# CYCLEGAN-VC2: IMPROVED CYCLEGAN-BASED NON-PARALLEL VOICE CONVERSION

*Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, Nobukatsu Hojo*

NTT Communication Science Laboratories, NTT Corporation, Japan

## ABSTRACT

Non-parallel voice conversion (VC) is a technique for learning the mapping from source to target speech without relying on parallel data. This is an important task, but it has been challenging due to the disadvantages of the training conditions. Recently, CycleGAN-VC has provided a breakthrough and performed comparably to a parallel VC method without relying on any extra data, modules, or time alignment procedures. However, there is still a large gap between the real target and converted speech, and bridging this gap remains a challenge. To reduce the gap, we propose CycleGAN-VC2, which is an improved version of CycleGAN-VC incorporating three new techniques: an improved objective (two-step adversarial losses), improved generator (2-1-2D CNN), and improved discriminator (PatchGAN). We evaluated our method on a non-parallel VC task and analyzed the effect of each technique in detail. An objective evaluation showed that these techniques help bring the converted feature sequence closer to the target in terms of both global and local structures, which we assess by using Mel-cepstral distortion and modulation spectra distance, respectively. A subjective evaluation showed that CycleGAN-VC2 outperforms CycleGAN-VC in terms of naturalness and similarity for every speaker pair, including intra-gender and inter-gender pairs.[1]

***Index Terms—*** Voice conversion (VC), non-parallel VC, generative adversarial networks (GANs), CycleGAN, CycleGAN-VC

## 1. INTRODUCTION

Voice conversion (VC) is a technique for transforming the non/para-linguistic information of given speech while preserving the linguistic information. VC has great potential for application to various tasks, such as speaking aids [1, 2] and the conversion of style [3, 4] and pronunciation [5].

One successful approach to VC is statistical, and many methods in this vein have been proposed: Gaussian mixture model (GMM)-based methods [6, 7, 8], neural network (NN)-based methods (including restricted Boltzmann machines (RBMs) [9, 10], feed forward NNs [11, 12, 13], recurrent NNs (RNNs) [14, 15], convolutional NNs (CNNs) [5], and generative adversarial networks (GANs) [5]), and exemplar-based methods using non-negative matrix factorization (NMF) [16, 17].

Many VC methods (including the above-mentioned) are categorized as parallel VC, which relies on the availability of parallel utterance pairs of the source and target speakers. However, collecting such data is often laborious or impractical. Even if obtaining such data is feasible, many VC methods require a time alignment procedure as a pre-process, which may occasionally fail and requires careful pre-screening or manual correction. To overcome these restrictions, this paper focuses on non-parallel VC, which does not rely on parallel utterances, transcriptions, or time alignment procedures.

In general, non-parallel VC is quite challenging and is inferior to parallel VC in terms of quality due to the disadvantages of the training conditions. To alleviate these severe conditions, several studies have incorporated an extra module (e.g., an automatic speech recognition (ASR) module [18, 19]) or extra data (e.g., parallel utterance pairs among reference speakers [20, 21, 22, 23]). Although these additional modules or data are helpful for training, preparing them imposes other costs and thus limits application. To avoid such additional costs, recent studies have examined the use of stochastic NNs (e.g., an RBN [24] and variational autoencoders (VAEs) [25, 26]), which embed the acoustic features into common low-dimensional space with the supervision of speaker identification. It is noteworthy that they are free from extra data, modules, and time alignment procedures. However, one limitation is that they need to approximate data distribution explicitly (e.g., Gaussian is typically used), which tends to cause over-smoothing through statistical averaging.

To overcome these limitations, recent studies [25, 27, 28] have incorporated GANs [29], which can learn a generative distribution close to the target without explicit approximation, thus avoiding the over-smoothing caused by statistical averaging. Among these, in contrast to some of the frame-by-frame methods [25, 28], which have difficulty in learning time dependencies, CycleGAN-VC [27] (published in [30]) makes it possible to learn a sequence-based mapping function by using CycleGAN [31, 32, 33] with a gated CNN [34] and identity-mapping loss [35]. This allows sequential and hierarchical structures to be captured while preserving linguistic information. With this improvement, CycleGAN-VC performed comparably to a parallel VC method [7].

However, even using CycleGAN-VC, there is still a challenging gap to bridge between the real target and converted speech. To reduce the gap, we propose CycleGAN-VC2, which is an improved version of CycleGAN-VC incorporating three new techniques: an improved objective (two-step adversarial losses), improved generator (2-1-2D CNN), and improved discriminator (PatchGAN). We analyzed the effect

[1] The converted speech samples are provided at http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc2/index.html.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

of each technique on the Spoke (i.e., non-parallel VC) task of the Voice Conversion Challenge 2018 (VCC 2018) [36]. An objective evaluation showed that the proposed techniques help bring the converted acoustic feature sequence closer to the target in terms of global and local structures, which we assess by using Mel-cepstral distortion and modulation spectra distance, respectively. A subjective evaluation showed that CycleGAN-VC2 outperforms CycleGAN-VC in terms of naturalness and similarity for every speaker pair, including intra-gender and inter-gender pairs.

In Section 2 of this paper, we review the conventional CycleGAN-VC. In Section 3, we describe CycleGAN-VC2, which is an improved version of CycleGAN-VC incorporating three new techniques. In Section 4, we report the experimental results. We conclude in Section 5 with a brief summary and mention of future work.

## 2. CONVENTIONAL CYCLEGAN-VC

### 2.1. Objective: One-Step Adversarial Loss

Let $x \in \mathbb{R}^{Q \times T_x}$ and $y \in \mathbb{R}^{Q \times T_y}$ be acoustic feature sequences belonging to source $X$ and target $Y$, respectively, where $Q$ is the feature dimension and $T_x$ and $T_y$ are the sequence lengths. The goal of CycleGAN-VC is to learn mapping $G_{X \rightarrow Y}$, which converts $x \in X$ into $y \in Y$, without relying on parallel data. Inspired by CycleGAN [31], which was originally proposed in computer vision for unpaired image-to-image translation, CycleGAN-VC uses an adversarial loss [29] and cycle-consistency loss [37]. Additionally, to encourage the preservation of linguistic information, CycleGAN-VC also uses an identity-mapping loss [35].

**Adversarial loss:** To make a converted feature $G_{X \rightarrow Y}(x)$ indistinguishable from a target $y$, an adversarial loss is used:

$$\mathcal{L}_{adv}(G_{X \rightarrow Y}, D_Y) = \mathbb{E}_{y \sim P_Y(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim P_X(x)}[\log(1 - D_Y(G_{X \rightarrow Y}(x)))], \quad (1)$$

where discriminator $D_Y$ attempts to find the best decision boundary between real and converted features by maximizing this loss, and $G_{X \rightarrow Y}$ attempts to generate a feature that can deceive $D_Y$ by minimizing this loss.

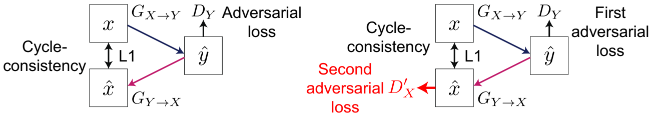
**Cycle-consistency loss:** The adversarial loss only restricts $G_{X \rightarrow Y}(x)$ to follow the target distribution and does not guarantee the linguistic consistency between input and output features. To further regularize the mapping, a cycle-consistency loss is used:

$$\begin{aligned} &\mathcal{L}_{cyc}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) \\ &= \mathbb{E}_{x \sim P_X(x)}[\|G_{Y \rightarrow X}(G_{X \rightarrow Y}(x)) - x\|_1] \\ &+ \mathbb{E}_{y \sim P_Y(y)}[\|G_{X \rightarrow Y}(G_{Y \rightarrow X}(y)) - y\|_1], \end{aligned} \quad (2)$$

where forward-inverse and inverse-forward mappings are simultaneously learned to stabilize training. This loss encourages $G_{X \rightarrow Y}$ and $G_{Y \rightarrow X}$ to find an optimal pseudo pair of $(x, y)$ through circular conversion, as shown in Fig. 1(a).

**Identity-mapping loss:** To further encourage the input preservation, an identity-mapping loss is used:

$$\mathcal{L}_{id}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) = \mathbb{E}_{y \sim P_Y(y)}[\|G_{X \rightarrow Y}(y) - y\|_1] + \mathbb{E}_{x \sim P_X(x)}[\|G_{Y \rightarrow X}(x) - x\|_1]. \quad (3)$$



**(a) One-step adversarial loss** **(b) Two-step adversarial losses (proposed)**

**Fig. 1**. Comparison of objectives

**Full objective:** The full objective is written as

$$\begin{aligned} \mathcal{L}_{full} &= \mathcal{L}_{adv}(G_{X \rightarrow Y}, D_Y) + \mathcal{L}_{adv}(G_{Y \rightarrow X}, D_X) \\ &+ \lambda_{cyc}\mathcal{L}_{cyc}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) + \lambda_{id}\mathcal{L}_{id}(G_{X \rightarrow Y}, G_{Y \rightarrow X}), \end{aligned} \quad (4)$$

where $\lambda_{cyc}$ and $\lambda_{id}$ are trade-off parameters. In this formulation, an adversarial loss is used once for each cycle, as shown in Fig. 1(a). Hence, we call it a *one-step adversarial loss*.

### 2.2. Generator: 1D CNN

CycleGAN-VC uses a *one-dimensional (1D) CNN* [5] for the generator to capture the overall relationship along with the feature direction while preserving the temporal structure. This can be viewed as the direct temporal extension of a frame-by-frame model that captures such features' relationship only per frame. To capture the wide-range temporal structure efficiently while preserving the input structure, the generator is composed of downsampling, residual [38], and upsampling layers, as shown in Fig. 2(a). The other notable point is that CycleGAN-VC uses a gated CNN [34] to capture the sequential and hierarchical structures of acoustic features.

### 2.3. Discriminator: FullGAN

CycleGAN-VC uses a 2D CNN [5] for the discriminator to focus on a 2D structure (i.e., 2D spectral texture [39]). More precisely, as shown in Fig. 3(a), a fully connected layer is used at the last layer to determine the realness considering the input's overall structure. Such a model is called *FullGAN*.

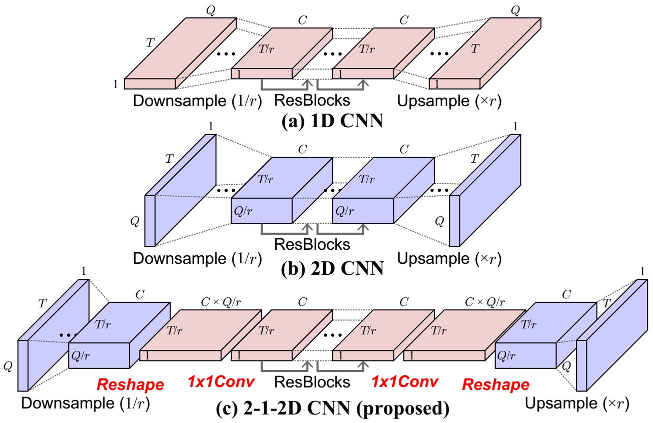
## 3. CYCLEGAN-VC2

### 3.1. Improved Objective: Two-Step Adversarial Losses

One well-known problem for statistical models is the oversmoothing caused by statistical averaging. The adversarial loss used in Eq. 4 helps to alleviate this degradation, but the cycle-consistency loss formulated as L1 still causes oversmoothing. To mitigate this negative effect, we introduce an additional discriminator $D'_X$ and impose an adversarial loss on the circularly converted feature, as

$$\mathcal{L}_{adv2}(G_{X \rightarrow Y}, G_{Y \rightarrow X}, D'_X) = \mathbb{E}_{x \sim P_X(x)}[\log D'_X(x)] + \mathbb{E}_{x \sim P_X(x)}[\log(1 - D'_X(G_{Y \rightarrow X}(G_{X \rightarrow Y}(x))))]. \quad (5)$$

Similarly, we introduce $D'_Y$ and impose an adversarial loss $\mathcal{L}_{adv2}(G_{Y \rightarrow X}, G_{X \rightarrow Y}, D'_Y)$ for the inverse-forward mapping. We add these two adversarial losses to Eq. 4. In this improved objective, we use adversarial losses twice for each cycle, as shown in Fig. 1(b). Hence, we call them *two-step adversarial losses*.

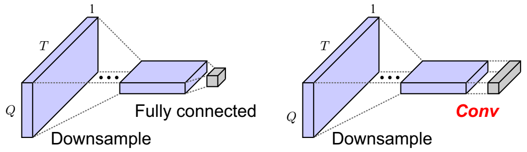


(a) 1D CNN

(b) 2D CNN

(c) 2-1-2D CNN (proposed)

**Fig. 2**. Comparison of generator network architectures. Red and blue blocks indicate 1D and 2D convolution layers, respectively. $r$ indicates a downsampling or upsampling rate.



(a) FullGAN

(b) PatchGAN (proposed)

**Fig. 3**. Comparison of discriminator network architectures

### 3.2. Improved Generator: 2-1-2D CNN

In a VC framework [5, 27] (including CycleGAN-VC), a 1D CNN (Fig. 2(a)) is commonly used as a generator, whereas in a postfilter framework [39, 40], a 2D CNN (Fig. 2(b)) is more preferred. These choices are related to the pros and cons of each network. A 1D CNN is more feasible for capturing dynamical change, as it can capture the overall relationship along with the feature dimension. In contrast, a 2D CNN is better suited for converting features while preserving the original structures, as it restricts the converted region to local. Even using a 1D CNN, residual blocks [38] can mitigate the loss of the original structure, but we find that downsampling and upsampling (which are necessary for effectively capturing the wide-range structures) become a severe cause of this degradation. To alleviate it, we have developed a network architecture called a *2-1-2D CNN*, shown in Fig. 2(c). In this network, 2D convolution is used for downsampling and upsampling, and 1D convolution is used for the main conversion process (i.e., residual blocks). To adjust the channel dimension, we apply $1 \times 1$ convolution before or after reshaping the feature map.

### 3.3. Improved Discriminator: PatchGAN

In previous GAN-based speech processing models [39, 40, 5, 27], FullGAN (Fig. 3(a)) has been extensively used. However, recent studies in computer vision [41, 42] indicate that the wide-range receptive fields of the discriminator require more parameters, which causes difficulty in training. Inspired by this, we replace FullGAN with *PatchGAN* [43, 41, 42] (Fig. 3(b)), which uses convolution at the last layer and determines the realness on the basis of the patch. We experimentally examine its effect for non-parallel VC in Section 4.2.

## 4. EXPERIMENTS

### 4.1. Experimental Conditions

**Dataset:** We evaluated our method on the **Spoke** (i.e., **non-parallel VC**) task of the VCC 2018 [36], which includes recordings of professional US English speakers. We selected a subset of speakers so as to cover all inter-gender and intra-gender conversions: VCC2SF3 (***SF***), VCC2SM3 (***SM***), VCC2TF1 (***TF***), and VCC2TM1 (***TM***), where S, T, F, and M indicate source, target, female, and male, respectively. In the following, we use the abbreviations in the parenthesis (e.g., ***SF***). Combinations of 2 sources (***SF*** or ***SM***) × 2 targets (***TF*** or ***TM***) were used for evaluation. Each speaker has sets of 81 (about 5 minutes; **relatively few** for VC) and 35 sentences for training and evaluation, respectively. In the Spoke task, the source and target speakers have a **different** set of sentences (**no overlap**) so as to evaluate in a non-parallel setting. The recordings were downsampled to 22.05 kHz for this challenge. We extracted 34 Mel-cepstral coefficients (MCEPs), logarithmic fundamental frequency ($\log F_0$), and aperiodicities (APs) every 5 ms by using the WORLD analyzer [44].

**Conversion process:** The proposed method was used to convert MCEPs ($Q = 34+1$ dimensions including 0th coefficient).[2] The objective of these experiments was to analyze the quality of the converted MCEPs; therefore, for the other parts, we used typical methods similar to the baseline of the VCC 2018 [36]. Specifically, in inter-gender conversion, a vocoder-based VC method was used. $F_0$ was converted by using logarithm Gaussian normalized transformation [45], APs were directly used without modification, and the WORLD synthesizer [44] was used to synthesize speech. In intra-gender conversion, we used a vocoder-free VC method [46]. More precisely, we calculated differential MCEPs by taking the difference between the source and converted MCEPs. For a similar reason, we did not use any postfilter [39, 40, 47] or powerful vocoder such as the WaveNet vocoder [48, 49]. Incorporating them is one possible direction of future work.

**Training details:** The implementation was almost the same as that of CycleGAN-VC except that the improved techniques were incorporated. The details of the network architectures are given in Fig. 4. For a pre-process, we normalized the source and target MCEPs to zero-mean and unit-variance by using the statistics of the training sets. To stabilize training, we used a least squares GAN (LSGAN) [50]. To increase the randomness of training data, we randomly cropped a segment (128 frames) from a randomly selected sentence instead of using an overall sentence directly. We used the Adam optimizer [51] with a batch size of 1. We trained the networks for $2 \times 10^5$ iterations with learning rates of 0.0002 for the generator and 0.0001 for the discriminator and with momentum term $\beta_1$ of 0.5. We set $\lambda_{cyc} = 10$ and $\lambda_{id} = 5$. We used $\mathcal{L}_{id}$ only for the first $10^4$ iterations to guide the learning direction. **Note that we did not use any extra data, modules, or time alignment procedures for training.**

[2] For reference, the converted speech samples, in which the proposed method was used to convert all acoustic features (namely, MCEPs, band APs, continuous $\log F_0$, and voice/unvoice indicator), are provided at http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc2/index.html. Even in this challenging setting, CycleGAN-VC2 works reasonably well.

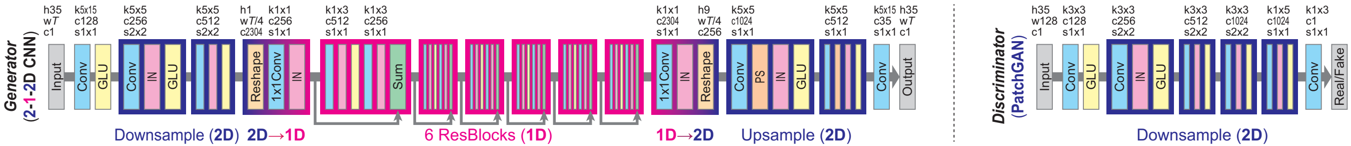


**Fig. 4**. Network architectures of generator and discriminator. In input, output, and reshape layers, h, w, and c denote height, width, and number of channels, respectively. In each convolution layer, k, c, and s denote kernel size, number of channels, and stride size, respectively. IN, GLU, and PS indicate instance normalization [52], gated linear unit [34], and pixel shuffler [42], respectively. Since the generator is fully convolutional [53], it can take input of arbitrary length $T$.

**Table 1**. Comparison of MCD [dB]

| No. | **Method** | | | **Intra-gender** | | **Inter-gender** | |
|---|---|---|---|---|---|---|---|
| | CycleGAN-VC2 | | | SF-TF | SM-TM | SM-TF | SF-TM |
| | Adv. | *G* | *D* | | | | |
| 1 | 1Step | 2-1-2D | Patch | 6.86±.04 | 6.32±.06 | 7.36±.04 | 6.28±.04 |
| 2 | 2Step | 1D | Patch | 6.86±.04 | 6.73±.08 | 7.77±.07 | 6.41±.01 |
| 3 | 2Step | 2D | Patch | 7.01±.07 | 6.63±.03 | 7.63±.03 | 6.73±.04 |
| 4 | 2Step | 2-1-2D | Full | 7.01±.07 | 6.45±.05 | 7.41±.04 | 6.51±.02 |
| 5 | **2Step** | **2-1-2D** | **Patch** | **6.83±.01** | **6.31±.03** | **7.22±.05** | **6.26±.03** |
| 6 | CycleGAN-VC [27] | | | 7.37±.03 | 6.68±.07 | 7.68±.05 | 6.51±.05 |
| 7 | Frame-based CycleGAN [28] | | | 8.85±.07 | 7.27±.11 | 8.86±.27 | 8.51±.36 |

**Table 2**. Comparison of MSD [dB]

| No. | **Method** | | | **Intra-gender** | | **Inter-gender** | |
|---|---|---|---|---|---|---|---|
| | CycleGAN-VC2 | | | SF-TF | SM-TM | SM-TF | SF-TM |
| | Adv. | *G* | *D* | | | | |
| 1 | 1Step | 2-1-2D | Patch | 1.60±.02 | 1.63±.05 | 1.54±.03 | 1.56±.04 |
| 2 | 2Step | 1D | Patch | 3.31±.36 | 4.26±.37 | 2.04±.21 | 5.03±.32 |
| 3 | 2Step | 2D | Patch | 1.57±.07 | 1.54±.01 | 1.46±.03 | 1.66±.07 |
| 4 | 2Step | 2-1-2D | Full | 1.52±.02 | 1.56±.04 | 1.47±.01 | 1.67±.06 |
| 5 | **2Step** | **2-1-2D** | **Patch** | **1.49±.01** | **1.53±.02** | **1.45±.00** | **1.52±.01** |
| 6 | CycleGAN-VC [27] | | | 2.42±.08 | 2.66±.08 | 2.21±.13 | 2.65±.15 |
| 7 | Frame-based CycleGAN [28] | | | 3.78±.26 | 2.77±.10 | 3.32±.06 | 3.61±.15 |

### 4.2. Objective Evaluation

As discussed in previous studies [7, 39], it is fairly complex to design a single metric that can assess the quality of converted MCEPs comprehensively. Alternatively, we used two metrics to assess the local and global structures. To measure global structural differences, we used the Mel-cepstral distortion (MCD), which measures the distance between the target and converted MCEP sequences. To measure the local structural differences, we used the modulation spectra distance (MSD), which is defined as the root mean square error between the target and converted logarithmic modulation spectra of MCEPs averaged over all MCEP dimensions and modulation frequencies. For both metrics, smaller values indicate that target and converted MCEPs are more similar.

We list the MCD and MSD in Tables 1 and 2, respectively. To eliminate the effect of initialization, we report the average and standard deviation scores over three random initializations. To analyze the effect of each technique, we conducted ablation studies on ***CycleGAN-VC2*** (no. 5 is the full model). We also compared ***CycleGAN-VC2*** with two state-of-the-art methods: ***CycleGAN-VC*** [27] and ***frame-based CycleGAN*** [28] (our reimplementation; we additionally used $\mathcal{L}_{id}$ for stabilizing training). The comparison of one-step and two-step adversarial losses (nos. 1, 5) indicates that this technique is particularly effective for improving MSD. The comparisons of generator (nos. 2, 3, 5) and discriminator (nos. 4, 5) network architectures indicate that they contribute to improving both MCD and MSD. Finally, the comparison to the baselines (nos. 5, 6, 7) verifies that by incorporating the three proposed techniques, we achieve state-of-the-art performance in terms of MCD and MSD for every speaker pair.

### 4.3. Subjective Evaluation

We conducted listening tests to evaluate the quality of converted speech. ***CycleGAN-VC*** [27] was used as the baseline. To measure naturalness, we conducted a mean opinion score (MOS) test (5: excellent and 1: bad), in which we included the target speech as a reference (MOS for ***TF*** and ***TM*** are 4.8). Ten sentences were randomly selected from the evaluation sets. To measure speaker similarity, we conducted an XAB test, where "A" and "B" were speech converted by the baseline and proposed methods, and "X" was target speech. We selected ten sentence pairs randomly from the evaluation sets and presented all pairs in both orders (AB and BA) to eliminate bias in the order of stimuli. For each sentence pair, the listeners were asked to select their preferred one ("A" or "B") or to opt for "Fair." Ten listeners participated in these listening tests. We show the MOS for naturalness and the preference scores for speaker similarity in Figs. 5 and 6, respectively. These results show that ***CycleGAN-VC2*** outperforms ***CycleGAN-VC*** in terms of both naturalness and similarity for every speaker pair. Particularly, ***CycleGAN-VC*** is difficult to apply to a vocoder-free VC framework [46] (used in ***SF-TF*** and ***SM-TM***), as this framework is sensitive to conversion error due to the usage of differential MCEPs. However, the MOS indicates that ***CycleGAN-VC2*** works relatively well in such a difficult setting.

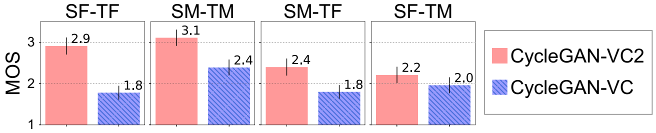


**Fig. 5**. MOS for naturalness with 95% confidence intervals

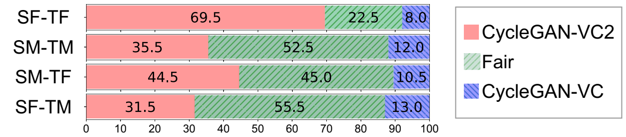


**Fig. 6**. Average preference score (%) on speaker similarity

## 5. CONCLUSION

To advance the research on non-parallel VC, we have proposed CycleGAN-VC2, which is an improved version of CycleGAN-VC incorporating three new techniques: an improved objective (two-step adversarial losses), improved generator (2-1-2D CNN), and improved discriminator (PatchGAN). The experimental results demonstrate that CycleGAN-VC2 outperforms CycleGAN-VC in both objective and subjective measures for every speaker pair. Our proposed techniques are not limited to one-to-one VC, and adapting them to other settings (e.g., many-to-many VC [54]) and other applications [1, 2, 4, 3, 5] remains an interesting future direction.

**Acknowledgements:** This work was supported by JSPS KAKENHI 17H01763.

# 6. REFERENCES

[1] Alexander B Kain, John-Paul Hosom, Xiaochuan Niu, Jan P. H. van Santen, Melanie Fried-Oken, and Janice Staehely, "Improving the intelligibility of dysarthric speech," *Speech Commun.*, vol. 49, no. 9, pp. 743–759, 2007.

[2] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Commun.*, vol. 54, no. 1, pp. 134–146, 2012.

[3] Zeynep Inanoglu and Steve Young, "Data-driven emotion conversion in spoken English," *Speech Commun.*, vol. 51, no. 3, pp. 268–283, 2009.

[4] Tomoki Toda, Mikihiro Nakagiri, and Kiyohiro Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 9, pp. 2505–2517, 2012.

[5] Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *Proc. Interspeech*, 2017, pp. 1283–1287.

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

[7] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[8] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 912–921, 2010.

[9] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1859–1872, 2014.

[10] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "Voice conversion based on speaker-dependent restricted Boltzmann machines," *IEICE Trans. Inf. Syst.*, vol. 97, no. 6, pp. 1403–1410, 2014.

[11] Srinivas Desai, Alan W Black, B Yegnanarayana, and Kishore Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 954–964, 2010.

[12] Seyed Hamidreza Mohammadi and Alexander Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *Proc. SLT*, 2014, pp. 19–23.

[13] Keisuke Oyamada, Hirokazu Kameoka, Takuhiro Kaneko, Hiroyasu Ando, Kaoru Hiramatsu, and Kunio Kashino, "Non-native speech conversion with consistency-aware recursive network and generative adversarial network," in *Proc. APSIPA ASC*, 2017, pp. 182–188.

[14] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted Boltzmann machines for voice conversion," in *Proc. Interspeech*, 2014, pp. 2278–2282.

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

[16] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exampler-based voice conversion using sparse representation in noisy environments," *IEICE Trans. Inf. Syst.*, vol. E96-A, no. 10, pp. 1946–1953, 2013.

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

[18] Feng-Long Xie, Frank K Soong, and Haifeng Li, "A KL divergence and DNN-based approach to voice conversion without parallel training sentences," in *Proc. Interspeech*, 2016, pp. 287–291.

[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 *Proc. ICASSP*, 2018, pp. 5274–5278.

[20] Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 3, pp. 952–963, 2006.

[21] Chung-Han Lee and Chung-Hsien Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *Proc. ICSLP*, 2006, pp. 2254–2257.

[22] Tomoki Toda, Yamato Ohtani, and Kiyohiro Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *Proc. Interspeech*, 2006, pp. 2446–2449.

[23] Daisuke Saito, Keisuke Yamamoto, Nobuaki Minematsu, and Keikichi Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Proc. Interspeech*, 2011, pp. 653–656.

[24] Toru Nakashika, Tetsuya Takiguchi, and Yasuhiro Minami, "Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2032–2045, 2016.

[25] 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 *Proc. Interspeech*, 2017, pp. 3364–3368.

[26] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," in *arXiv preprint arXiv:1808.05092*. Aug. 2018.

[27] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," in *arXiv preprint arXiv:1711.11293*. Nov. 2017.

[28] Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *Proc. ICASSP*, 2018, pp. 5279–5283.

[29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Proc. NPIS*, 2014, pp. 2672–2680.

[30] Takuhiro Kaneko and Hirokazu Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. EUSIPCO*, 2018, pp. 2114–2118.

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2223–2232.

[32] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. ICCV*, 2017, pp. 2849–2857.

[33] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. ICML*, 2017, pp. 1857–1865.

[34] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, 2017, pp. 933–941.

[35] Yaniv Taigman, Adam Polyak, and Lior Wolf, "Unsupervised cross-domain image generation," in *Proc. ICLR*, 2017.

[36] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Speaker Odyssey*, 2018, pp. 195–202.

[37] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros, "Learning dense correspondence via 3D-guided cycle consistency," in *Proc. CVPR*, 2016, pp. 117–126.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[39] Takuhiro Kaneko, Hirokazu Kameoka, Nobukatsu Hojo, Yusuke Ijima, Kaoru Hiramatsu, and Kunio Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. ICASSP*, 2017, pp. 4910–4914.

[40] Takuhiro Kaneko, Shinji Takaki, Hirokazu Kameoka, and Junichi Yamagishi, "Generative adversarial network-based postfilter for STFT spectrograms," in *Proc. Interspeech*, 2017, pp. 3389–3393.

[41] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 5967–5976.

[42] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. CVPR*, 2016, pp. 1874–1883.

[43] Chuan Li and Michael Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. ECCV*, 2016, pp. 702–716.

[44] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[45] Kun Liu, Jianping Zhang, and Yonghong Yan, "High quality voice conversion through phoneme-based linear mapping functions with STRAIGHT for Mandarin," in *Proc. FSKD*, 2007, pp. 410–414.

[46] Kazuhiro Kobayashi, Tomoki Toda, and Satoshi Nakamura, "F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential," in *Proc. SLT*, 2016, pp. 693–700.

[47] Kou Tanaka, Takuhiro Kaneko, Nobukatsu Hojo, and Hirokazu Kameoka, "Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks," in *Proc. SLT*, 2018, pp. 632–639.

[48] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio," in *arXiv preprint arXiv:1609.03499*. Sep. 2016.

[49] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017, pp. 1118–1122.

[50] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Proc. ICCV*, 2017, pp. 2794–2802.

[51] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[52] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Instance normalization: The missing ingredient for fast stylization," in *arXiv preprint arXiv:1607.08022*. July 2016.

[53] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2015, pp. 3431–3440.

[54] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. SLT*, 2018, pp. 266–273.