ON THE STUDY OF GENERATIVE ADVERSARIAL NETWORKS FOR CROSS-LINGUAL VOICE CONVERSION

Berrak Sisman[1,2] , *Mingyang Zhang*[1], *Minghui Dong*[2], *Haizhou Li*[1]

[1]National University of Singapore, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore

ABSTRACT

Cross-lingual voice conversion (VC) aims to convert the source speaker's voice to sound like that of the target speaker, when the source and target speakers speak different languages. In this paper, we propose to use Generative Adversarial Networks (GANs) for cross-lingual voice-conversion. We further the studies on Variational Autoencoding Wasserstein GAN (VAW-GAN) and cycle-consistent adversarial network (CycleGAN), that are known to be effective for mono-lingual voice conversion. As cross-lingual voice conversion needs to converts the voice across different phonetic system, it is more challenging than mono-lingual voice conversion. By using VAW-GAN and CycleGAN, we successfully convert the speaker identity while carrying over the source speaker's linguistic content. The proposed idea is unique in the sense that it neither relies on bilingual data and their alignment, nor any external process, such as ASR. Moreover, it works with limited amount of training data of any two languages. To our best knowledge, this is the first comprehensive study of Generative Adversarial Networks in cross-lingual voice conversion. In the experiments, we achieve high-quality converted voice, that performs equally well or better than mono-lingual voice conversion.

Index Terms— cross-lingual voice conversion, generative models, variational autoencoders, generative adversarial networks

1. INTRODUCTION

Voice conversion (VC) converts one speaker's voice to sound like that of another. It has enabled many applications such as personalized speech synthesis, spoofing attacks, and dubbing of movies.

Most of the existing VC techniques are designed for mono-lingual voice conversion, where the source and target speakers use the same language. In this paper, we will work on cross-lingual voice conversion, that changes the speaker voice from the source to target speaker, who doesn't speak the source language [1–8]. This is a more challenging task than the mono-lingual voice conversion because the source and target speakers speak in two different phonetic systems, therefore, parallel data is not available. Fig. 1 depicts the difference of training resources between cross-lingual and mono-lingual voice conversion.

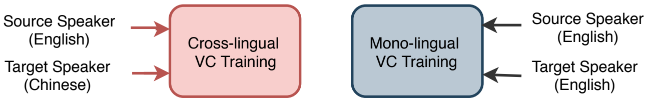

Fig. 1. Cross-lingual voice conversion is trained with source and target speech, that are in different languages, whereas mono-lingual voice conversion is trained in the same language.

The early studies of mono-lingual voice conversion relied on parallel training data to convert spectral frames from source to target speakers. The techniques include Vector Quantization (VQ) [9], codebook mapping [10], Gaussian Mixture Model (GMM) [11, 12], partial least square regression [13], dynamic kernel partial least squares regression (DKPLS) [14], and non-negative matrix factorization (NMF) [15–23].

However, parallel training data is not always possible in practice. Therefore, many have studied how to train a conversion model with non-parallel training data, such as the joint use of DBLSTM and i-vector [24], variational auto-encoder [25], and DBLSTM based Recurrent Neural Networks [5, 26]. Recently, Generative Adversarial Networks [27] such as VAW-GAN [28], CycleGAN [29–31] and StarGAN [32] eliminate the need of parallel training data, and yet achieve high quality converted voice. Generative Adversarial Networks have also been shown to be very effective in translating an image from a source domain to a target domain in the absence of parallel data [33–35], that motivates the cross-lingual voice conversion study in this paper.

The prior work on cross-lingual voice conversion includes codebook mapping [7] and GMM [8] that achieve good quality voice. However, such approaches use training data of two

Speech Samples: https://sites.google.com/view/berraksisman/

This research is supported by Programmatic grant no. A18A2b0046 from the Singapore Governments Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain) and by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-100E-2018-006).

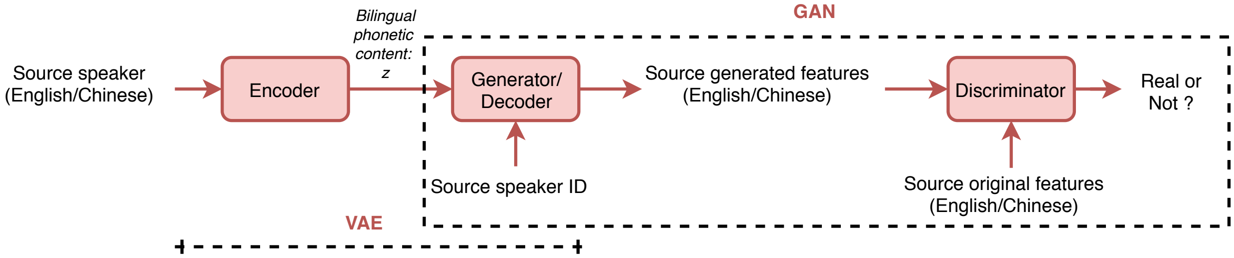
GAN
Bilingual phonetic content: z
Source speaker (English/Chinese)
Encoder
Generator/ Decoder
Source generated features (English/Chinese)
Discriminator
Real or Not ?
Source speaker ID
Source original features (English/Chinese)
VAE

Fig. 2. The training phase of VAW-GAN for cross-lingual voice conversion. Both English and Chinese speech data are used to train the encoder, generator/decoder and discriminator. The function of the encoder is similar to a phone recognizer, that learns to represent the phonetic content of two languages, while the generator/decoder behaves as a synthesizer.

languages spoken by the same bilingual speaker. As bilingual speakers are not easy to find, others have studied warping functions between phones or acoustic classes of two phonetic systems [1, 2], that don't require bilingual training data.

More recently, techniques that are based on Phonetic Posteriograms (PPGs) were reported for cross-lingual voice conversion [5, 6]. PPGs, derived from a speech recognition system, represent the posterior probability of the speech frame with respect to the phonetic classes [36, 37] that are believed to be speaker independent. Therefore, PPGs may serve as the bridge between the speakers. However, the use of PPGs has its own limitation. PPGs are language specific, furthermore, their quality also depends on the performance of speech recognition system. Another technique resorts to finding source-target frame pairs from non-parallel utterances, for instance, unit selection [38, 39] and the iterative frame alignment methods [3, 4]. However, their performances remain to be improved due to their inaccurate alignments [4].

In this paper, we propose to use Generative Adversarial Networks for cross-lingual voice conversion. With that, we eliminate the need of any external process (such as ASR), any alignment technique or bilingual training data. We focus on Wasserstein generative adversarial network (VAW-GAN) and cycle-consistent adversarial network (CycleGAN) to achieve high-quality cross-lingual voice conversion.

We note that both VAW-GAN and Cycle-GAN have been shown to achieve high-quality mono-lingual voice conversion, but haven't yet been explored for cross-lingual voice-conversion applications. VAW-GAN focuses on explaining the observations with latent variables instead of learning a pairwise transformation function, hence it doesn't require aligned parallel corpus during training. On the other hand, CycleGAN learns the forward and inverse mappings simultaneously using adversarial and cycle-consistency losses. This makes it possible to find an optimal pseudo pair, even from cross-lingual training data, that will be reported in experiments.

The main contributions of this paper are 1) we devise the generative adversarial networks for cross-lingual voice conversion; 2) we propose to use VAW-GAN and CycleGAN to eliminate the need of any external processes, or any alignment technique, that may cause degradation in voice quality; 3) we report extensive comparison of VAW-GAN and CycleGAN, that includes mono-lingual vs cross-lingual voice conversion. To our best knowledge, this paper reports the first attempt to use Generative Adversarial Networks in cross-lingual voice conversion.

The rest of the paper is organized as follows: In Section 2, we explain the Variational Autoencoders and the novel idea of using Variational Autoencoding Wasserstein Generative Adversarial Networks (VAW-GAN) for cross-lingual VC. In Section 3, we present the novel idea of using Cycle-consistent Adversarial Networks for cross-lingual VC. We provide the experimental setup, objective and subjective test results, and a discussion on GANs for cross-lingual voice conversion in Section 4. We conclude the paper in Section 5.

2. VARIATIONAL AUTOENCODING WASSERSTEIN GENERATIVE ADVERSARIAL NETWORKS (VAW-GAN)

2.1. Voice Conversion with Variational Autoencoders

It has been reported that VAEs [25] generate high-quality voice conversion with mono-lingual nonparallel training data. VAEs consist of two parts: 1) encoder, that is similar to a phone recognizer to infer the phonetic content; and 2) decoder, that operates as a synthesizer. Unfortunately, the simplicity of VAE induces inaccuracy in the synthesis model. This defect originates from the fallible assumption that the observed data is normally distributed and uncorrelated across dimensions. Such assumption gave us a defective learning objective, leading to muffled converted voices [28].

By incorporating VAEs with a GAN objective into the decoder, it was reported that mono-lingual voice conversion performance was improved [28]. GANs produce sharper spectra in general because they optimize a loss function between two distributions in a more direct fashion. An interesting attempt was to use the VAE decoder as the GAN generator to form a VAE-GAN for image generation [40]. More recently, variational autoencoding Wasserstein GAN (VAW-GAN) has been used in mono-lingual voice conversion [28]. In this paper, we will further the study of VAW-GAN towards cross-lingual

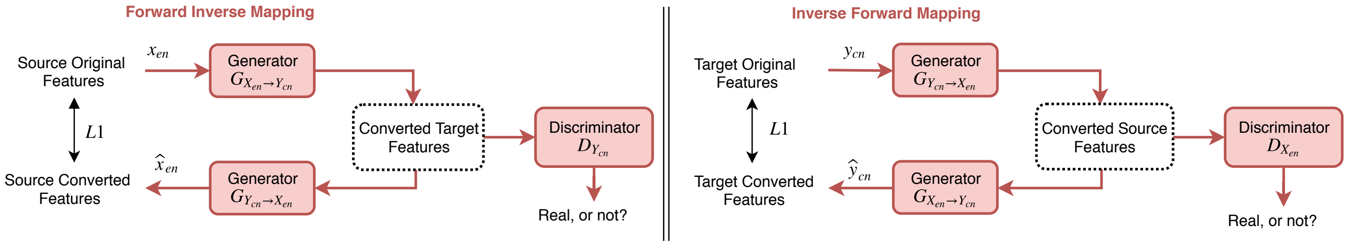
Forward Inverse Mapping
Source Original Features
x_{en}
Generator $G_{X_{en}\rightarrow Y_{cn}}$
Converted Target Features
Discriminator $D_{Y_{cn}}$
Real, or not?
$L1$
Source Converted Features
$\widehat{x}_{en}$
Generator $G_{Y_{cn}\rightarrow X_{en}}$
Inverse Forward Mapping
Target Original Features
y_{cn}
Generator $G_{Y_{cn}\rightarrow X_{en}}$
Converted Source Features
Discriminator $D_{X_{en}}$
Real, or not?
$L1$
Target Converted Features
$\widehat{y}_{cn}$
Generator $G_{X_{en}\rightarrow Y_{cn}}$

Fig. 3. Training phase of CycleGAN with cycle-consistency loss for cross-lingual voice conversion where $\hat{x}_{en}$ is equal to $G_{Y_{cn}\rightarrow X_{en}}\left(G_{X_{en}\rightarrow Y_{cn}(x_{en})}\right)$, and $\hat{y}_{cn}$ is equal to $G_{X_{en}\rightarrow Y_{cn}}\left(G_{Y_{cn}\rightarrow X_{en}(y_{cn})}\right)$.

voice conversion.

2.2. Cross-lingual Voice Conversion with VAW-GAN

We believe that the encoder-decoder structure of VAW-GAN can allow for the learning of the mapping between the phonetic systems of two different languages, through a bilingual phonetic content.

With the encoder, speech frames that belong to the similar or same phoneme class can hinge on a similar *phonetic content*, that is denoted as z. The proposed *phonetic content* is bilingual in the sense that it is exposed to two languages during training, therefore, serves as the bridge between two languages. The *phonetic content* z of VAW-GAN also plays a role as the bridge between the speakers, that is similar to PPG in mono-lingual voice conversion [20, 26] and TTS [5]. We believe that, by training the VAE with two languages, we can build a better bridge between the speakers, without the need of PPGs.

The implementation of VAW-GAN is unique in cross-lingual voice conversion. We train the same encoder-decoder structure for bidirectional conversion of two languages by taking the speech of both languages as the input. The same system works for both English and Chinese inputs.

As the encoder is exposed to input frames in both languages during training, it transforms the input into the phonetic content z that is not only speaker independent, but also bilingual. The decoder of VAW-GAN is conditioned on the speaker identity input to generate target speech. We use a one-hot vector to represent the speaker identity during training. The training phase of the entire process is summarized in Figure 2.

Overall, the proposed VAW-GAN consists of three components, that are 1) encoder, 2) generator/decoder, and 3) discriminator. While the generator/decoder minimizes the loss, the discriminator maximizes it. The encoder structure of VAW-GAN helps us to find some phonetic clusters that represent both languages. In addition, the generative and discriminative structure of GAN helps us to achieve high quality cross-lingual voice conversion. The proposed idea is more attractive than the state-of-the-art in cross-lingual voice conversion [1, 2, 6–8, 26], as it eliminates the need of any external processes, such as ASR, any alignment technique, and the need of bilingual training data.

3. CYCLE-CONSISTENT ADVERSARIAL NETWORKS FOR CROSS-LINGUAL VC

3.1. Cycle-Consistent Adversarial Networks

Cycle Consistent Adversarial Networks have been successfully used in many applications, such as image-to-image translation [33–35]. Image-to-image translation is to learn the mapping between an input image and an output image using a training set of aligned image pairs. We believe that image-to-image translation and cross-lingual voice conversion face a similar challenge that is to find a mapping from a source domain to a target domain without the need of parallel training data. As reported in [33], CycleGAN is known to achieve remarkable results on several tasks where paired training data does not exist, including collection style transfer, object transfiguration, season transfer, etc.

We consider that the task of converting the speaking voice from source to target while preserving the linguistic content is similar to that of translating an image from horse to zebra, while preserving the structure of horse and changing the color of horse to that of zebra [33], both in the absence of parallel training examples.

3.2. Cross-lingual VC with CycleGAN

In cross-lingual voice conversion, we need to preserve linguistic content while capturing the sequential and hierarchical structures through a bi-directional mapping, that motivates the use of CycleGAN. Similar to VAW-GAN, we would like to eliminate the need of any external process, such as ASR, and the need of bilingual training data.

A CycleGAN learns forward and inverse mappings simultaneously using adversarial and cycle-consistency losses. This makes it possible to find an optimal pseudo pair from unpaired cross-lingual data. Furthermore, the adversarial loss contributes to reducing over-smoothing of the converted feature sequence. We configure a CycleGAN with gated CNNs and train it with an identity-mapping loss. This allows us to preserve the linguistic content of the source speaker.

In short, a cross-lingual mapping can be learned using three loss functions that are adversarial loss, cycle-consistency loss and identity-mapping loss. We next describe their role in cross-lingual voice conversion. In all equations, source is assumed to be an English speaker, denoted as *en* and target is assumed to be a Chinese speaker, denoted as *cn*. Our goal is to learn a mapping from source $x_{en} \in X_{en}$ to target $y_{cn} \in Y_{cn}$ without relying on parallel data.

3.2.1. *Adversarial loss:*

In cross-lingual voice conversion, we optimize the distribution of the converted data as close as possible to the distribution of target data, that is from a different language. The objective function can be written as follows:

$$\begin{aligned}&\mathcal{L}_{adv}(G_{X_{en}\rightarrow Y_{cn}}, D_{Y_{cn}}) = E_{y_{cn}\sim P_{Data(y_{cn})}}\left[\log D_{Y_{cn(y_{cn})}}\right]\\&+E_{x_{en}\sim P_{Data(x_{en})}}\left[\log\left(1-D_{Y_{cn}}\left(G_{X_{en}\rightarrow Y_{cn(x_{en})}}\right)\right)\right]\end{aligned} \tag{1}$$

The closer the distribution of converted data becomes to that of target data, the smaller the loss (Eq. (1)) becomes, thus, higher similarity of the output voice to the target speaker.

3.2.2. *Cycle-consistency loss:*

The adversarial loss only tells us whether $G_{X_{en}\rightarrow Y_{cn}}$ follows the target-data distribution and does not help preserve the contextual information of x_{en}. In mono-lingual voice conversion, CycleGAN [29, 30] introduces two additional terms, that are the adversarial loss $\mathcal{L}_{adv}(G_{Y_{cn}\rightarrow X_{en}}, D_{X_{en}})$ for inverse mapping $G_{Y_{cn}\rightarrow X_{en}}$, and the cycle-consistency loss. The training phase of the proposed approach for cross-lingual voice conversion that uses CycleGAN with cycle-consistency loss is given in Figure 3. We define the cycle-consistency loss as follows:

$$\begin{aligned}&\mathcal{L}_{cyc}\left(G_{X_{en}\rightarrow Y_{cn}}, G_{Y_{cn}\rightarrow X_{en}}\right)\\&=E_{x_{en}\sim P_{Data(x_{en})}}\left[||G_{Y_{cn}\rightarrow X_{en}}\left(G_{X_{en}\rightarrow Y_{cn(x_{en})}}\right)-x_{en}||_1\right]\\&+E_{y_{cn}\sim P_{Data(y_{cn})}}\left[||G_{X_{en}\rightarrow Y_{cn}}\left(G_{Y_{cn}\rightarrow X_{en(y_{cn})}}\right)-y_{cn}||_1\right]\end{aligned} \tag{2}$$

These additional terms $G_{X_{en}\rightarrow Y_{cn}}$ and $G_{Y_{cn}\rightarrow X_{en}}$ encourage (x_{en}, y_{cn}) pairs with similar (or even same) contextual information, therefore, establishing the phonetic mapping between two languages. This is very important in cross-lingual voice conversion as the phonetic systems of source and target languages are different, and finding such pairs can be challenging.

3.2.3. *Identity-mapping loss:*

A cycle-consistency loss provides constraints on a structure; however, it would not suffice to guarantee that the mappings always preserve linguistic content. To explicitly preserve the linguistic content without relying on external processes, such as ASR, we incorporate an identity-mapping loss.

$$\begin{aligned}&\mathcal{L}_{id}\left(G_{X_{en}\rightarrow Y_{cn}}, G_{Y_{cn}\rightarrow X_{en}}\right)\\&=E_{x_{en}\sim P_{Data(x_{en})}}\left[||G_{Y_{cn}\rightarrow X_{en}}(x_{en})-x_{en}||\right]\\&+E_{y_{cn}\sim P_{Data(y_{cn})}}\left[||G_{X_{en}\rightarrow Y_{cn}}(y_{cn})-y_{cn}||\right]\end{aligned} \tag{3}$$

The studies on CycleGAN [29, 30, 33] have showed the effectiveness of this loss for color preservation in image-to-image translation and linguistic content preservation in mono-lingual voice conversion. Hence, we have good reason to expect that it preserves the rendering of the language identity.

4. EXPERIMENTS

We conduct both objective and subjective experiments to assess the performance of our proposed cross-lingual voice conversion approaches. We use CMU database [41,42], that consists of English data, and Blizzard Challenge 2010 database [43], that consists of Mandarin Chinese data. We train both VAW-GAN and CycleGAN with nonparallel data in two different languages to learn the spectral mapping between two speakers of different languages. For fundamental frequency (F0), we perform the traditional linear conversion by normalizing the mean and variance of the source speech to those of target [17]. We use VAW-GAN [28] and CycleGAN [29] in mono-lingual voice conversion as the reference baselines because they render high quality voice.

4.1. Experimental Setup

In VAW-GAN, the input of the encoder was 513-dimension spectral envelope. The encoder was a 5-layer 1D Convolutional Neural Networks (CNN) with a kernel size of 7 and a stride of 3 followed by a fully connected layer, and the output channels were $\{16, 32, 64, 128, 256\}$. The dimension of the latent vector space was set to 64, and the dimension of the speaker embedding was set to 10. Then these two vectors were merged to a 171-dimension vector by fully connection layer. For GAN, the generator was a 4-layer 1D Convolutional Neural Networks (CNN) with kernel sizes of $\{9, 7, 7, 1025\}$ and strides of $\{3, 3, 3, 1\}$, and the output channels were $\{32, 16, 8, 1\}$. The target of the generator was also the 513-dimension spectral envelope. The discriminator was a 3-layer 1D Convolutional Neural Networks (CNN) with kernel sizes of $\{7, 7, 115\}$ and a stride of 3 followed by a fully connected layer, and the output channels were $\{16, 32, 64\}$. The network was trained by using RMSProp with a learning rate of 1e-5. The batch size was set to 256 and ran for 45 epochs. We note that we train only one network that can perform both English to Chinese and Chinese to English conversion, through the bilingual phonetic content z.

In CycleGAN, we designed the generator using a one-dimensional (1D) CNN to capture the relationship among the

Framework	Language Information	Training Data	MCD [source]	MCD [target]
Mono-lingual VAW-GAN [28]	en-en	50-50 utt. 200-200 utt.	6.89 7.13	6.28 5.88
Mono-lingual CycleGAN [29]	en-en	50-50 utt. 200-200 utt.	7.09 7.26	5.98 5.67
Cross-lingual VAW-GAN *(proposed)*	en-cn and cn-en	50-50 utt. 200-200 utt.	6.65 6.75	NA NA
Cross-lingual CycleGAN *(proposed)*	en-cn and cn-en	50-50 utt. 200-200 utt.	6.81 6.91	NA NA

Table 1. A comparison between VAW-GAN and CycleGAN for mono-lingual (en-en) and cross-lingual (en-cn, cn-en) voice conversion. All experiments are conducted with non-parallel training data. Note that it is desired to have lower $MCD[target]$ and higher $MCD[source]$.

overall features while preserving the temporal structure. The target of the generator was also the 513-dimension spectral envelope. We designed the discriminator using a 2D CNN to focus on a 2D spectral texture. As a pre-process, we normalized the source and target MCEPs per dimension. We set λ_{cyc} =10 and λ_{id} =5. We trained the network using the Adam optimizer with a batch size of 1. We empirically set the initial learning rates to 0.0002 for the generator and 0.0001 for the discriminator.

4.2. Objective Evaluation

As an objective evaluation metric, we use the Mel-cepstral distortion (MCD) [44]. For mono-lingual voice conversion, we report the MCD between the converted spectrum and original source spectrum, denoted as $MCD[source]$, and MCD between converted spectrum and the original target spectrum, denoted as $MCD[target]$. However, in cross-lingual voice conversion, it is not feasible to calculate $MCD[target]$ as we do not have target speaker's voice in source language. Therefore, we only calculate $MCD[source]$ to show the effect. The MCD between two frames are calculated as follows,

$$MCD = \frac{10}{\log 10}\sqrt{2\sum_{m=1}^{24}(c(m) - c_{cv}(m))^2} \qquad (4)$$

where $c_{cv}(m)$ and $c(m)$ are the m^{th} coefficients of the converted MCCs, and the comparing MCCs, respectively. We extracted 24 Mel-cepstral coefficients (MCEPs) to calculate the MCD values frame-by-frame over all the paired frames. We note that it is desired to have lower $MCD[target]$ and higher $MCD[source]$.

We report the $MCD[source]$ and $MCD[target]$ values of 8 different settings in Table 1. Firstly, we observe that both VAW-GAN and CycleGAN can achieve good performance with limited training data. Secondly, we note that in cross-lingual voice conversion, CycleGAN outperforms VAW-GAN in all cases by achieving higher $MCD[source]$. Last but not least, the proposed cross-lingual CycleGAN framework, that uses 200 utterances, outperforms the mono-lingual VAW-GAN framework, that uses 50 utterances.

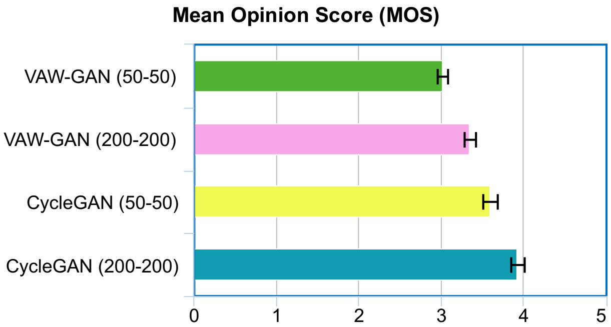

Fig. 4. Mean opinion scores with 95% confidence intervals for CycleGAN and VAW-GAN with 50-50 and 200-200 non-parallel training data.

We would like to highlight that evaluating cross-lingual voice conversion with $MCD[target]$ is not straightforward because the target speaker doesn't speak the source language [6]. Therefore, we can only use $MCD[source]$ as an indirect indicator to evaluate the converted voice quality. We next conduct extensive subjective tests to confirm our findings.

4.3. Subjective Evaluation

We conduct four listening experiments to assess the performance of generative adversarial networks for cross-lingual voice conversion in terms of voice quality and speaker similarity. 15 English speakers and 10 Chinese speakers participated in the listening tests. Each subject listens to 50 converted utterances of his/her speaking language. The p-values are calculated in a similar way that is reported in [46].

We first evaluate the sound quality of the converted voices with mean opinion score (MOS) between VAW-GAN and CycleGAN cross-lingual voice conversion systems, that is reported in Figure 4. The listeners rate the quality of the converted voice using a 5-point scale: 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad. It is observed that CycleGAN clearly outperforms VAW-GAN counterpart with the same amount of training data. It is worth mentioning that CycleGAN trained on 50 utterances even outperforms VAW-GAN trained on 200 training utterances.

We next conduct preference test, that is reported in Figure 5, to compare CycleGAN with VAW-GAN for cross-lingual

Framework	Language Information	Training data	Best (%)	Worst (%)	Not Preferred (%)
Mono-lingual VAW-GAN [28]	en-en	50-50 utt.	18	45	37
Cross-lingual CycleGAN	en-cn and cn-en	50-50 utt.	9	55	36
		200-200 utt.	73	0	27

Table 2. Voice quality assessment using Best-Worst percentages on an aggregate level [45]. As we do not have bilingual data from any speaker, this experiment does not consider speaker similarity, and focus only on voice quality.

Mono-lingual CycleGAN	Cross-lingual CycleGAN
$(53.6 \pm 2.2)\,\%$	$(46.4 \pm 2.7)\,\%$

Table 3. The preference test in terms of voice quality and speaker similarity between mono-lingual and cross-lingual CycleGAN. 200 nonparallel data have been used to train the networks. The p-value [46] is $2.062e^{-17}$.

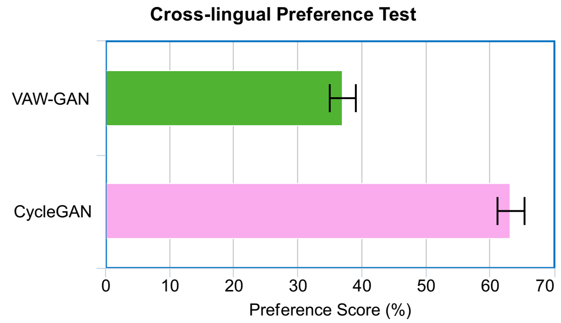

Fig. 5. The speaker similarity preference test between the proposed CycleGAN and VAW-GAN for cross-lingual voice conversion. 200 nonparallel cross-lingual data have been used to train the networks. The p-value [46] is $2.178e^{-21}$.

voice conversion, in terms of speaker similarity. It is observed that CycleGAN outperforms VAW-GAN also in similarity test.

Note that CycleGAN consistently outperforms VAW-GAN in cross-lingual voice conversion, we are now interested in comparing cross-lingual CycleGAN results with its mono-lingual VAW-GAN counterparts. We choose the mono-lingual VAW-GAN (en-en, 50-50 utt) that we used in objective test as the reference baseline. As Best-Worst Scaling (BWS) [45] method is effective to provide ranking of a long list of listening samples, we conduct BWS experiment over the converted speech from three systems, the mono-lingual VAW-GAN (en-en, 50-50 utt) reference baseline, the cross-lingual Cycle-GANs (50-50 utt) and cross-lingual Cycle-GANs (200-200 utt).

As shown in Table 2, the CycleGAN system for cross-lingual voice conversion (200-200 utt) is chosen as the best for 73.0 % of the time and never chosen as the worst. It clearly outperforms the mono-lingual VAW-GAN reference baseline. Moreover, it is shown that CycleGAN can also work with limited data. Even with 50 cross-lingual utterances, CycleGAN achieves comparable results with mono-lingual VAW-GAN that is trained on 50 utterances. This experiment further validates ability of CycleGAN in cross-lingual voice conversion.

We further choose the mono-lingual CycleGAN as a benchmarking reference because CycleGAN [29] was reported to show remarkable mono-lingual voice conversion quality. We conduct a preference test that is reported in Table 3, to compare mono-lingual CycleGAN with cross-lingual CycleGAN, both of which are trained on 200-200 training utterances. The results show that listeners find the two systems are comparable. While the mono-lingual CycleGAN is slightly better than cross-lingual one, we have not forgotten that cross-lingual conversion is a more challenging task.

4.4. Discussion on the Effectiveness of GANs in Cross-lingual Voice Conversion

This paper shows that the generative adversarial networks can perform high-quality cross-lingual voice conversion even with limited data. We note that CycleGAN and VAW-GAN shares a similar motivation regarding the use of adversarial loss function. We have benchmarked the proposed cross-lingual CycleGAN against different reference baselines to show its advantage. It is observed that CycleGAN consistently achieves better results than the VAW-GAN. The results suggest that the cycle-consistency and the identity-mapping losses of CycleGAN allows the network to optimize the phonetic mapping between two languages, that has an advantage over VAW-GAN, that depends on the bilingual *phonetic content z* to find this mapping.

We have also obtained some good preliminary results with the proposed GAN frameworks from bidirectional mapping to many-to-many mappings across different language domains, as a step towards *multilingual voice conversion across speakers and languages*. We will report the findings in the near future.

5. CONCLUSION

We have studied the generative adversarial networks to perform cross-lingual voice conversion, that is more challenging than mono-lingual voice conversion. We propose to use VAW-GAN and CycleGAN to learn a mapping between two speakers, who speak different languages. The proposed approaches produce high quality converted voice without the need of any bilingual data, alignment, or external processes (such as ASR). Moreover, they perform well with very limited training data. We have benchmarked the proposed cross-lingual GAN systems with the state-of-the-art mono-lingual voice conversion implementation, that show comparable results with the same amount of training data.

6. REFERENCES

[1] David Sundermann, Hermann Ney, and H Hoge, "Vtln-based crosslanguage voice conversion," *IEEE ASRU*, 2003.

[2] Yao Qian, Ji Xu, and Frank K Soong, "A frame mapping based hmm approach to cross-lingual voice transformation," *ICASSP*, 2011.

[3] Daniel Erro and Asuncion Moreno, "Frame alignment method for cross-lingual voice conversion," *INTERSPEECH*, 1972.

[4] Daniel Erro, Asuncin Moreno, and Antonio Bonafonte, "Inca algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, 2015.

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

[6] Yi Zhou, Xiaohai Tian, Haihua Xu, Rohan Kumar Das, and Haizhou Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgrams and average modeling," *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2019.

[7] 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*, 1991.

[8] Mikiko Mashimo, Tomoki Toda, Hiromichi Kawanami, Kiyohiro Shikano, and Nick Campbell, "Cross-language voice conversion evaluation using bilingual databases," *IPSJ Journal*, 2002.

[9] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 655–658, 1988.

[10] Kiyohiro Shikano, Satoshi Nakamura, and Masanobu Abe, "Speaker Adaptation and Voice Conversion by Codebook Mapping," *IEEE International Sympoisum on Circuits and Systems*, pp. 594–597, 1991.

[11] 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.

[12] Kazuhiro Kobayashi, Shinnosuke Takamichi, Satoshi Nakamura, and Tomoki Toda, "The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016," *in INTERSPEECH*, 2016.

[13] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[14] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[15] Berrak Şişman, Haizhou Li, and Kay Chen 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.

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

[17] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[18] Ryo Aihara, Kenta Masaka, Tetsuya Takiguchi, and Yasuo Ariki, "Parallel dictionary learning for multimodal voice conversion using matrix factorization," *In INTERSPEECH*, pp. 27–40, 2016.

[19] Zeyu Jin, Adam Finkelstein, Stephen Di Verdi, Jingwan Lu, and Gautham J Mysore, "Cute: a concatenative method for voice conversion using exemplar- based unit selection," *In ICASSP*, 2016.

[20] Berrak Çişman, Haizhou Li, and Kay Chen 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.

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

[22] 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*, 2019.

[23] Berrak Sisman and Haizhou Li, “Wavelet analysis of speaker dependent and independent prosody for voice conversion,” *INTERSPEECH*, 2018.

[24] Jie Wu, Zhizheng Wu, and Lei Xie, “On the use of I-vectors and average voice model for voice conversion without parallel data,” *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[25] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, “Voice conversion from non-parallel corpora using variational auto-encoder,” *APSIPA ASC*, 2016.

[26] 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.

[27] Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi 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.

[28] 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,” *arXiv:1704.00849 [cs.CL]*, 2017.

[29] Takuhiro Kaneko and Hirokazu Kameoka, “Parallel-data-free voice conversion using cycle-consistent adversarial networks,” *arXiv*, 2017.

[30] Fuming Fang, Junichi Yamagishi, Echizen I, and Jaime Lorenzo-Trueba, “High-Quality Nonparallel Voice Conversion Based on Cycle-Consistent Adversarial Network,” *IEEE ICASSP*, 2018.

[31] Jaime Lorenzo-Trueba, Fuming Fang, Xin Wang, Isao Echizen, Junichi Yamagishi, and Tomi Kinnunen, “Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama’s voice using GAN, WaveNet and low-quality found data,” *arXiv:1803.00860 [eess.AS]*, 2018.

[32] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, “StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks,” *arXiv:1806.02169 [cs.SD]*, 2018.

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” *ICCV*, 2017.

[34] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, “Multimodal Unsupervised Image-to-Image Translation,” *ECCV*, 2018.

[35] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman, “Toward Multimodal Image-to-Image Translation,” *NIPS*, 2017.

[36] Timothy J Hazen, Wade Shen, and Christopher White, “Query-by-example spoken term detection using phonetic posteriorgram templates,” *In IEEE ASRU*, pp. 421–426, 2009.

[37] Keith Kintzley, Aren Jansen, and Hynek Hermansky, “Event selection from phone posteriorgrams using matched filters,” *In INTERSPEECH*, pp. 1905–1908, 2011.

[38] David Sundermann, Harald Hoge, Antonio Bonafonte, Hermann Ney, Alan Black, and Shri Narayanan, “Text-independent voice conversion based on unit selection,” *IEEE ICASSP*, 2006.

[39] Hao Wang, Frank Soong, and Helen Meng, “A spectral space warping approach to cross-lingual voice transformation in hmm-based tts,” *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2015.

[40] Larsen A B L, S K Snderby, H Larochelle, and O Winther, “Autoencoding beyond pixels using a learned similarity metric,” *JMLR Workshop and Conference Proceedings*, 2016.

[41] Alan W Black and John Kominek, “CMU ARCTIC databases for speech synthesis,” *Carnegie Mellon University*, 2003.

[42] Alan W Black and John Kominek, “The CMU Arctic Speech Databases,” *Fifth ISCA ITRW on Speech Synthesis*, 2004.

[43] Simon King and Vasilis Karaiskos, “The Blizzard Challenge 2010,” *CSTR, University of Edinburgh, UK*, 2010.

[44] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” *Communications, Computers and Signal Processing*, pp. 125–128, 1993.

[45] Terry N. Flynn and Anthony A. J. Marley, “Best worst scaling: Theory and methods,” *Handbook of choice modelling, Edward Elgar Publishing*, pp. 178–201, 2014.

[46] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” *2016 IEEE ICME*, 2016.