

StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion

Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, Nobukatsu Hojo

NTT Communication Science Laboratories, NTT Corporation, Japan

takuhiro.kaneko.tb@hco.ntt.co.jp

Abstract

Non-parallel multi-domain voice conversion (VC) is a technique for learning mappings among multiple domains without relying on parallel data. This is important but challenging owing to the requirement of learning multiple mappings and the non-availability of explicit supervision. Recently, StarGAN-VC has garnered attention owing to its ability to solve this problem only using a single generator. However, there is still a gap between real and converted speech. To bridge this gap, we rethink conditional methods of StarGAN-VC, which are key components for achieving non-parallel multi-domain VC in a single model, and propose an improved variant called StarGAN-VC2. Particularly, we rethink conditional methods in two aspects: training objectives and network architectures. For the former, we propose a source-and-target conditional adversarial loss that allows all source domain data to be convertible to the target domain data. For the latter, we introduce a modulation-based conditional method that can transform the modulation of the acoustic feature in a domain-specific manner. We evaluated our methods on non-parallel multi-speaker VC. An objective evaluation demonstrates that our proposed methods improve speech quality in terms of both global and local structure measures. Furthermore, a subjective evaluation shows that StarGAN-VC2 outperforms StarGAN-VC in terms of naturalness and speaker similarity.¹

Index Terms: voice conversion (VC), non-parallel VC, multi-domain VC, generative adversarial networks (GANs), StarGAN-VC

1. Introduction

Voice conversion (VC) is a technique for converting the non/para-linguistic information between source and target speech while preserving the linguistic information. VC has been studied intensively owing to its high potential for various applications, such as speaking aids [1, 2] and style [3, 4] and pronunciation [5] conversion.

One well-established approach to VC involves statistical methods based on Gaussian mixture models (GMMs) [6, 7, 8], neural networks (NNs) (including restricted Boltzmann machines (RBMs) [9, 10], feed forward NNs (FNNs) [11, 12, 13], recurrent NNs (RNNs) [14, 15], convolutional NNs (CNNs) [5], attention networks [16, 17], and generative adversarial networks (GANs) [5]), and exemplar-based methods using non-negative matrix factorization (NMF) [18, 19].

Many VC methods (including the above-mentioned) are categorized as parallel VC, which learns a mapping using the training data of parallel utterance pairs. However, obtaining

such data is often time-consuming or impractical. Moreover, even if such data are obtained, most VC methods rely on a time alignment procedure, which occasionally fails and requires other painstaking processes, i.e., careful pre-screening or manual correction.

As a solution, non-parallel VC has begun to be studied. Non-parallel VC, which is comparable to parallel VC, is generally quite challenging to achieve owing to its disadvantageous training conditions. To mitigate this difficulty, several studies have used additional data (e.g., parallel utterance pairs among reference speakers [20, 21, 22, 23]) or extra modules (e.g., automatic speech recognition (ASR) modules [24, 25]). These additional data and extra modules are useful for simplifying training but require other costs for preparation. To avoid such additional costs, recent studies have introduced probabilistic deep generative models, such as an RBM [26], variational autoencoders (VAEs) [27, 28]), and GANs [27, 29]. Among them, CycleGAN-VC [29] (published [30] and further improved [31]) shows promising results by configuring CycleGAN [32, 33, 34] with a gated CNN [35] and identity-mapping loss [36]. This makes it possible to learn a sequence-based mapping function without relying on parallel data. With this improvement, CycleGAN-VC performs comparably to parallel VC [7].

Along with non-parallel VC, another practically important issue is non-parallel multi-domain VC, i.e., learning mappings among multiple domains (e.g., multiple speakers) without relying on parallel data. This problem is challenging in terms of scalability because typical VC methods (including CycleGAN-VC) are designed to learn a one-to-one mapping; therefore, they require the learning of multiple generators to achieve multi-domain VC. For this problem, StarGAN-VC [37] provides a promising solution by extending CycleGAN-VC to a conditional setting and incorporating domain codes. Through this extension, StarGAN-VC makes it possible to achieve non-parallel multi-domain VC by only using a single generator while maintaining the advantage of CycleGAN-VC. The subjective evaluation [37] demonstrates that StarGAN-VC outperforms another state-of-the-art method, i.e., VAE/GAN-VC [27].

However, even using StarGAN-VC, there is still an insurmountable gap between real and converted speech. To bridge this gap, we rethink conditional methods of StarGAN-VC, which are key components for solving non-parallel multi-domain VC using a single generator, and propose an improved variant called StarGAN-VC2. In particular, we rethink conditional methods in two aspects: training objectives and network architectures. For the former, we propose a source-and-target conditional adversarial loss, which encourages all source domain data to be converted into the target data. For the latter, we introduce a modulation-based conditional method that can transform the modulation of acoustic features in a domain-dependent manner. We examined the performance of the proposed methods on the multi-speaker VC task using the Voice

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

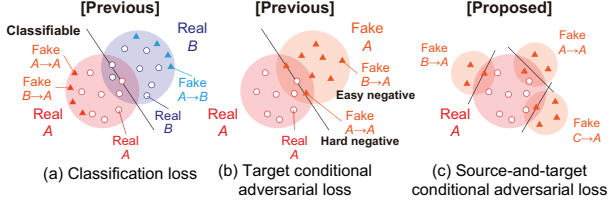


Figure 1: Comparison of conditional methods in training objectives. “A” and “B” denote the domain codes and “A → B” represents the data converted from “A” to “B.” Circle and triangle markers denote real and fake data, respectively. (a) In the classification loss, G prefers to generate classifiable (i.e., far from the decision boundary) data. (b) In the target conditional adversarial loss, D needs to simultaneously handle hard negative samples (e.g., conversion between the same speaker $A \rightarrow A$) and easy negative samples (e.g., conversion between completely different speakers $B \rightarrow A$). (c) The proposed source-and-target conditional adversarial loss can bring all the converted data close to the target data in both source-wise and target-wise manners.

Conversion Challenge 2018 (VCC 2018) dataset [38]. An objective evaluation demonstrates that the proposed methods effectively bring the converted acoustic feature sequence close to the target one in terms of both global and local structure measures. A subjective evaluation shows that StarGAN-VC2 outperforms StarGAN-VC in terms of both naturalness and speaker similarity.

In Section 2, we review the conventional StarGAN-VC. In Section 3, we describe the proposed StarGAN-VC2. In Section 4, we report the experimental results. We conclude in Section 5 with a brief summary and mention of future work.

2. Conventional StarGAN-VC

2.1. Training objectives

The aim of StarGAN-VC is to obtain a single generator G that learns mappings among multiple domains (e.g., multiple speakers). To achieve this, StarGAN-VC extends CycleGAN-VC to a conditional setting with a domain code (e.g., a speaker identifier). More precisely, StarGAN-VC learns a generator G that converts an input acoustic feature \mathbf{x} into an output feature \mathbf{x}' conditioned on the target domain code c' , i.e., $G(\mathbf{x}, c') \rightarrow \mathbf{x}'$. Here, let $\mathbf{x} \in \mathbb{R}^{Q \times T}$ be an acoustic feature sequence where Q is the feature dimension and T is the sequence length, and let $c \in \{1, \dots, N\}$ be the corresponding domain code where N is the number of domains. Inspired by StarGAN [39], which was originally proposed in computer vision for multi-domain image-to-image translation, StarGAN-VC solves this problem by using an adversarial loss [40], classification loss [41], and cycle-consistency loss [42]. Additionally, inspired by CycleGAN-VC [29], StarGAN-VC also uses an identity-mapping loss [36] to preserve linguistic composition.

Adversarial loss: The adversarial loss is used to render the converted feature indistinguishable from the real target feature:

$$\mathcal{L}_{t-adv} = \mathbb{E}_{(\mathbf{x}, c) \sim P(\mathbf{x}, c)} [\log D(\mathbf{x}, c)] + \mathbb{E}_{\mathbf{x} \sim P(\mathbf{x}), c' \sim P(c')} [\log(1 - D(G(\mathbf{x}, c'), c'))], \quad (1)$$

where D is a target conditional discriminator [43]. By maximizing this loss, D attempts to learn the best decision boundary between the converted and real acoustic features conditioned on the target domain codes (c and c'). In contrast, G attempts to

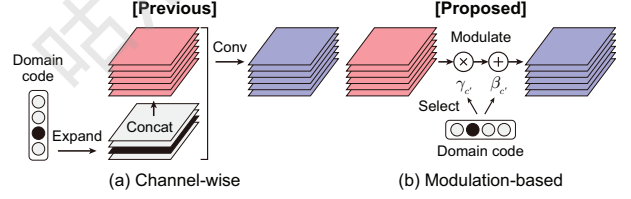


Figure 2: Comparison of conditional methods in generator networks. We consider the case when convolutional networks are used. Such networks are commonly used in state-of-the-art VC models (e.g., CycleGAN-VC [29] and StarGAN-VC [37]).

render the converted feature indistinguishable from real acoustic features conditioned on c' by minimizing this loss.

Classification loss: The aim of StarGAN-VC is to synthesize the acoustic feature that belongs to the target domain. To achieve this, the classification loss is used. First, the classifier C is trained for real acoustic features:

$$\mathcal{L}_{cls}^r = \mathbb{E}_{(\mathbf{x}, c) \sim P(\mathbf{x}, c)} [-\log C(c|\mathbf{x})], \quad (2)$$

where C attempts to classify a real acoustic feature \mathbf{x} to the corresponding domain c by minimizing this loss. Subsequently, G is optimized for C :

$$\mathcal{L}_{cls}^f = \mathbb{E}_{\mathbf{x} \sim P(\mathbf{x}), c' \sim P(c')} [-\log C(c'|G(\mathbf{x}, c'))], \quad (3)$$

where G attempts to generate an acoustic feature that is classified to the target domain c' by minimizing this loss.

Cycle-consistency loss: Although the adversarial loss and classification loss encourage a converted acoustic feature to become realistic and classifiable, respectively, they do not guarantee that the converted feature will preserve the input composition. To mitigate this problem, the cycle-consistency loss is used:

$$\mathcal{L}_{cyc} = \mathbb{E}_{(\mathbf{x}, c) \sim P(\mathbf{x}, c), c' \sim P(c')} [\|\mathbf{x} - G(G(\mathbf{x}, c'), c)\|_1]. \quad (4)$$

This cyclic constraint encourages G to find out an optimal source and target pair that does not compromise the composition.

Identity-mapping loss: To impose a further constraint on the input preservation, the identity-mapping loss is used:

$$\mathcal{L}_{id} = \mathbb{E}_{(\mathbf{x}, c) \sim P(\mathbf{x}, c)} [\|\mathbf{x} - G(\mathbf{x}, c)\|_1]. \quad (5)$$

Full objective: The full objective is written as

$$\mathcal{L}_D = -\mathcal{L}_{t-adv}, \quad (6)$$

$$\mathcal{L}_C = \lambda_{cls} \mathcal{L}_{cls}^r, \quad (7)$$

$$\mathcal{L}_G = \mathcal{L}_{t-adv} + \lambda_{cls} \mathcal{L}_{cls}^f + \lambda_{cyc} \mathcal{L}_{cyc} + \lambda_{id} \mathcal{L}_{id}, \quad (8)$$

where D , C , and G are optimized by minimizing \mathcal{L}_D , \mathcal{L}_C , and \mathcal{L}_G , respectively.

2.2. Network architectures

Regarding the network architectures, this study focuses on the conditional method in the generator. Hence, here we review the StarGAN-VC generator network architecture. As shown in Figure 2(a), StarGAN-VC incorporates conditional information into the generator in a channel-wise manner, i.e., first creates the one-hot vector indicating the domain code, subsequently expands the one-hot vector to the feature map size, and finally concatenates it to the feature map. Concatenated features are convoluted together and propagated to the next layer.

3. StarGAN-VC2

3.1. Rethinking conditional method in training objectives

We first rethink a conditional method in training objectives. As described in Section 2.1, StarGAN-VC uses two conditional methods to make the converted feature belonging to the target domain: the classification loss (Equations 2 and 3) and the target conditional adversarial loss (Equation 1). We illustrate their training strategies in Figure 1(a) and (b), respectively.

In the classification loss (Figure 1(a)), via Equation 2, the decision boundary (black line) is learned among real-data domains (e.g., between “Real A” and “Real B” in Figure 1(a)). For this decision boundary, G attempts to generate easily “classifiable” data via Equation 3. This means that G prefers to generate data that are far from the decision boundary even when the real data exist around the decision boundary. As discussed elsewhere [41, 44, 45], this prevents G from covering the whole real data distribution. In VC, this may result in a partial conversion.

Meanwhile, the target conditional adversarial loss (Figure 1(b)) encourages the generated data close to the real data conditioned on the target domain code. As discussed in the previous study [44], this objective prevents G from leaning towards generating only classifiable data. However, a possible difficulty is that this loss needs to simultaneously handle diverse data, including hard negative samples (e.g., conversion between the same speaker $A \rightarrow A$ in Figure 1(b)) and easy negative samples (e.g., conversion between completely different speakers $B \rightarrow A$ in Figure 1(b)). This unfair condition makes it difficult to bring all the converted data close to real target data.

To solve this problem, we develop a source-and-target conditional adversarial loss defined as

$$\begin{aligned} \mathcal{L}_{st-adv} = & \mathbb{E}_{(\mathbf{x}, c) \sim P(\mathbf{x}, c), c' \sim P(c')} [\log D(\mathbf{x}, c', c)] \\ & + \mathbb{E}_{(\mathbf{x}, c) \sim P(\mathbf{x}, c), c' \sim P(c')} [\log D(G(\mathbf{x}, c, c'), c, c')], \end{aligned} \quad (9)$$

where $c' \sim P(c')$ is randomly sampled independently of real data. Differently from Equation 1, both G and D are conditioned on the source code c' in addition to the target code c . We call such G and D a source-and-target conditional generator and discriminator, respectively. As shown in Figure 1(c), by using both source and target domain codes as conditional information, this loss encourages all the converted data to be close to real data in both source-wise and target-wise manners. This resolves the unfair training condition in the target conditional adversarial loss (Figure 1(b)) and allows all the source domain data to be converted into the target domain data.

One possible disadvantage of the source-and-target conditional generator is that this requires the availability of the source code in inference, which is not required in the conventional StarGAN-VC. However, note that speaker recognition has been actively studied (e.g., [46]), and this problem can be alleviated by using it as a pre-process.

3.2. Rethinking conditional method in networks

As indicated by previous studies on VC postfilters (e.g., global variance [7] and modulation spectrum [47] postfilters), accurate modulation translation is important to achieve high-quality VC. Particularly, to achieve multi-domain VC only using a single generator, a framework must be incorporated that can conduct diverse domain-specific modulations effectively. For this challenge, a channel-wise conditional method (Figure 2(a)) is not effective because the concatenated conditional information can be additively used in a convolutional procedure but cannot

be multiplicatively used to modulate features. To alleviate this problem, we introduce a modulation-based conditional method, which can directly modulate features in a domain-dependent manner. In particular, we introduce a conditional instance normalization (CIN) [48], which was originally proposed in computer vision for style transfer. As shown in Figure 2(b), given the feature \mathbf{f} , CIN conducts the following procedure:

$$\text{CIN}(\mathbf{f}; c') = \gamma_{c'} \left(\frac{\mathbf{f} - \mu(\mathbf{f})}{\sigma(\mathbf{f})} \right) + \beta_{c'}, \quad (10)$$

where $\mu(\mathbf{f})$ and $\sigma(\mathbf{f})$ are the average and standard deviation of \mathbf{f} that are calculated over for each instance. $\gamma_{c'}$ and $\beta_{c'}$ are domain-specific scale and bias parameters that allow the modulation to be transformed in a domain-specific manner. These parameters are learnable and optimized through training.

In the above, we explain the case when the generator is conditioned on the target domain code c' . When using a source-and-target conditional generator introduced in Equation 9, we replace $\gamma_{c'}$ and $\beta_{c'}$ with $\gamma_{c, c'}$ and $\beta_{c, c'}$, respectively, which are selected depending on both the source c and target c' .

4. Experiments

4.1. Experimental conditions

Dataset: We evaluated our method on the multi-speaker VC task using VCC 2018 [38], which contains recordings of professional US English speakers. Following the StarGAN-VC study [37], we selected a subset of speakers as covering all inter- and intra-gender conversions: VCC2SF1, VCC2SF2, VCC2SM1, and VCC2SM2, where F and M indicate female and male speakers, respectively. Thus, the number of domains N is set to 4. Our goal is to learn $4 \times 3 = 12$ different source-and-target mappings in a single model. Each speaker has sets of 81 and 35 sentences for training and evaluation, respectively. 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 [52].

Conversion process: In these experiments, we focused on analyzing the performance in MCEP conversion. Hence, we applied the proposed method only to MCEP conversion,² and for the other parts, we used typical methods, i.e., converted $\log F_0$ using logarithmic Gaussian normalized transformation [53], directly used APs, and synthesized speech using the WORLD vocoder [52]. To examine the pure effect of the proposed methods, we did not use any postfilter [54, 55, 56] or powerful vocoder such as the WaveNet vocoder [57, 58]. Incorporating them remains possible future work.

Implementation details: We designed the network architectures on the basis of CycleGAN-VC2 [31], i.e., we used a 2-1-2D CNN in G and a 2D CNN in D . We formulate D using the projection discriminator [44]. In the pre-experiment, we found that skip connections in residual blocks [59] result in partial conversion. Thus, we removed them in G . The details of the network architectures are given in Figure 3. For a GAN objective, we used a least squares GAN [60]. We conducted speaker-wise normalization for a pre-process. We trained the networks

²For reference, the converted speech samples, in which the proposed method was applied 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/stargan-vc2/index.html>. Even in this challenging setting, StarGAN-VC2 works reasonably well.

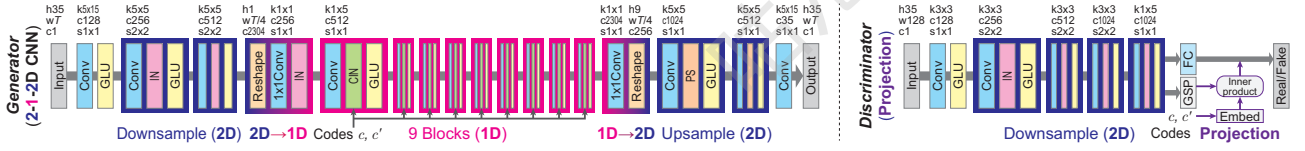


Figure 3: *Generator and discriminator network architectures. In input, output, and reshape layers, h, w, and c represent height, width, and number of channels, respectively. In each convolution layer, k, c, and s denote kernel size, number of channels, and stride, respectively. IN, GLU, PS, and GSP indicate instance normalization [49], gated linear unit [35], pixel shuffler [50], and global sum pooling, respectively. The generator is fully convolutional [51]. This allows an arbitrary length T to be input in inference.*

Table 1: *Comparison of MCD and MSD among models using different conditional methods in training objectives. We fix the conditional method in G network as modulation-based.*

| Objective | MCD [dB] | MSD [dB] |
|--|----------------------------------|----------------------------------|
| \mathcal{L}_{cls} | $7.73 \pm .07$ | $1.96 \pm .03$ |
| \mathcal{L}_{t-adv} | $7.21 \pm .16$ | $2.87 \pm .51$ |
| $\mathcal{L}_{t-adv} + \mathcal{L}_{cls}$ (StarGAN-VC) | $7.11 \pm .10$ | $2.41 \pm .13$ |
| \mathcal{L}_{st-adv} (StarGAN-VC2) | $6.90 \pm .07$ | $1.89 \pm .03$ |

Table 2: *Comparison of MCD and MSD among models using different conditional methods in G networks. We fix the conditional method in the training objective as \mathcal{L}_{st-adv} .*

| G network | MCD [dB] | MSD [dB] |
|--------------------------------|----------------|----------------------------------|
| Channel-wise (StarGAN-VC) | $6.90 \pm .08$ | $2.55 \pm .20$ |
| Modulation-based (StarGAN-VC2) | $6.90 \pm .07$ | $1.89 \pm .03$ |

using the Adam optimizer [61] with a batch size of 8, in which we used a randomly cropped segment (128 frames) as one instance. The number of iterations was set to 3×10^5 , learning rates for G and D were set to 0.0002 and 0.0001, respectively, and the momentum term was set to 0.5. We set $\lambda_{cyc} = 10$, $\lambda_{id} = 5$, and $\lambda_{cls} = 1$. We used \mathcal{L}_{id} only for the first 10^4 iterations to stabilize the training at the beginning.

4.2. Objective evaluation

We conducted an objective evaluation to validate the advantages of the proposed conditional methods over other conditional methods. The same as the previous study [31], we used two evaluation metrics for comprehensive analysis: the Mel-cepstral distortion (MCD), which measures the global structural differences by calculating the distance between the target and converted MCEPs, and the modulation spectra distance (MSD), which measures the local structural differences by computing the distance between the target and converted logarithmic modulation spectra of MCEPs. For both metrics, a smaller value indicates that the target and converted features are more similar.

We conducted comparative studies in two aspects: training objectives and network architectures, which are listed in Tables 1 and 2, respectively. We have calculated the scores averaged over three models trained with different initializations to eliminate the effect of initialization. In Table 1, the proposed source-and-target conditional loss \mathcal{L}_{st-adv} outperforms the other losses in terms of both the MCD and MSD. This indicates that the proposed loss is useful for improving the feature quality in terms of both the global and local structure measures. In Table 2, the proposed modulation-based conditional method outperforms the conventional channel-wise conditional method in terms of the MSD. This indicates that the proposed architecture is particularly useful for improving the local structure. Through these experiments, we empirically confirm that the proposed conditional methods in objectives and networks effectively bring the converted acoustic feature sequence close to the target one.

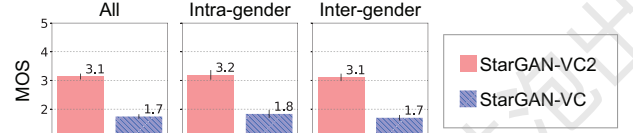


Figure 4: *MOS for naturalness with 95% confidence intervals*

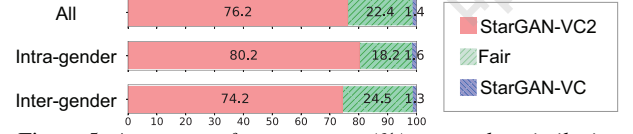


Figure 5: *Average preference scores (%) on speaker similarity*

4.3. Subjective evaluation

We conducted listening tests to analyze the performance compared with StarGAN-VC [37], which is a state-of-the-art multi-domain non-parallel VC. To measure naturalness, we conducted a mean opinion score (MOS) test (5: excellent to 1: bad), in which we included the analysis-synthesized speech (which is the upper limit of the converted speech) as a reference (MOS: 4.2). For each model, we generated 36 sentences (4×3 source-target combinations $\times 3$ sentences). We conducted an XAB test to measure speaker similarity. Here, “X” was target speech and “A” and “B” were speech converted by StarGAN-VC and StarGAN-VC2, respectively. For each model, we generated 24 sentences (4×3 source-target combinations $\times 2$ sentences). To eliminate bias in the order of stimuli, we presented all pairs in both orders (AB and BA). For each sentence pair, the listeners were asked to select their preferred one (“A” or “B”) or “Fair.” 12 well-educated English speakers participated in the tests.

Figures 4 and 5 show the MOS for naturalness and the preference scores for speaker similarity, respectively. We summarized the results on the basis of three categories: all conversion, inter-gender conversion, and intra-gender conversion. These results empirically demonstrate that StarGAN-VC2 outperforms StarGAN-VC in terms of both naturalness and speaker similarity for every category.

5. Conclusions

To advance the research on multi-domain non-parallel VC, we have rethought conditional methods in StarGAN-VC in two aspects: training objectives and network architectures. We developed a source-and-target conditional adversarial loss for the former and a modulation-based conditional method for the latter and have proposed StarGAN-VC2 incorporating them. The empirical studies on non-parallel multi-speaker VC demonstrate that StarGAN-VC2 outperforms StarGAN-VC in both objective and subjective measures. StarGAN-VC2 is a general model for multi-domain VC and is not limited to multi-speaker VC. Adapting it to other tasks (e.g., multi-emotion VC and multi-pronunciation VC) remains a promising future direction.

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

6. References

- [1] A. B. Kain, J.-P. Hosom, X. Niu, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Commun.*, vol. 49, no. 9, pp. 743–759, 2007.
- [2] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Commun.*, vol. 54, no. 1, pp. 134–146, 2012.
- [3] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Commun.*, vol. 51, no. 3, pp. 268–283, 2009.
- [4] T. Toda, M. Nakagiri, and K. 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] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *Proc. Interspeech*, 2017, pp. 1283–1287.
- [6] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.
- [7] T. Toda, A. W. Black, and K. 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] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 912–921, 2010.
- [9] 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 Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1859–1872, 2014.
- [10] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on speaker-dependent restricted Boltzmann machines," *IEICE Trans. Inf. Syst.*, vol. 97, no. 6, pp. 1403–1410, 2014.
- [11] S. Desai, A. W. Black, B. Yegnanarayana, and K. 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] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *Proc. SLT*, 2014, pp. 19–23.
- [13] K. Oyamada, H. Kameoka, T. Kaneko, H. Ando, K. Hiramatsu, and K. Kashino, "Non-native speech conversion with consistency-aware recursive network and generative adversarial network," in *Proc. APSIPA ASC*, 2017, pp. 182–188.
- [14] T. Nakashika, T. Takiguchi, and Y. Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted Boltzmann machines for voice conversion," in *Proc. Interspeech*, 2014, pp. 2278–2282.
- [15] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 4869–4873.
- [16] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. ICASSP*, 2019, pp. 6805–6809.
- [17] H. Kameoka, K. Tanaka, T. Kaneko, and N. Hojo, "ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion," in *arXiv preprint arXiv:1811.01609*, Nov. 2018.
- [18] R. Takashima, T. Takiguchi, and Y. Ariki, "Exampler-based voice conversion using sparse representation in noisy environments," *IEICE Trans. Inf. Syst.*, vol. E96-A, no. 10, pp. 1946–1953, 2013.
- [19] Z. Wu, T. Virtanen, E. S. Chng, and H. 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.
- [20] A. Mouchtaris, J. Van der Spiegel, and P. 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] C.-H. Lee and C.-H. Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *Proc. ICSLP*, 2006, pp. 2254–2257.
- [22] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *Proc. Interspeech*, 2006, pp. 2446–2449.
- [23] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Proc. Interspeech*, 2011, pp. 653–656.
- [24] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN-based approach to voice conversion without parallel training sentences," in *Proc. Interspeech*, 2016, pp. 287–291.
- [25] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *Proc. ICASSP*, 2018, pp. 5274–5278.
- [26] T. Nakashika, T. Takiguchi, and Y. 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.
- [27] 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," in *Proc. Interspeech*, 2017, pp. 3364–3368.
- [28] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 9, pp. 1432–1443, 2019.
- [29] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," in *arXiv preprint arXiv:1711.11293*, Nov. 2017.
- [30] —, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. EUSIPCO*, 2018, pp. 2114–2118.
- [31] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved cyclegan-based non-parallel voice conversion," in *Proc. ICASSP*, 2019, pp. 6820–6824.
- [32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2223–2232.
- [33] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. ICCV*, 2017, pp. 2849–2857.
- [34] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. ICML*, 2017, pp. 1857–1865.
- [35] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, 2017, pp. 933–941.
- [36] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. ICLR*, 2017.
- [37] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. SLT*, 2018, pp. 266–273.
- [38] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Speaker Odyssey*, 2018, pp. 195–202.
- [39] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. CVPR*, 2018, pp. 8789–8797.
- [40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.
- [41] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. ICML*, 2017, pp. 2642–2651.
- [42] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3D-guided cycle consistency," in *Proc. CVPR*, 2016, pp. 117–126.
- [43] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, Nov. 2014.
- [44] T. Miyato and M. Koyama, "cGANs with projection discriminator," in *Proc. ICLR*, 2018.
- [45] T. Kaneko, Y. Ushiku, and T. Harada, "Class-distinct and class-mutual image generation with GANs," in *Proc. BMVC*, 2019.
- [46] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *Proc. ICASSP*, 2015, pp. 4814–4818.
- [47] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," in *Proc. ICASSP*, 2014, pp. 290–294.
- [48] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," in *Proc. ICLR*, 2017.
- [49] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," in *arXiv preprint arXiv:1607.08022*, July 2016.
- [50] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. CVPR*, 2016, pp. 1874–1883.
- [51] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2015, pp. 3431–3440.
- [52] M. Morise, F. Yokomori, and K. 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.
- [53] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with STRAIGHT for Mandarin," in *Proc. FSKD*, 2007, pp. 410–414.
- [54] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. ICASSP*, 2017, pp. 4910–4914.
- [55] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for STFT spectrograms," in *Proc. Interspeech*, 2017, pp. 3389–3393.
- [56] K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka, "Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks," in *Proc. SLT*, 2018, pp. 632–639.
- [57] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *arXiv preprint arXiv:1609.03499*, Sept. 2016.
- [58] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017, pp. 1118–1122.
- [59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
- [60] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. ICCV*, 2017, pp. 2794–2802.
- [61] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.