_0 = 5ms$. These time scales were originally proposed in [38].

Given 10 wavelet components, $\hat{W}_i(f,t)$, that are the converted version of CWT components for target speaker, we can recompose signal $\hat{f}$ by the following formula [41]:

$$\hat{f}(t) = \sum_{i=1}^{10} \hat{W}_i(f,t)(i + 2.5)^{-5/2} \quad (4)$$

As can be observed in Fig. 2(a) and 2(b) where two speakers read the same text, high scale (e.g. scale 9 and 10) coefficients are similar between speakers, while low scale (e.g. scale 1 and 2) coefficients are speaker-dependent [33].

With the multi-scale CWT decomposition, we can now represent a F0 sequence using a sequence of CWT coefficient frames, in parallel to spectral frames of an utterance. We expect to train a prosody mapping function to learn the mapping between individual CWT coefficients, in particular to reflect, the speaker style transfer. Fig. 3(c) and 4 show an example of English-Chinese training pair from two different speakers. We will study the prosody mapping in Section III.

B. CycleGAN for Style Transfer

Generative adversarial networks (GANs) [42]–[44] provide a way of representing and modelling the high-dimensional distribution of data. GANs learn deep representation and generate many different acceptable answers by implicitly modelling the high-dimensional distribution of data [45]. GANs consist of two neural networks, a generator and a discriminator, which compete with each other. The generator creates samples following the same distribution of the training set to fool the discriminator while the discriminator distinguishes whether they are real samples from training set or fake generated samples [46], [47]. GANs have shown remarkable performance
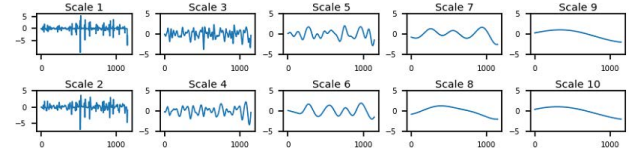

Fig. 4. 10-scale CWT coefficients of a Chinese utterance spoken by the target speaker: "法国人民深深铭记着将军对法兰西民族的丰功伟绩".

in the fields of computer version [48]–[53], natural language processing [54], [55], speech synthesis [28], [56], [57] and voice conversion [10], [19], [27].

CycleGAN, a successful implementation of GANs, was first proposed for unpaired image-to-image translation [50], and then applied to other non-parallel style transfer tasks, such as speech synthesis and mono-lingual voice conversion [26], [58], [59]. CycleGAN has also been used for spectrum mapping in cross-lingual voice conversion [10]. In this paper, we extend the idea and formulate a CycleGAN that learns prosody mapping of CWT coefficients without the need of parallel or bilingual data. Despite different languages, as shown in Fig. 3(c) and Fig. 4, the low scale CWT coefficients of F0 between the two utterances could be very different, and the high scale CWT coefficients of F0 are similar. We expect that CycleGAN is able to learn the differentiated mapping relationship between low and high scale coefficients.

III. CYCLEGAN BASED CROSS-LINGUAL VOICE CONVERSION

In this section, we propose a CycleGAN-based cross-lingual voice conversion framework that effectively learns a mapping between source and target speakers from two different languages. We also use CWT to decompose F0 into 10 different time scales for both source and target speakers, ranging from the micro-prosody level to the whole utterance level, with the aim to describe the prosody in different time resolutions.

A. Training of Spectrum and Prosody Conversion with Cycle-GANs

We first use WORLD vocoder to extract spectral features and fundamental frequency (F0) from source and target waveform. It is noted that the extracted F0 features are discontinuous due to the voiced/unvoiced activities in an utterance, as shown in Fig. 3(a). Since CWT is sensitive to the gaps in the extracted F0 contour [31], [39], the following preprocessing steps are necessary [7], [33]: 1) linear interpolation over unvoiced regions; 2) transformation of F0 from linear to a logarithmic scale; and 3) normalizing the resulting F0 to zero mean and unit variance. The continuous F0 signals after preprocessing is shown in Fig. 3(b). After the pre-processing, we obtain the 10-scale CWT-F0 coefficients, and a 10-dimensional prosody frame.

For spectrum conversion, we extract 24-dimensional Melcepstral coefficients (MCEPs) features for spectral Cycle-GAN model, denoted as *Spectrum-CycleGAN*. As for prosody conversion, we use 10-dimensional CWT-F0 coefficients as
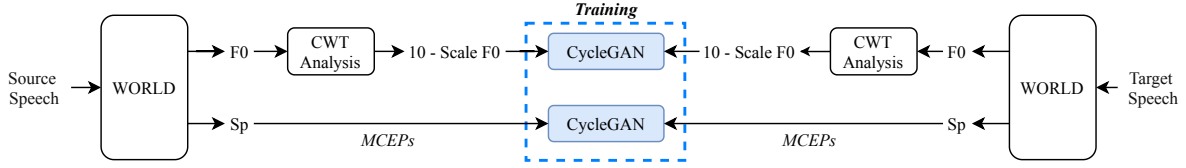
Fig. 5. The training phase of the proposed CycleGAN-based cross-lingual VC framework. Blue box represents the cross-lingual VC model for prosody conversion called *F0-CycleGAN* and yellow box is the cross-lingual VC model for spectrum conversion called *MCEP-CycleGAN*.
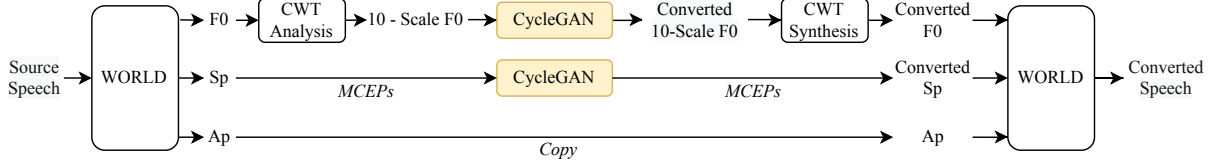

Fig. 6. The run-time conversion phase of the proposed *Spectrum-Prosody-CycleGAN* framework. Colored boxes represent the trained models in Fig. 5.

prosodic features for prosodic CycleGAN model, and denoted as *Prosody-CycleGAN*. The same training procedure applies to the training of both models.

During training, CycleGAN is capable of learning the forward and inverse mappings simultaneously through the non-parallel training data and trying to find an optimal pseudo pair for the spectrum and prosody conversion. The training phase of the proposed framework is shown in Fig. 5, that includes a spectrum and a prosody modeling pipeline. Previous study on emotional voice conversion [4] has shown that separate training of the spectrum and prosody can achieve better performance than joint training. Therefore, we train the two CycleGANs separately.

We assume that source and target speaker speak two different languages: English and Chinese respectively, which are denoted as en and cn. The goal of CycleGAN is to learn a mapping between the source $x_{en} \in X_{en}$ and the target $y_{cn} \in Y_{cn}$. CycleGAN is incorporated with three main loss functions: adversarial loss, cycle-consistency loss, and identity-mapping loss [26].

An adversarial loss measures how distinguishable between the data distribution of converted data and source or target data. For the forward mapping, it is defined as:

$$L_{ADV}(G_{X_{en} \to Y_{cn}}, D_{Y_{cn}}) =$$
$$\mathbb{E}_{y_{cn}} P_{Y_{cn}(y_{cn})}[\log(D_{Y_{cn}}(y_{cn}))]$$
$$+\mathbb{E}_{x_{en} P_{X_{en}}(x_{en})}[\log(1 - D_{Y_{cn}}(G_{X_{en} \to Y_{cn}}(x_{en})))] \quad (5)$$

Since the adversarial loss only restricts $G_{X \to Y}(x)$ to follow the target distribution, a cycle-consistency loss is proposed in order to guarantee the consistency of the contextual information between input and output. It is defined as:

$$L_{CYC}(G_{X_{en} \to Y_{cn}}, G_{Y_{cn} \to X_{en}}) =$$
$$\mathbb{E}_{x_{en} \sim P(x_{en})}[\|G_{Y_{cn} \to X_{en}}(G_{X_{en} \to Y_{cn}}(x_{en})) - x_{en}\|_1]$$
$$+\mathbb{E}_{y_{cn} \sim P(y_{cn})}[\|G_{X_{en} \to Y_{cn}}(G_{Y_{cn} \to X_{en}}(y_{cn})) - y_{cn}\|_1] \quad (6)$$

Cycle-consistency loss encourages the forward mapping $G_{X_{en} \to Y_{cn}}$ and the inverse mapping $G_{Y_{cn} \to X_{en}}$ to find an optimal pseudo pair of (x_{en}, y_{cn}) through circular conversion.

In order to preserve linguistic information without any external processes, identity mapping loss is defined as:

$$L_{ID}(G_{X_{en} \to Y_{cn}}, G_{Y_{cn} \to X_{en}}) =$$
$$\mathbb{E}_{x_{en} \sim P(x_{en})}[\|G_{Y_{cn} \to X_{en}}(x_{en}) - x_{en}\|]$$
$$+\mathbb{E}_{y_{cn} \sim P(y_{cn})}[\|G_{X_{en} \to Y_{cn}}(y_{cn}) - y_{cn}\|] \quad (7)$$

With these three loss, we expect CycleGAN to learn a bi-directional mapping between the spectrum and prosody distributions in different languages, from source and target speakers.

B. Run-time Conversion

The conversion phase of the proposed framework is illustrated in Fig. 6. Similar to that of training phase, WORLD vocoder is used to extract spectral features, F0 and aperiodicities (APs) of source speech. We then encode 24-dimensional MCEPs spectral features, and decompose F0 into 10-scales, denoted as CWT-F0 coefficients. The 24-dimensional MCEPs and 10-dimensional CWT-F0 coefficients are converted by the trained *Spectrum-CycleGAN* and *Prosody-CycleGAN* respectively. We reconstruct F0 from the converted 10-dimensional CWT-F0 coefficients through CWT synthesis given in Fig. 6, using Eq. 4.

At run-time, we present the combined results of spectrum conversion and prosody conversion to the vocoder for reconstruction of speech waveform. Therefore, we call the proposed framework *Spectrum-Prosody-CycleGAN*. It is noted that APs are directly copied from source to that of target speaker without any modification [60].

IV. EXPERIMENTS

In this section, we conduct experiments to assess the performance of our proposed CycleGAN-based cross-lingual voice conversion framework for spectrum and prosody. We use VCC 2016 dataset [61] and Blizzard Challenge 2010 database [62],

which consists of English and Chinese speech data respectively. We choose four female English speakers (SF3, TF2, TF1, SF1), and one female Chinese speaker for experiments. We conduct cross-lingual voice conversion both from English to Chinese, denoted as $en2cn$; and from Chinese to English, denoted as $cn2en$, respectively. For $en2cn$, we conduct two experiments and choose SF3 and TF2 as the source speakers. For $cn2en$, we conduct two experiments and choose SF1 and TF1 as the source speakers, respectively. We note that bilingual data are required for MCD calculation in cross-lingual VC [8], [10], hence we only conduct subjective experiments to show the effectiveness of our proposed framework.

We build two systems for a comparative study: 1) proposed *Spectrum-Prosody-CycleGAN*, and 2) baseline *Spectrum-CycleGAN* [10], where spectrum is converted with CycleGAN and fundamental frequency (F0) is converted through the logarithm Gaussian (LG) normalized transformation [63]. In *Spectrum-Prosody-CycleGAN*, we perform CWT to decompose F0 into 10 different time scales and train two CycleGAN networks using non-parallel data with two different languages to learn the mappings of spectral and prosody features between source and target speaker.

A. Experimental Setup

To evaluate the systems under the condition of non-parallel and limited amount of data, we use 81 non-parallel utterances from source and target speakers with English and Chinese for training, and 54 utterances for evaluation. It is noted that our proposed cross-lingual VC method is trained under the disadvantageous condition (non-parallel and limited amount of data).

The speech data is downsampled to 16kHz, and 24-dimensional Mel-cepstral coefficients (MCEPs), fundamental frequency (F0), and aperiodicity (APs) are then extracted every 5 ms using WORLD vocoder [64]. For both frameworks, we extract 24-dimensional MCEPs and one-dimensional F0 features for each frame. We further obtain 10-dimensional CWT-based F0 features from one-dimensional F0 features with CWT analysis in our proposed framework. It is noted that APs are directly copied from the source to target without any modification.

In both systems, the generator uses a one-dimensional (1D) CNN to capture the relationship among the overall features while preserving the temporal structure. The 1D CNN is incorporated with down-sampling, residual,and up-sampling layers. We design the discriminator using a 2D CNN to focus on a 2D spectral texture. In the training phase, we set λ_{CYC}=10. and only use L_{ID} for the first 10^4 iterations. The Adam optimizer [65] with a batch size of 1 is used to train the networks. The initial learning rate for the generators is set to 0.0002 while that of discriminators are set to 0.0001. The learning rate keeps the same for the first 2×10^5 iterations, and linearly decays over the second 2×10^5 iterations. The momentum term β_1 is set to 0.5.

TABLE I
A COMPARISON OF THE MOS RESULTS BETWEEN BASELINE AND OUR PROPOSED METHOD FOR ENGLISH-TO-CHINESE (EN2CN) AND CHINESE-TO-ENGLISH (CN2EN).

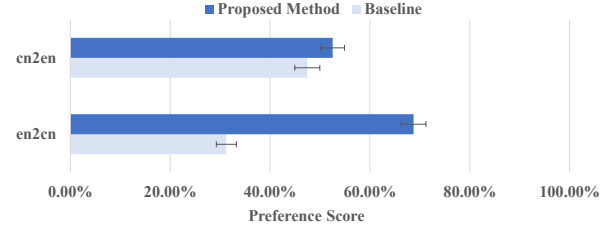| Framework	Mean Opinion Score (MOS)	
	$en2cn$	$cn2en$
Spectrum-CycleGAN & LG [10]	2.91 ± 0.33	3.30 ± 0.27
Spectrum-Prosody-CycleGAN	**2.92 ± 0.27**	**3.34 ± 0.29**

Fig. 7. The XAB preference results with 95% confidence interval between the baseline and the proposed cross-lingual framework for speaker similarity.

B. Evaluations

We conduct two listening tests to assess system performance of in terms of voice quality and speaker similarity. For each test, we conduct cross-lingual VC from English to Chinese and from Chinese to English respectively, which denoted as $en2cn$ and $cn2en$. 15 bilingual native Chinese speakers participated in all the listening tests and listened to 120 converted utterances in total.

1) Mean Opinion Score Tests: To evaluate the converted speech quality, we first conduct mean opinion score (MOS) test between the baseline and our proposed method. 15 sentences are randomly selected from the evaluation set. In MOS test, each subject are asked to score each sample in a five-point-scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). As shown in Table I, our proposed method achieves comparable results with the baseline for both $en2cn$ and $cn2en$.

2) Similarity Listening Tests: We consider that how similar the converted speech is to the target speech reflects the effect of prosody conversion. Therefore, we conduct XAB preference test between our proposed framework and the baseline framework in terms of speaker similarity. The subjects are asked to listen to the reference target samples and the converted samples of the baseline and our proposed framework, and choose the one which sounds closer to the target sample. We expect that the proposed prosody conversion approach will remarkably improve the speaker similarity, as prosody contains the information of speaking style.

As shown in Fig. 7, we observe that our proposed method clearly outperforms the baseline framework in both experiments English to Chinese ($en2cn$), and Chinese to English ($cn2en$) experiments. The results suggest that CWT is an effective way of prosody modeling for cross-lingual voice conversion, and through CycleGAN we can learn a prosody mapping between source and target that speak different languages. The results also validate our idea to use CycleGAN to learn the differentiated mapping relationship among low and

high scale coefficients of F0.

V. Conclusion

In this paper, we propose a novel parallel-data-free cross-lingual voice conversion framework. We convert the spectrum and prosody based on CycleGAN with non-parallel and limited amount of training data. Moreover, We also provide a non-linear method which uses CWT to describe the prosody in different time scales for cross-lingual voice conversion. Experimental results show the effectiveness of our proposed framework in terms of voice quality and speaker similarity.

VI. Acknowledgement

References

[1] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *arXiv preprint arXiv:2008.03648*, 2020.

[2] R. Liu, B. Sisman, J. Li, F. Bao, G. Gao, and H. Li, "Teacher-student training for robust tacotron-based tts," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[3] R. Liu, B. Sisman, F. Bao, G. Gao, and H. Li, "Wavetts: Tacotron-based tts with joint time-frequency domain loss," *arXiv preprint arXiv:2002.00417*, 2020.

[4] K. Zhou, B. Sisman, and H. Li, "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 230–237.

[5] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion," *arXiv preprint arXiv:2005.07025*, 2020.

[6] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[7] B. Sisman, M. Zhang, and H. 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.

[8] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6790–6794.

[9] M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and N. Campbell, "Cross-language voice conversion evaluation using bilingual databases," 2002.

[10] B. Sisman, M. Zhang, M. Dong, and H. Li, "On the study of generative adversarial networks for cross-lingual voice conversion," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 144–151.

[11] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *ArXiv*, vol. abs/1711.11293, 2017.

[12] M. Zhang, B. Sisman, L. Zhao, and H. Li, "Deepconversion: Voice conversion with limited parallel training data," *Speech Communication*, 06 2020.

[13] K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," *1991., IEEE International Sympoisum on Circuits and Systems*, pp. 594–597 vol.1, 1991.

[14] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pp. 655–658 vol.1, 1988.

[15] 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.

[16] B. Çişman, H. Li, and K. C. Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 677–684.

[17] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder." in *Interspeech*, 2018, pp. 1978–1982.

[18] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1859–1872, 2014.

[19] B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "Adaptive wavenet vocoder for residual compensation in gan-based voice conversion," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 282–289.

[20] 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.

[21] Y. Qian, J. Xu, and F. K. Soong, "A frame mapping based hmm approach to cross-lingual voice transformation," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5120–5123.

[22] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[23] H. Wang, F. Soong, and H. Meng, "Aa spectral space warping approach to cross-lingual voice transformation in hmm-based tts," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4874–4878.

[24] D. Erro and A. Moreno, "Frame alignment method for cross-lingual voice conversion," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[25] 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, 2009.

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

[27] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5279–5283.

[28] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. 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 preprint arXiv:1803.00860*, 2018.

[29] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[30] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional

style transfer for voice conversion with a new emotional speech dataset," 2020.

[31] H. Ming, D. Huang, M. Dong, H. Li, L. Xie, and S. Zhang, "Fundamental frequency modeling using wavelets for emotional voice conversion," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 804–809.

[32] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Gmm-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, no. 5, pp. 134–138, 2012.

[33] B. Sisman and H. Li, "Wavelet analysis of speaker dependent and independent prosody for voice conversion." in *Interspeech*, 2018, pp. 52–56.

[34] E. Helander, J. Nurminen, and M. Gabbouj, "Analysis of lsf frame selection in voice conversion," in *International conference on Speech and Computer*. Citeseer, 2007, pp. 651–656.

[35] E. Morley, E. Klabbers, J. P. v. Santen, A. Kain, and S. H. Mohammadi, "Synthetic f0 can effectively convey speaker id in delexicalized speech," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[36] M. Vainio, A. Suni, D. Aalto *et al.*, "Continuous wavelet transform for analysis of speech prosody," *TRASP 2013-Tools and Resources for the Analysys of Speech Prosody, An Interspeech 2013 satellite event, August 30, 2013, Laboratoire Parole et Language, Aix-en-Provence, France, Proceedings*, 2013.

[37] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[38] A. S. Suni, D. Aalto, T. Raitio, P. Alku, M. Vainio *et al.*, "Wavelets for intonation modeling in hmm speech synthesis," in *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*. ISCA, 2013.

[39] B. Şişman, H. Li, and K. C. Tan, "Transformation of prosody in voice conversion," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1537–1546.

[40] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, "Exemplar-based sparse representation of timbre and prosody for voice conversion," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5175–5179.

[41] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion," 2016.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[43] K. E. AK, "Deep learning approaches for attribute manipulation and text-to-image synthesis," Ph.D. dissertation, PhD thesis, 2019.

[44] K. E. Ak, J. H. Lim, J. Y. Tham, and A. A. Kassim, "Attribute manipulation generative adversarial networks for fashion images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10541–10550.

[45] K. Emir Ak, J. Hwee Lim, J. Yew Tham, and A. Kassim, "Semantically consistent hierarchical text to fashion image synthesis with an enhanced-attentional generative adversarial network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[46] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[47] K. E. Ak, N. Xu, Z. Lin, and Y. Wang, "Incorporating reinforced adversarial learning in autoregressive image generation," *arXiv preprint arXiv:2007.09923*, 2020.

[48] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[49] H. Zou, K. E. Ak, and A. A. Kassim, "Edge-gan: Edge conditioned multi-view face image generation," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 2401–2405.

[50] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[51] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 521–530.

[52] K. E. Ak, Y. Sun, and J. H. Lim, "Learning cross-modal representations for language-based image manipulation," in *Proceedings of the IEEE ICIP*, 2020, pp. 0–0.

[53] K. E. Ak, A. A. Kassim, J. Hwee Lim, and J. Yew Tham, "Learning attribute representations with localization for flexible fashion search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7708–7717.

[54] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," *arXiv preprint arXiv:1701.06547*, 2017.

[55] Z. Yang, W. Chen, F. Wang, and B. Xu, "Improving neural machine translation with conditional sequence generative adversarial nets," *arXiv preprint arXiv:1703.04887*, 2017.

[56] H. Guo, F. K. Soong, L. He, and L. Xie, "A new gan-based end-to-end tts training algorithm," *arXiv preprint arXiv:1904.04775*, 2019.

[57] Y. Zhao, S. Takaki, H.-T. Luong, J. Yamagishi, D. Saito, and N. Minematsu, "Wasserstein gan and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a wavenet vocoder," *IEEE Access*, vol. 6, pp. 60478–60488, 2018.

[58] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[59] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.

[60] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4910–4914.

[61] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016." in *Interspeech*, 2016, pp. 1632–1636.

[62] S. Kinga and V. Karaiskosb, "The blizzard challenge 2010," 2009.

[63] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin," in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, vol. 4. IEEE, 2007, pp. 410–414.

[64] M. Morise, F. Yokomori, and K. 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.

[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.