CycleGAN Voice Conversion of Spectral Envelopes using Adversarial Weights

Rafael Ferro, Nicolas Obin, Axel Roebel

CycleGAN Voice Conversion of Spectral Envelopes using Adversarial Weights

Rafael Ferro
IRCAM, CNRS, Sorbonne Université
STMS Lab
Paris, France
rafael.ferro@ircam.fr

Nicolas Obin
IRCAM, CNRS, Sorbonne Université
STMS Lab
Paris, France
nicolas.obin@ircam.fr

Axel Roebel
IRCAM, CNRS, Sorbonne Université
STMS Lab
Paris, France
axel.roebel@ircam.fr

***Résumé*—This paper tackles GAN optimization and stability issues in the context of voice conversion. First, to simplify the conversion task, we propose to use spectral envelopes as inputs. Second we propose two adversarial weight training paradigms, the generalized weighted GAN and the generator impact GAN, both aim at reducing the impact of the generator on the discriminator, so both can learn more gradually and efficiently during training. Applying an energy constraint to the cycleGAN paradigm considerably improved conversion quality. A subjective experiment conducted on a voice conversion task on the voice conversion challenge 2018 dataset shows first that despite a significantly reduced network complexity, the proposed method achieves state-of-the-art results, and second that the proposed weighted GAN methods outperform a previously proposed one.**

***Index Terms*—voice conversion, cycleGAN, GAN stability, adversarial weights**

I. Introduction

A. Related works

Voice identity conversion (VC) consists in modifying the voice of a source speaker so as to be perceived as the one of a target speaker. Over the past few years, VC has largely gained in popularity and in quality [1], [2], in particular with the development of neural voice conversion algorithms [3], [4]. VC consists in learning a conversion function between the acoustic space of a source and a target speaker. This conversion function is generally learned from a pre-aligned database (parallel VC) in which the source and the target speakers pronounce the same set of sentences, so that a direct correspondence between the frames of the source and target speakers can be established. Unfortunately, this constraint reduces the amount of available recordings of source and target speakers.

Modern VC is mainly based on neural architectures with the particular objective to extend the VC from parallel to non-parallel speech databases. The main advantage of non-parallel VC is that it provides the flexibility to learn the conversion from "on-the-fly" speech databases, which can more easily handle large amount of data and accommodate multiple speakers. These architectures includes Variational Autoencoders (VAEs) [5]–[7], Generative Adversarial Networks (GANs) [8]–[12], Phonetic PosteriorGrams (PPGs) [13] and sampleRNNs [14] among others. The use of GAN architectures [15] for VC is inspired by advances conducted in the fields of image generation and manipulation. The cycleGAN is a particular configuration of GANs which has been specifically formulated to learn transformations between two different domains or between unaligned or unpaired datasets with application to image-to-image translation [16].

The cycleGAN-VC is the extension of the cycleGAN to the VC task, which has become a standard in non-parallel VC [8], [10]. The main idea behind the application of the cycleGAN to VC is that the cycle-consistency encourages the preservation of the phonetic content through the cycle while learning to modify the speaker identity. Despite the advances accomplished in non-parallel VC, cycleGAN-VC still suffers from important limitations which conduct to conversion of mitigated quality. One main limitation is due to the well-known stability issues of the GAN [17], [18]. This issue, combined to the limited amount of data available in the VC task, can lead to severe degradation of the voice conversion quality and the naturalness of the converted speech.

Tackling GAN stability issues has motivated the development of multiple ideas and heuristics [17], such as the popular Wasserstein GAN [19] or the LSGAN [20]. More recently, so as to tackle stability issues, the weighted GAN has been introduced [21]. This novel approach, based on game theory, instead of equally weighted "fake" samples, more attention is given to samples that fool the discriminator. A particular form of weighted GAN has been recently applied to VC. [22].

B. Contribution of the paper

This paper proposes two contributions. First, we propose to use spectral envelopes as inputs instead of using cepstral coefficients as in [10]. The use of the spectral envelope is assumed to simplify the task for the convolutional networks that will be required to only slightly move the spectral formants. Thus we assume to not need an extremely deep network and we will show that we can achieve similar performance with a significantly smaller network. Second, we propose to introduce a constraint as an additional training loss that enforces to preserve energy contour of the converted speech signal. This allows us to add a generator loss not depending on adversarial training, which is expected to contribute to the stability of

The research in this paper has been funded by the ANR project TheVoice : ANR-17-CE23-0025

the training. Third, we present a novel method to tackle the stability issue of GAN training, exploring a novel weighted GAN approach. We achieve this by adding a weight to the loss of the discriminator, giving more weight to "true" samples rather than to "fake" ones.

II. Preliminary Works on CycleGAN VC

A. Generative Adversarial Networks

A Generative Adversarial Network (GAN) [15] is a neural network system composed by a generator G and a discriminator D, in which the discriminator is trained to discriminate real samples from generated samples, while the generator is trained to generate real-like samples, using the discriminator as the decision rule. The objective can be written as :

$$\begin{aligned}\min_G \max_D \mathcal{L}(D,G) =& \mathbb{E}_{x\sim p(X)}[\log D(x)] \\ &+\mathbb{E}_{z\sim p(Z)}[\log(1-D(G(z)))]\end{aligned} \tag{1}$$

where : x is a sample from a distribution $p(X)$ to be modeled, z is a sample generated from a random distribution $p(Z)$, and $\mathbb{E}_{x\sim p(X)}$ represents the expected value of x given the distribution $p(X)$.

B. Cycle Generative Adversarial Networks

In the cycleGAN architecture [16], [23], a generator $G_{X\to Y}$ reads data from a dataset X and learns to map it into its respective position in a dataset Y, and vice versa for a generator $G_{Y\to X}$. If X and Y represent languages, this system should be analogous to two translators. To train these generators, the cycleGAN framework uses two adversarially trained discriminators to discriminate respectively any $x \in X$ in relation to $G_{Y\to X}(y)$ for any $y \in Y$ and any $y \in Y$ in relation to $G_{X\to Y}(x)$ for any $x \in X$. Since $G_{Y\to X}(G_{X\to Y}(x))$ should be equal to x, and $G_{X\to Y}(G_{Y\to X}(y))$ should be equal to y, a loss named cycle-consistent loss is added to enforce this constraint. In the following equations we state the total objective of the cycleGAN, where $\mathbb{E}_{y\sim P_{Data}(y)}$ represents the expected value for the distribution Y and $\mathbb{E}_{Y\sim P_{Data}(X)}$ represents the expected value for the distribution X.

The following equation describes the adversarial loss for the discriminator D_Y (the equation for D_X is analogous) :

$$\begin{aligned}\mathcal{L}_{adv}(G_{X\to Y},D_Y) = &\mathbb{E}_{y\sim P_{Data}(y)}[\log(D_Y(y))] \\ &+\mathbb{E}_{x\sim P_{Data}(x)}[\log(1-D_Y(G_{X\to Y}(x)))].\end{aligned} \tag{2}$$

The following equation describes the cycle-consistency loss, using L1 norm :

$$\begin{aligned}\mathcal{L}_{cyc}(G_{X\to Y},G_{Y\to X}) = & \\ &\mathbb{E}_{x\sim P_{Data}(x)}[||G_{Y\to X}(G_{X\to Y}(x))-x||_1] \\ &+\mathbb{E}_{y\sim P_{Data}(y)}[||G_{X\to Y}(G_{Y\to X}(y))-y||_1].\end{aligned} \tag{3}$$

The following equation describes the total objective of the cycleGAN, where λ_c represents the weight for the cycle-consistency loss :

$$\begin{aligned}\mathcal{L}_{full} =& \mathcal{L}_{adv}(G_{X\to Y},D_Y) + L_{adv}(G_{Y\to X},D_X) \\ &+\lambda_c L_{cyc}(G_{X\to Y},G_{Y\to X})\end{aligned} \tag{4}$$

The cycleGAN-VC, introduced by [10], is trained to convert a source speaker Mel Frequency Cepstral Coefficients (MFCCs) into a target speaker MFCCs, so as to perform VC. Their discriminators task is therefore to discriminate whether the conversions belong to their respective target speaker identity or not. In particular, so as to adapt the original cycleGAN framework, they used Gated CNNs as well as an identity-mapping loss, which is reported to encourage phonetic invariance. However, we did not find these two additional ideas beneficial.

III. GAN with Adversarial Weights

Generative Adversarial Networks are reported to be difficult to train. One problem are vanishing gradients when the discriminator achieves perfect discrimination, or when the generator is able to perfectly fool the discriminator, though it is producing nonsense. Another problem is the instability that is due to the fact that the discriminator is trained to systematically reject generated examples independent of their quality. In the case when the generator generates target samples covering only a small part of the target space the discriminator will improve its objective by means of pushing the generator out of the target space even if it has to wrongly classify some of the real samples as well. As a result the discriminator will push the generator away from the target space hindering the generator to converge.

To solve these issues, many ideas have been proposed, such as the DCGAN architecture [18]. [24] discussed mini-batch discrimination, historical averaging, one-sided label smoothing and virtual batch normalization. Also, new losses have been proposed, such as the Wasserstein GAN [19] and the LSGAN [20].

A. Weighted GAN System

Recently, so as to tackle CycleGAN optimization, Paul et al. developed the weStarGAN [22], implementing the weighted GAN idea [21] to the starGAN-VC architecture [8]. They do so by multiplying sample-wise a coefficient to the generator loss. Their idea is to give less weight to samples poorly produced by the generator, so that the generator has a stronger motivation to produce samples similar to the real data. For each sample j they compute its respective normalised weight $w_{j,g}$, while introducing a hyper-parameter η_{gen} :

$$\begin{aligned}w_{j,g} &= e^{\eta min(0,D_j)} \\ w_{j,g} &= \frac{w_{j,g}}{\sum_{j=1}^{m} w_{j,g}}, j = 1,...,m.\end{aligned} \tag{5}$$

This coefficient is then multiplied sample-wise by the generator loss. Note that samples that the discriminator does not correctly discriminate have a higher weight in the generator loss. A hypothesis assumed in [22] is that the discriminator is "faithful" : that quantitatively it returns on average values above 0.5 when the samples come from the real distribution and below 0.5 when fake samples are fed to the Discriminator. This structure is our starting point in our research on the stability and optimization of the GAN.

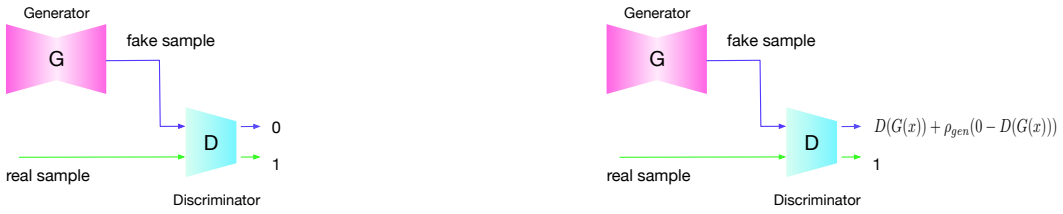

FIGURE 1. Comparison of GAN with and without generator impact. On left : vanilla GAN, on right : proposed GAN with weighted labels.

B. Generalized Weighted GAN

Since in practice the discriminator might not be "faithful", in order to encourage its "faithfulness", we propose to train the discriminator rather by fairly produced samples than by poorly produced ones. This helps to reduces the above mentioned instability caused by training the discriminator to judge generated data, no matter how good it is, as *equally* wrong. To do so, we further develop Paul et al. idea, by also applying weights to the discriminator loss. We followed the exact same procedure : we introduce a hyper-parameter η_{dis} and for each sample we compute its respective normalised weight $w_{j,d}$, following the same procedure as in Eq. (5). This coefficient is then multiplied sample-wise by the discriminator loss. Thus samples perceived as poor by the discriminator have less weighting in the discriminator loss. We name this system as generalized weighted GAN (geweGAN).

C. Generator Impact GAN

Finally, we introduce a novel weighted GAN technique, also encouraging the discriminator "faithfulness". In the proposed approach, the idea is to encourage the training of the discriminator by weighting samples produced by the generator by a constant generator impact hyperparameter ρ_{gen}, so that the discriminator is rather encouraged by the "real" distribution than by the generated one. We name this system the generator impact GAN (gimGAN), since we can control the impact of the generator on the discriminator loss. The idea is equivalent to replacing the "hard" label 0 scored by the discriminator on the samples produced by the generator $D(G(x))$, by a "soft" label calculated as :

$$D(G(x)) + \rho_{gen}(0 - D(G(x))) \in [0, 1] \tag{6}$$

where : $\rho_{gen}(0 - D(G(x)))$ represents the current distance of the discriminator relatively to the target "hard" label 0, weighted by generator impact hyperparameter ρ_{gen}. Once again, the motivation behind is to tackle the fact that in the vanilla GAN, the discriminator is trained to consider samples produced by the generator as "fake" ones, whatever its quality is, which might have the effect to discourage the generator to converge to the correct distribution. The principle is illustrated in Figure 1.

D. Energy constraint

Additionally, we further apply an energy constraint to the cycleGAN architecture. This constraint enforces preservation of the energy contour of the original source speech signal during conversion, and avoids incoherence between source and converted envelope in the converted speech signal. Furthermore, by means of providing stable feedback to the generator, this constraint is expected to reduce the instability of the GAN training. To achieve this, we impose a reconstruction loss on the amplitude mean for each frame, on both generators. Since we work with spectral envelopes as inputs, we just add the following term to the total loss, where λ_c represents the weight for the cycle-consistency loss :

$$\begin{aligned}\mathcal{L}_e =& \lambda_e \mathbb{E}_{x \sim P_{Data}(x)}[||\sum_{t=0}^{T} G_{X \to Y}(x) - x||_1] \\ &+ \lambda_e \mathbb{E}_{x \sim P_{Data}(x)}[||\sum_{t=0}^{T} G_{Y \to X}(y) - y||_1]\end{aligned} \tag{7}$$

This constraint enforces preservation of the energy contour of the original source speech signal during conversion, and avoids incoherence between source and converted envelope in the converted speech signal. Further, by means of providing stable feedback to the generator, this constraint is expected to reduce the instability of the GAN training.

IV. EXPERIMENT

The proposed CycleGAN architectures have been trained and evaluated, using the VCC2018 database [2]. The VCC2018 training corpus contains 80 short sentences per speaker, sampled at 16 kHz and quantified on 16 bits. For the evaluation set, we used the first 5 sentences, whose length was superior to 2s.

Our architecture was inspired by the DCGAN and by the cycleGAN-VC [10], [11]. For both generators, we used a two-layer encoder with convolutions, followed by a two-layer bottleneck with convolutions and then a two-layer decoder with transposed convolutions. For both the encoder and the decoder, we applied a kernel size 2 and a stride size 2, so as to avoid the checkerboard effect [25], noticing that consistently better results were obtained when this undesirable effect was avoided. For the bottleneck, we applied kernel size 3 and and stride size 1. The overall generator has respectfully 256, 512, 512 , 512, 256 and 1 filters. We applied instance

TABLE I
MOS AND 95% CONFIDENCE INTERVAL OBTAINED FOR THE DIFFERENT VC SYSTEMS.

	Male-to-Male (MTM)		Female-to-Female (FTF)		TOTAL	
Speech Signal Class	Similarity	Naturalness	Similarity	Naturalness	Similarity	Naturalness
orig : target	4.96 ± 0.09	4.91 ±0.12	5.00 ±0.00	5.00 ±0.00	4.98 ±0.04	4.96 ±0.05
conv : weGAN	2.52 ± 0.42	2.15 ±0.23	2.29 ±0.40	2.13 ±0.34	2.42 ±0.30	2.16 ±0.20
conv : geweGAN	2.80 ± 0.76	2.20 ±0.30	**3.32 ±0.35**	3.00 ±0.40	3.10 ±0.47	2.56 ±0.28
conv : gimGAN	3.21 ± 0.37	2.19 ±0.42	3.09 ±0.46	**3.41 ±0.36**	**3.15 ±0.30**	**2.81 ±0.33**
conv : cycleGAN-VC (baseline)	**3.33 ± 0.46**	**2.43 ±0.42**	2.75 ±0.33	2.06 ±0.28	3.05 ±0.31	2.27 ±0.27

normalization, followed by a ReLU at the end of each layer. Since spectral envelopes explicitly contain formant information, the task becomes much easier allowing us to implant a rather small generator, compared to traditional cycleGAN implementations. [10], [11] For both discriminators, we used 4 convolutional layers with a filter size 2, a kernel size 2, with respectfully 64, 128, 256 and 512 filters. These four layers were followed by two fully connected layers, with 512 and 1 neurons respectively. We applied instance normalization, followed by a LeakyReLU at the end of each layer, except for the last one. Our inputs were of size 32 and 128, for frequency bins and time frames respectively. We chose λ_c and λ_e to respectfully be 0.3 and 1. By lowering the traditional values for the cycle-consistency loss, we rather force GAN learning than cycle-consistency learning. This means that more weight is given to the identity learning, the task of the GAN, than to the phoneme reconstruction learning, the cycle-consistency task. In fact, otherwise results were found to be closer to a reconstruction, rather than a conversion. We applied a batch size 1 and we used least squares error for the discriminator loss, introduced in the LSGAN [20] and optimized it with the Adam algorithm, as in [10], [11] for a total of 800k iterations, with a generator learning rate of 0.0002 and a discriminator learning rate of 0.0001. We implemented three GAN variations that all share the same network structure and input representation. The first variation, uses the Weighted GAN paradigm (weGAN) from [22] with $\eta_{gen} = 0.1$. For the second variation, we implemented the generalized weighted GAN paradigm (geweGAN) for both the discriminator and the generator losses, as discussed above, with $\eta_{gen} = 0.9$and $\eta_{dis} = 0.9$. For the third variation, we implemented the generator impact weighted GAN paradigm (gimGAN), as discussed above, with $\rho_{gen} = 0.9$. Finally, we also implemented cycleGAN-VC [10] as a baseline.

Similarly to previous research on VC, the proposed VC is focused on spectral voice conversion only. The VC is based on a source/filter decomposition of the speech signal, in which the excitation of the source speaker is preserved during conversion and only the spectral envelope conversion is learned and modified. The analysis/synthesis engine relies on superVP, an extended phase vocoder developed by IRCAM[1]. The spectral envelope is estimated from the short-term Fourier transform (STFT) by using the True Envelope algorithm [26]. The Mel spectral envelope is then computed by integrating the estimated spectral envelope over 32 Mel filters in which the energy of each Mel filter is normalized to unity.

1. www.forumnet.ircam.fr/product/supervp-max-en/

A. Experimental setups

The experiment consisted into the judgment by listeners of singing voice samples, based on the similarity to the target singer and the naturalness of the singer, as used for the voice conversion 2018 challenge [2]. Conversion were processed for all sentences contained in the test set. For the perceptual experiment, short excerpts were used and presented to the participants (around 5s.). We chose SF4 and TM4 as the source and TF3, TF4, TM3 and TM4 as the target. We evaluated the naturalness and speaker similarity of the converted samples, with a mean opinion score (MOS) test.

During the experiment, 15 short speech samples, original source and target speakers, and converted source-to-target speaker (each having duration of about 5s) were randomly selected from the test set, and presented to the participant in a random order. For each speech sample, the participant has the possibility to listen to an excerpt of the original target speaker. Then the participant is asked to rate the naturalness of the converted speech sample and its similarity to the target speaker. The experiment was conducted on-line, encouraging the use of headphones and quiet environment. 15 individuals participated in the experiment.

V. RESULTS AND DISCUSSION

The results of the perceptual evaluation are presented in table I. An overall result is that the original target speaker is consistently qualified to have high similarity and quality. In looking into the average results for all speaker pairs it is easy to see that the gimGAN and geweGAN variants perform significantly better than weGAN variant. This shows that the proposed GAN modifications used for training geweGAN and gimGAN have a positive impact on similarity and naturalness. Note that the energy constraint had an important contribution to these results. Without this constraint the similarity and naturalness rating are about 0.5 points lower (results not shown). Finally, the cycleGAN-VC baseline achieves approximately the same performance in similarity, while gimGAN and wegeGAN are perceived as more natural with gimGAN achieving overall the best results where differences in naturalness are more pronounced and differences in similarity remain marginal. Note however, that the geweGAN and gimGAN networks achieve this performance with about 3 times less parameters, which results in considerably shorter training and

inference times, which is expected to be a result of directly working on the spectral envelope. Looking into the sub groups for MTM and FTF conversions, one can notice that the best network changes with gender. While cycleGAN-VC baseline is best for MTM conversion, wegeGAN has been evaluated best in the FTF case. These differences in the two gender groups indicate a relatively strong dependency of the conversion performance on the speaker.

VI. Conclusion

The present paper investigates Voice Conversion with Generative Adversarial Networks (GAN), particularly with the cycleGAN paradigm, addressing optimization and stability issues. First, we use spectral envelopes as inputs. Second, so as to optimize and to address the stability issues of the GAN training we propose a generalization of the weighted GAN, the geweGAN, and a similar approach, the gimGAN. Our conducted experiment shows first that both proposed methods are able to score a better performance than the previously proposed weighted GAN. Second, it shows that the proposed method performs similarly to the cycleGAN-VC baseline on similarity and considerably outperforms it on quality. Furthermore, these results were achieved using a significantly smaller network, which significantly reduces training time. Finally, an additional energy constraint to the loss was found to be essential for similarity and naturalness learning. For future work, we plan to combine the proposed generalized weighted GAN with the generator impact GAN, so as to further improve the stability of the GAN training procedure.

Références

[1] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "The voice conversion challenge 2016," in *Interspeech*, 2016.

[2] 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 *Speaker Odyssey*, 2018.

[3] Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, and Nobukatsu Hojo, "AttS2s-VC : Sequence-to-Sequence Voice Conversion with Attention and Context Preservation Mechanisms," *arXiv :1811.04076 [cs, eess, stat]*, 2018.

[4] Hirokazu Kameoka, Kou Tanaka, Takuhiro Kaneko, and Nobukatsu Hojo, "ConvS2s-VC : Fully convolutional sequence-to-sequence voice conversion," *arXiv :1811.01609 [cs, eess, stat]*, 2018.

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

[6] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks," *arXiv :1704.00849 [cs]*, 2017, arXiv : 1704.00849.

[7] Wen-Chin Huang, Yi-Chiao Wu, Hsin-Te Hwang, Patrick Lumban Tobing, Tomoki Hayashi, Kazuhiro Kobayashi, Tomoki Toda, Yu Tsao, and Hsin-Min Wang, "Refined WaveNet Vocoder for Variational Autoencoder Based Voice Conversion," *arXiv :1811.11078 [cs, eess]*, 2018.

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

[9] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "StarGAN-VC2 : Rethinking Conditional Methods for StarGAN-Based Voice Conversion," *arXiv :1907.12279 [cs, eess, stat]*, July 2019.

[10] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks," *arXiv :1711.11293 [cs, eess, stat]*, 2017.

[11] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "CycleGAN-VC2 : Improved CycleGAN-based Non-parallel Voice Conversion," *arXiv :1904.04631 [cs, eess, stat]*, Apr. 2019.

[12] Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," *arXiv :1804.00425 [cs, eess, stat]*, 2018, arXiv : 1804.00425.

[13] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[14] Cong Zhou, Michael Horgan, Vivek Kumar, Cristina Vasco, and Dan Darcy, "Voice Conversion with Conditional SampleRNN," in *Interspeech*, 2018.

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *arXiv :1406.2661 [cs, stat]*, 2014.

[16] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[17] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved Techniques for Training GANs," *arXiv :1606.03498 [cs]*, 2016.

[18] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *arXiv :1511.06434 [cs]*, Nov. 2015.

[19] Martin Arjovsky, Soumith Chintala, and Leon Bottou, "Wasserstein Generative Adversarial Networks," *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[20] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley, "Least Squares Generative Adversarial Networks," *arXiv :1611.04076 [cs]*, 2016.

[21] Yannis Pantazis, Dipjyoti Paul, Michail Fasoulakis, and Yannis Stylianou, "Training Generative Adversarial Networks with Weights," *arXiv :1811.02598 [cs, stat]*, Nov. 2018, arXiv : 1811.02598.

[22] Dipjyoti Paul, Yannis Pantazis, and Yannis Stylianou, "Non-Parallel Voice Conversion Using Weighted Generative Adversarial Networks," in *Interspeech 2019*. Sept. 2019, pp. 659–663, ISCA.

[23] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong, "DualGAN : Unsupervised Dual Learning for Image-to-Image Translation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2868–2876.

[24] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved Training of Wasserstein GANs," *CoRR*, 2017.

[25] Augustus Odena, Vincent Dumoulin, and Chris Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016.

[26] Axel Röbel, Fernando Villavicencio, and Xavier Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1343–1350, 2007.