CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks

Takuhiro Kaneko and Hirokazu Kameoka
NTT Communication Science Laboratories, NTT Corporation, Japan
{kaneko.takuhiro, kameoka.hirokazu}@lab.ntt.co.jp

***Abstract*—We propose a non-parallel voice-conversion (VC) method that can learn a mapping from source to target speech without relying on parallel data. The proposed method is particularly noteworthy in that it is general purpose and high quality and works without any extra data, modules, or alignment procedure. Our method, called CycleGAN-VC, uses a cycle-consistent adversarial network (CycleGAN) with gated convolutional neural networks (CNNs) and an identity-mapping loss. 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 non-parallel data. Furthermore, the adversarial loss can bring the converted speech close to the target one on the basis of indistinguishability without explicit density estimation. This allows to avoid over-smoothing caused by statistical averaging, which occurs in many conventional statistical model-based VC methods that represent data distribution explicitly. We configure a CycleGAN with gated CNNs and train it with an identity-mapping loss. This allows the mapping function to capture sequential and hierarchical structures while preserving linguistic information. We evaluated our method on a non-parallel VC task. An objective evaluation showed that the converted feature sequence was near natural in terms of global variance and modulation spectra, which are structural indicators highly correlated with subjective evaluation. A subjective evaluation showed that the quality of the converted speech was comparable to that obtained with a Gaussian mixture model-based parallel VC method even though CycleGAN-VC is trained under disadvantageous conditions (non-parallel and half the amount of data).**

***Index Terms*—voice conversion, non-parallel conversion, generative adversarial networks, CycleGAN, gated CNN**

I. Introduction

Voice conversion (VC) is a technique to modify non/para-linguistic information of speech while preserving linguistic information. This technique can be applied to various tasks such as speaker-identity modification for text-to-speech systems [2], speaking assistance [3]–[5], emotion/expressiveness conversion [6], [7], and pronunciation conversion [8].

Voice conversion can be formulated as a regression problem of estimating a mapping function from source to target speech. One successful approach involves statistical methods using a Gaussian mixture model (GMM) [9]–[11]. Neural network (NN)-based methods, such as a restricted Boltzmann machine (RBM) [12], [13], feed forward NN [14]–[16], recurrent NN (RNN) [17], [18], convolutional NN (CNN) [8], and generative adversarial network (GAN) [8], and exemplar-based methods, such as non-negative matrix factorization (NMF) [19], [20], have also recently been proposed.

Many VC methods including those mentioned above typically use temporally aligned parallel data of source and target speech as training data. If perfectly aligned parallel data are available, obtaining the mapping function becomes relatively simple; however, collecting such data can be a painstaking process in real application scenarios. Even though we could collect such data, we need to perform automatic time alignment, which may occasionally fail. This can be problematic since misalignment involved in parallel data can cause speech-quality degradation; thus, careful pre-screening and manual correction may be required [21].

These facts motivated us to consider a VC problem that does not rely on parallel data. In this paper, we propose a non-parallel VC method, which is particularly noteworthy in that it (1) does not require any extra data, such as transcripts and reference speech, and extra modules, such as an automatic speech-recognition (ASR) module, (2) is not prone to over-smoothing, which is known to be one of the main factors leading to speech-quality degradation, and (3) captures a spectrotemporal structure without any alignment procedure.

To satisfy these requirements, our method[1], called *CycleGAN-VC*, uses a cycle-consistent adversarial network (CycleGAN) [22] (i.e., DiscoGAN [23] or DualGAN [24]) with gated CNNs [25] and an identity-mapping loss [26]. The CycleGAN was originally proposed for unpaired image-to-image translation. With this model, forward and inverse mappings are simultaneously learned using an adversarial loss [27] and cycle-consistency loss [28]. This makes it possible to find an optimal pseudo pair from non-parallel data. Furthermore, the adversarial loss can bring the converted speech close to the target one on the basis of indistinguishability without explicit density estimation. This allows to avoid over-smoothing caused by statistical averaging [8], [16], [29]–[31], which occurs in many conventional statistical model-based VC methods that represent data distribution explicitly (e.g., Gaussian distribution). To use a CycleGAN for non-parallel VC, we configure a network using gated CNNs and train it with an identity-mapping loss. This allows the mapping function to capture sequential and hierarchical structures while preserving linguistic information.

We evaluated our method on a non-parallel VC task using the Voice Conversion Challenge 2016 (VCC 2016) dataset [32]. An objective evaluation showed that the converted feature sequence was near natural in terms of global variance (GV) [10] and modulation spectra (MS) [33], which are structural indicators highly correlated with subjective evaluation.

[1]A preprint version of this paper has already been shared publicly [1].

ISBN 978-90-827970-1-5 © EURASIP 2018

A subjective evaluation showed that the speech quality was comparable to that obtained with a GMM-based parallel VC method [10] even though *CycleGAN-VC* is trained under disadvantageous conditions (non-parallel and half the amount of data). Although the GMM-based VC method is not state-of-the-art, these results are noteworthy because to the best of our knowledge, there was still a gap between the quality of previous non-parallel VC and GMM-based parallel VC.

This paper is organized as follows. In Section II, we describe related work. In Section III, we review the CycleGAN and explain our proposed method (*CycleGAN-VC*). In Section IV, we report on the experimental results. In Section V, we provide a discussion and conclude the paper.

II. Related work

Recently, several approaches for non-parallel VC have been proposed. One approach involves using an ASR module to find a pair of corresponding frames [34], [35]. This may work well if ASR performs robustly and accurately enough, but it requires a large amount of transcripts to train the ASR module. It would also be inherently difficult to capture non-verbal information that is an important factor for several VC tasks. Other approaches involve methods using an adaptation technique [36], [37] or incorporating a pre-constructed speaker space [38], [39]. These methods do not require parallel data between source and target speakers but require parallel data among reference speakers. A few attempts [40]–[43] have recently been made to develop methods that are completely free from parallel data and extra modules, similarly to ours. With these methods, it is assumed that source and target speech lie in the same low-dimensional embeddings. This would not only limit applicable data but also the converted data tend to lose detailed structures through embedding. In contrast, we learn a mapping function directly without embedding. This would make our method a general-purpose solution for various VC tasks. We note that generally speaking, **there is still a gap between the quality of parallel and non-parallel VC, and further studies are needed in this field**.

III. Non-parallel VC using CycleGAN

A. CycleGAN

Our goal is to learn a mapping from source $x \in X$ to target $y \in Y$ without relying on parallel data. We solve this problem based on a CycleGAN [22]. With CycleGAN, a mapping $G_{X \to Y}$ is learned using two losses, namely an adversarial loss [27] and cycle-consistency loss [28]. We illustrate the training procedure in Fig. 1.

Adversarial loss: An adversarial loss measures how distinguishable converted data $G_{X\to Y}(x)$ are from target data y. Hence, the closer the distribution of converted data $P_{G_{X\to Y}}(x)$ becomes to that of target data $P_{\text{Data}}(y)$, the smaller this loss becomes. This objective is written as

$$\begin{aligned}\mathcal{L}_{adv}(G_{X\to Y}, D_Y) = {} & \mathbb{E}_{y\sim P_{\text{Data}}(y)}[\log D_Y(y)] \\ & + \mathbb{E}_{x\sim P_{\text{Data}}(x)}[\log(1 - D_Y(G_{X\to Y}(x)))]. \quad (1)\end{aligned}$$

The generator $G_{X\to Y}$ attempts to generate data that can deceive the discriminator D_Y by minimizing this loss, whereas D_Y attempts not to be deceived by $G_{X\to Y}$ by maximizing this loss.

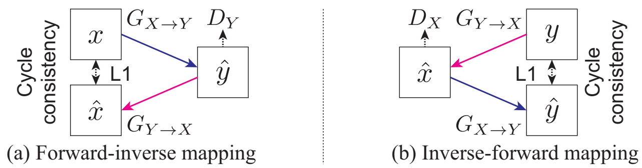

Fig. 1. Training procedure of CycleGAN

Cycle-consistency loss: Optimizing only the adversarial loss would not necessarily guarantee that the contextual information of x and $G_{X\to Y}(x)$ will be consistent. This is because the adversarial loss only tells us whether $G_{X\to Y}(x)$ follows the target-data distribution and does not help preserve the contextual information of x. The idea of CycleGAN [22] is to introduce two additional terms. One is an adversarial loss $\mathcal{L}_{adv}(G_{Y\to X}, D_X)$ for an inverse mapping $G_{Y\to X}$ and the other is a cycle-consistency loss, given as

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

These additional terms encourage $G_{X\to Y}$ and $G_{Y\to X}$ to find (x, y) pairs with the same contextual information.

Full objective: The full objective is written with trade-off parameter λ_{cyc}:

$$\begin{aligned}\mathcal{L}_{full} = {} & \mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \mathcal{L}_{adv}(G_{Y\to X}, D_X) \\ & + \lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}). \quad (3)\end{aligned}$$

B. CycleGAN for Non-parallel VC: CycleGAN-VC

To use a CycleGAN for non-parallel VC, we mainly made two modifications to the CycleGAN architecture: gated CNN [25] and identity-mapping loss [26].

Gated CNN: One of the characteristics of speech is that it has sequential and hierarchical structures, e.g., voiced/unvoiced segments and phonemes/morphemes. An effective way to represent such structures would be to use an RNN, but it is computationally demanding due to the difficulty of parallel implementations. Instead, we configure a CycleGAN using gated CNNs [25] that not only allow parallelization over sequential data but also achieve state-of-the-art in language modeling [25] and speech modeling [8]. In a gated CNN, gated linear units (GLUs) are used as an activation function. A GLU is a data-driven activation function, and the $(l+1)$-th layer output $\boldsymbol{H}_{l+1}$ is calculated using the l-th layer output $\boldsymbol{H}_l$ and model parameters $\boldsymbol{W}_l$, $\boldsymbol{V}_l$, $\boldsymbol{b}_l$, and $\boldsymbol{c}_l$,

$$\boldsymbol{H}_{l+1} = (\boldsymbol{H}_l * \boldsymbol{W}_l + \boldsymbol{b}_l) \otimes \sigma(\boldsymbol{H}_l * \boldsymbol{V}_l + \boldsymbol{c}_l), \quad (4)$$

where $\otimes$ is the element-wise product and σ is the sigmoid function. This gated mechanism allows the information to be selectively propagated depending on the previous layer states.

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 information. To encourage linguistic-information preservation

ISBN 978-90-827970-1-5 © EURASIP 2018

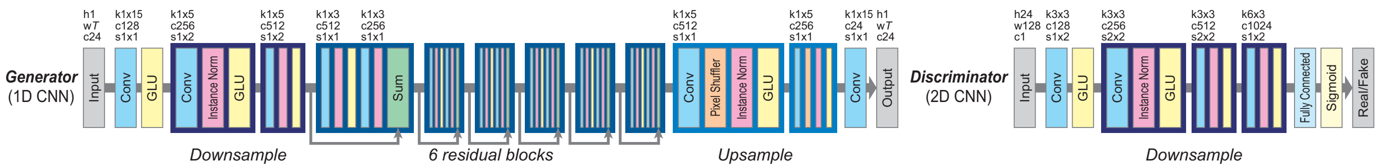

Fig. 2. Network architectures of generator and discriminator. In input or output layer, h, w, and c represent height, width, and number of channels, respectively. In each convolutional layer, k, c, and s denote kernel size, number of channels, and stride size, respectively. Since generator is fully convolutional [44], it can take input of arbitrary length T.

without relying on extra modules, we incorporate an identity-mapping loss [26],

$$\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{y\sim P_{\text{Data}}(y)}[\|G_{X\to Y}(y) - y\|_1] + \mathbb{E}_{x\sim P_{\text{Data}}(x)}[\|G_{Y\to X}(x) - x\|_1], \quad (5)$$

which encourages the generator to find the mapping that preserves composition between the input and output. In practice, weighted loss $\lambda_{id}\mathcal{L}_{id}$ with trade-off parameter λ_{id} is added to Eq. 3. Note that the original study on CycleGANs [22] showed the effectiveness of this loss for color preservation.

IV. Experiments

A. Experimental Conditions

We evaluated our method on a non-parallel VC task using the VCC 2016 dataset [32], which was recorded by professional US English speakers, including five females and five males. Following a previous study [41], we used a subset of speakers for evaluation. A pair of female (SF1) and male (SM1) speakers were selected as sources and another pair (TF2 and TM3) were selected as targets. The audio files for each speaker were manually segmented into 216 short parallel sentences (about 13 minutes). Among them, 162 and 54 sentences were provided as training and evaluation sets, respectively. To evaluate our method under a non-parallel condition, we divided the training set into two subsets without overlap. The first half 81 sentences were used for the source and the other 81 sentences were used for the target. This means that ***CycleGAN-VC* is trained under the disadvantageous condition (non-parallel and half the amount of data)**. The speech data were downsampled to 16 kHz, and 24 Mel-cepstral coefficients (MCEPs), logarithmic fundamental frequency ($\log F_0$), and aperiodicities (APs) were then extracted every 5 ms using the WORLD analysis system [45]. Among these features, we learned a mapping in the MCEP domain using our method. The F_0 was converted using logarithm Gaussian normalized transformation [46]. Aperiodicities were directly used without modification because a previous study [47] showed that converting APs does not significantly affect speech quality.

Implementation details: We designed a network based on the recent success in image modeling [22], [48], [49] and speech modeling [8], [29]. The network architecture is illustrated in Fig. 2. We designed the generator using a one-dimensional (1D) CNN [8] to capture the relationship among the overall features while preserving the temporal structure. Inspired by a previous study [49] for neural style transfer and super-resolution, we used the network that included downsampling, residual [50], and upsampling layers, as well as incorporating instance normalization [51]. We used pixel shuffler for upsampling, which is effective for high-resolution image generation [48]. We designed the discriminator using a 2D CNN [8] to focus on a 2D spectral texture [29].

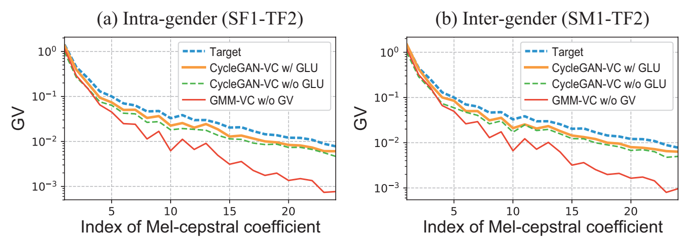

Fig. 3. Comparison of GV per MCEP. We omit *GMM-VC w/ GV* because it directly estimates GV.

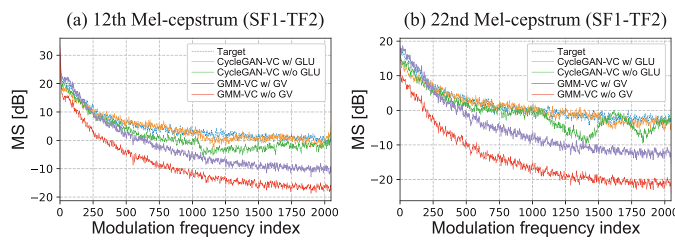

Fig. 4. Comparison of MS per modulation frequency

Training details: As a pre-process, we normalized the source and target MCEPs per dimension. To stabilize training, we used a least squares GAN [52], which replaces the negative log likelihood objective in $\mathcal{L}_{adv}$ by a least squares loss. We set $\lambda_{cyc} = 10$. We used $\mathcal{L}_{id}$ only for the first 10^4 iterations with $\lambda_{id} = 5$ to guide the learning process. To increase the randomness of each batch, we did not use a sequence directly and cropped a fixed-length segment (128 frames) randomly from a randomly selected audio file. We trained the network using the Adam optimizer [53] with a batch size of 1. We set the initial learning rates to 0.0002 for the generator and 0.0001 for the discriminator. We kept the same learning rate for the first 2×10^5 iterations and linearly decay over the next 2×10^5 iterations. We set the momentum term β_1 to 0.5.

B. Objective Evaluation

In these experiments, we focused on the conversion of MCEPs; therefore, we evaluated the quality of converted MCEPs. We compared our method (*CycleGAN-VC*) with a GMM-based parallel VC method (*GMM-VC*) [10]. Although *GMM-VC* is not a state-of-the-art method, we consider that it is a reasonable baseline because to the best of our knowledge, **there was still a gap between the quality of previous non-parallel VC and *GMM-VC***. Since this method requires parallel data, all the training data (162 sentences) for both source and target were used. This means that ***CycleGAN-VC* is trained under the disadvantageous condition (non-parallel and half the amount of data)**. As an ablation study, we examined *CycleGNA-VC* without GLUs. Instead of GLUs,

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE I
COMPARISON OF RMSE BETWEEN TARGET AND CONVERTED LOGARITHMIC MS AVERAGED OVER ALL MCEPS AND MODULATION FREQUENCIES [DB]. SMALL VALUE INDICATES CLOSENESS TO TARGET.

Method	SF1–TF2	SF1–TM3	SM1–TF2	SM1–TM3
CycleGAN-VC w/ GLU	**1.98**	**2.69**	**1.93**	**2.14**
CycleGAN-VC w/o GLU	3.34	2.99	3.17	2.94
GMM-VC w/ GV	7.59	9.41	8.69	9.67
GMM-VC w/o GV	13.56	14.90	14.17	14.53

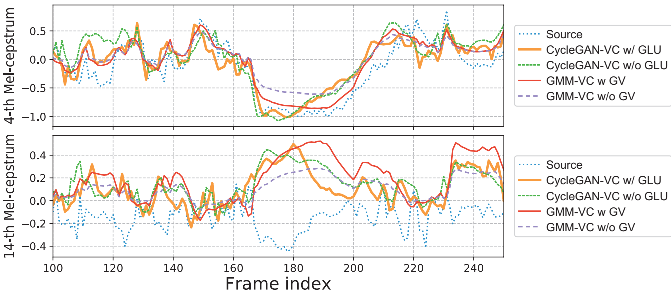

Fig. 5. Comparison of MCEP trajectories (SF1-TM3)

we used typical GAN activation functions, i.e., rectified linear units (ReLUs) [54] for the generator and leaky ReLUs [55], [56] for the discriminator. In the pre-experiment, we also examined *CycleGAN-VC* without an identity-mapping loss. This revealed that the lack of this loss tends to cause significant degradation, e.g., collapse of the linguistic structure.

Mel-cepstral distortion is a well-used measure to evaluate the quality of synthesized MCEPs, but recent studies [10], [29], [43] indicate the limitation of this measure: it tends to prefer over-smoothing because it internally assumes Gaussian distribution. Therefore, as alternatives, we used two structural indicators highly correlated with subjective evaluation: GV [10] and MS [33]. We show the comparison of GV in Fig. 3. We list the comparison of root mean squared error (RMSE) between target and converted logarithmic MS in Table I. We also show the comparison of MS per modulation frequency in Fig. 4. These results indicate that the MCEP sequences obtained with *CycleGAN-VC w/ GLU* are closest to the target in terms of GV and MS. We expect this is because (1) the adversarial loss contributes to avoiding over-smoothing, and (2) the GLU succeeds in representing sequential and hierarchical structures better than the ReLU and leaky ReLU. We show sample MCEP trajectories in Fig. 5. The trajectories of *CycleGAN-VC w/ GLU* have a similar global structure to those of *GMM-VC w/ GV* while preserving similar complexity to the source.

C. Subjective Evaluation

We conducted listening tests to evaluate the performance of converted speech[2]. By referring to the VCC 2016 [57], we evaluated the naturalness and speaker similarity of the converted samples. We compared our method with the official baseline of the VCC 2016 (*VCC-VC*)[3]. This is a GMM-based method [10] and trained using all the training data. To measure naturalness, we conducted a mean opinion score (MOS) test. As a reference, we used original (*Org*) and re-synthesized (*Resyn*; upper bound of our method) speeches of

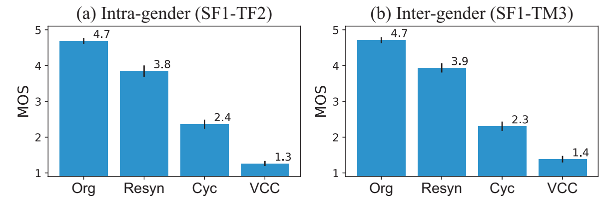

Fig. 6. MOS for naturalness with 95% confidence intervals (Org: *Original*, Resyn: *Re-synthesized*, Cyc: *CycleGAN-VC* (proposed), and VCC: *VCC-VC*)

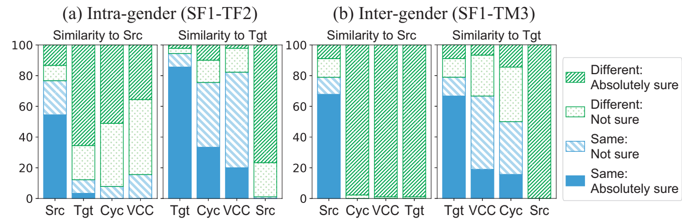

Fig. 7. Similarity to source speaker and to target speaker (Src: *Source*, Tgt: *Target*, Cyc: *CycleGAN-VC* (proposed), and VCC: *VCC-VC*)

target speakers. Twenty sentences longer than 2 s and shorter than 5 s were randomly selected from the evaluation sets. To measure speaker similarity, we used the same/different paradigm [57]. Ten sample pairs were randomly selected from the evaluation sets. There were nine participants who were well-educated English speakers. By referring to the study by [43], we evaluated on two subsets: intra-gender VC (*SF1–TF2*) and inter-gender VC (*SF1–TM3*). We show the MOS for naturalness in Fig. 6. The results indicate that *CycleGAN-VC* significantly outperformed *VCC-VC*. We show the similarity to a source speaker and to a target speaker in Fig. 7. The results indicate that *CycleGAN-VC* was slightly inferior to the *VCC-VC* in *SF1–TM3* but superior in *SF1–TF2*. Overall, *CycleGAN-VC* is comparable to *VCC-VC*. This is noteworthy because ***CycleGAN-VC* is trained under disadvantageous conditions (non-parallel and half the amount of data)**.

V. DISCUSSION AND CONCLUSIONS

We proposed a non-parallel VC method called *CycleGAN-VC*, which uses a CycleGAN with gated CNNs and an identity-mapping loss. This method can learn a sequence-based mapping function without any extra data, modules, and time alignment procedure. An objective evaluation showed that the MCEP sequences obtained with *CycleGAN-VC* are close to the target in terms of GV and MS. A subjective evaluation showed that the quality of converted speech was comparable to that obtained with the GMM-based parallel VC method even though *CycleGAN-VC* was trained under disadvantageous conditions (non-parallel and half the amount of data). However, there is still a margin between original and converted speeches. To fill the margin, we plan to apply our method to other features, such as STFT spectrograms [30], and other speech-synthesis frameworks, such as vocoder-free VC [58]. Furthermore, our proposed method is a general framework, and possible future work includes applying the method to other VC applications [2]–[8].

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 17H01763.

[2]We provide the converted speech samples at http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc/

[3]We used data at http://dx.doi.org/10.7488/ds/1575

REFERENCES

[1] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," in *arXiv preprint arXiv:1711.11293*, Nov. 30, 2017.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp. 285–288.

[3] A. B. Kain, J.-P. Hosom, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Commun.*, vol. 49, no. 9, pp. 743–759, 2007.

[4] 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.

[5] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 20, no. 9, pp. 2505–2517, 2012.

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

[7] O. Türk and M. Schröder, "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 965–973, 2010.

[8] 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.

[9] 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.

[10] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[11] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 912–921, 2010.

[12] 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.

[13] 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.

[14] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 954–964, 2010.

[15] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *Proc. SLT*, 2014, pp. 19–23.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, "On the impact of alignment on voice conversion performance," in *Proc. Interspeech*, 2008, pp. 1453–1456.

[22] 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.

[23] 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.

[24] 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.

[25] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, 2017, pp. 933–941.

[26] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. ICLR*, 2017.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NPIS*, 2014, pp. 2672–2680.

[28] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3D-guided cycle consistency," in *Proc. CVPR*, 2016, pp. 117–126.

[29] 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.

[30] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for STFT spectrograms," in *Proc. Interspeech*, 2017, pp. 3389–3393.

[31] Y. Saito, S. Takamichi, and H. Saruwatari, "Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis," in *Proc. ICASSP*, 2017, pp. 4900–4904.

[32] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *Proc. Interspeech*, 2016, pp. 1632–1636.

[33] 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.

[34] M. Zhang, J. Tao, J. Tian, and X. Wang, "Text-independent voice conversion based on state mapped codebook," in *Proc. ICASSP*, 2008, pp. 4605–4608.

[35] 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.

[36] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 14, no. 3, pp. 952–963, 2006.

[37] 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.

[38] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *Proc. Interspeech*, 2006, pp. 2446–2449.

[39] 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.

[40] 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.

[41] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. APSIPA ASC*, 2016, pp. 1–6.

[42] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, "Non-parallel voice conversion using i-vector PLDA: Towards unifying speaker verification and transformation," in *Proc. ICASSP*, 2017, pp. 5535–5539.

[43] 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.

[44] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2015, pp. 3431–3440.

[45] 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.

[46] 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.

[47] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. Interspeech*, 2006, pp. 2266–2269.

[48] 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.

[49] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[51] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," in *arXiv preprint arXiv:1607.08022*, July 27, 2016.

[52] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. ICCV*, 2017.

[53] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[54] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.

[55] A. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML Workshop*, 2013.

[56] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," in *Proc. ICML Workshop*, 2015.

[57] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the Voice Conversion Challenge 2016 evaluation results," in *Proc. Interspeech*, 2016, pp. 1637–1641.

[58] K. Kobayashi, T. Toda, and S. Nakamura, "F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential," in *Proc. SLT*, 2016, pp. 693–700.

ISBN 978-90-827970-1-5 © EURASIP 2018