Zero-shot Voice Conversion with Instance Normalization by Generative Adversarial Networks

Naibo Zhai, Lipeng Song and Ruichen Zhang

***Abstract*—Recently, Voice Conversion(VC) has received a lot of attention and research, especially the zero-shot many-to-many voice conversion. Many VC systems have been proposed to implement zero-shot conversion using an autoencoder framework with extracted timbre and content encoders. However, they do not convert well when confronted with an unseen speaker. To solve the problem, we propose an MD-VC model which introduces multiple discriminators based on generative adversarial networks. It improves the conversion quality through adversarial training. Both objective and subjective comparison experiments show that our model indeed outperforms other models, including Auto-VC, VQVC, and AdaIN-VC, on unseen speakers.**

***Index Terms*—voice conversion, zero-shot, non-parallel, generative adversarial networks.**

I. Introduction

VOICE conversion refers to converting the source speaker's voice into the target speaker's voice, changing the speaker's characteristics in the voice without changing the speech content, and realizing voice-to-voice mapping. This technology plays an important role in the maintenance of speaker-identity modification, speaking assistance, and voice quality enhancement.

Some of the first work in voice conversion is done on speakers who are only present in the speech training set. In recent works, for example, a CycleGAN-VC [1]–[3] framework based on generative adversarial networks is proposed, which uses CycleGAN [4], an adversarial network with a gated convolutional neural network and identity mapping loss for voice conversion. StarGAN-VC [5]–[7] uses Star Generative Adversarial Networks [8] for non-parallel many-to-many voice conversion. In StarGAN-VC, the generator is trained using adversarial loss to ensure that the mapping between each pair of attribute domains preserves the content information and does not require any information about the input utterance attributes at the time of prediction.

There are issues in converting speakers who do not exist in the training set in previous work, and given the difficulty in getting a significant number of available speaker corpus in real life, voice conversion in the form of fewer samples has become a current research trend. Recently, AutoVC [11] is based on the idea of StyleGAN [12], which uses an autoencoder with the bottleneck between the reconstruction quality of speech and speaker identity deconfliction, thus achieving zero-shot voice conversion. It is worth mentioning that AutoVC is one of the first works to achieve zero-shot conversion. AdaIN-VC [13] is based on the variational autoencoder (VAE) [14] framework, which uses the instance normalization (IN) strategy to achieve voice conversion between speakers. IN is a technique used in computer vision, but it has also been demonstrated to be utilized in the field of voice conversion. However, both AutoVC and AdaIN-VC are limited in their ability to synthesize audio quality due to the limitations of the models themselves. Furthermore, the stability of the pre-trained speaker encoder is doubted by some research [15]. Hence, analyzing the influences of pre-training on audio quality is of great importance.

Inspired by [16], we introduce a generative adversarial network with multiple discriminators based on autoencoder to achieve the zero-shot conversion, which is capable of performing non-parallel many-to-many conversions for speakers that do not appear in the training set. It takes speaker embedding as an input, untangles the speaker identity information and content by instance normalization in the content decoder, then synthesizes the converted speech by a generator. To improve the quality of synthesis, we use multiple discriminators and train adversarially on the reconstructed samples and their timbres in generative adversarial networks. In our empirical study, we demonstrate that it performs well when compared with the current state-of-the-art voice conversion schemes. Our main contributions are as follows:

- We introduce a multi-discriminator model based on generative adversarial networks which to further improve audio quality through adversarial training.
- We investigate and assess the effect of the pre-trained speaker encoder on our voice conversion model by modifying the encoder. Our experiments show that pre-training has indeed improved audio quality while its influence on stability is trivial.

The rest of the paper is organized as follows. In Section II, we introduce our framework architecture. In Section III, the model parameters, experimental setup, and results are presented, and Section IV concludes the paper.

II. Proposed Method

As shown in Fig. 1., the proposed framework has four modules: Speaker Encode E_T, which extracts arbitrary speaker embeddings; Content encoder E_C, which encodes the content; Generator G, which combines timbre and content to generate new speech; and Discriminator D, as shown in Fig. 2., has

Manuscript submitted on December 8, 2022. This work is supported in part by the National Key Research and Development Program of China under Grant 2021YFC3320100. *(Corresponding author: Lipeng Song.)*

The authors are with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China (e-mail: zhain218@gmail.com; slp880@sdu.edu.cn; zrc020413@gmail.com)

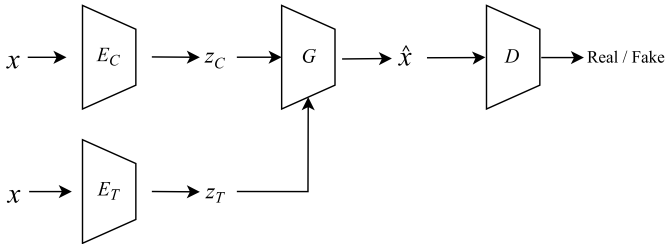

Fig. 1. The proposed zero-shot voice conversion framework.

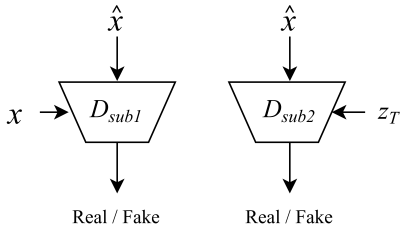

Fig. 2. The discriminator D is composed of two subparts, D_{sub1} and D_{sub2}, respectively.

two sub-discriminators, and the generated sample $\hat{x}$ will be delivered to both parts, D_{sub1} is used to discriminate real speech and generated speech, and D_{sub2} will extract the speaker embedding of generated speech and also discriminate the authenticity of timbre.

A. The Variational AutoEncoder Framework

The goal of VAE is to construct a model from the hidden variable z to the target data x, i.e., to construct $x = g(z)$, such that the learned target data has a probability distribution similar to the real data. The reconstruction loss and the KL divergence loss are typically included in VAE. Generator D will attempt to reconstruct the input Mel spectrogram x. The Reconstruction loss of its generator D for input can be written as:

$$L_{rec} = \mathbb{E}\left[\|\hat{x} - x\|_1^1\right] \tag{1}$$

The KL divergence loss is commonly used to constrain the distribution of speech content, and it can be written as:

$$L_{kl} = \frac{1}{2}\sum_{i=1}^{N}\left(\mu_i^2 + \sigma_i^2 - \log\left(\sigma_i^2\right) - 1\right) \tag{2}$$

where μ and σ are the mean and covariance of the content encoder output, respectively, N is the dimension of μ and σ.

B. The Generative Adversarial Networks Framework

Inspired by [16], we introduce a multi-discriminator approach to further improve speech quality. The discriminator consists of sub-discriminators, the first one is used to discriminate the authenticity of samples and the second one is used to discriminate the authenticity of timbre. For brevity, we use discriminator D to represent the two subparts uniformly. Then the discriminator loss of the GAN part can be defined as:

$$L_D = -\mathbb{E}\left[\|logD(x)\|\right] - \mathbb{E}\left[\|log(1 - D(\hat{x}))\|\right] \tag{3}$$

Mel-Spectrogram Loss In addition to the GAN objective, we try to promote the discriminator and generator effects of D_{sub1} and G, respectively, by adding Mel-spectrogram loss. In this paper, the Mel-spectrogram loss uses the L2 distance of the Mel-spectrogram between the real data and the generated data. This loss can be written as:

$$L_{mel} = \mathbb{E}\left[\|x - f_1(D_{sub1}(x))\|_2^2\right] \tag{4}$$

where f_1 is the output of the last layer of the sub-discriminator D_{sub1} network model.

Timbre Feature Matching Loss Similarly, for D_{sub2}, we try to improve the speaker similarity in the speech by adding feature matching of timbre. L2 distance is also used to represent timbre feature matching loss. This loss can be written as:

$$L_T = \mathbb{E}\left[\|z_T - f_2(D_{sub2}(G(x)))\|_2^2\right] \tag{5}$$

where f_2 is the output of the last layer of the sub-discriminator D_{sub2} network model.

Finally, the total loss of the whole training is as follows:

$$L_{spk} = \lambda_{rec}L_{rec} + \lambda_{kl}L_{kl} + \lambda_{mel}L_{mel} + \lambda_T L_T + L_D \tag{6}$$

where λ_{rec}, λ_{kl}, λ_{mel} and λ_T are custom parameters.

C. Separation of speaker identity and content with instance normalization

In the model, the content encoder E_C can be implemented by adding IN without affine transform to remove speaker identity information and retain speech content [13]. Given an input sequence Q, IN calculates the channel mean μ and the standard deviation σ. The equation is as follows:

$$IN(Q) = \frac{Q - \mu(Q)}{\sigma(Q)} \tag{7}$$

In the generator G, we use the adaptive instance normalization (AdaIN) strategy, which extends the capability of IN to richly exploit the information in the speaker embedding. The adaptive instance normalization equation is as follows:

$$AdaIN(Q, S) = f(S)IN(Q) + g(S) \tag{8}$$

where S is the input speaker embedding, and $f(S)$ and $g(S)$ are linear transformations for each channel.

III. Experiments

To evaluate the effectiveness of the proposed model, experiments were conducted for objective and subjective evaluation. And it is compared with the current state-of-the-art models AutoVC [11], AdaIN-VC [13], and VQVC [17]. Our VC demo can be seen at https://septembern.github.io/GANDEMO/index.html

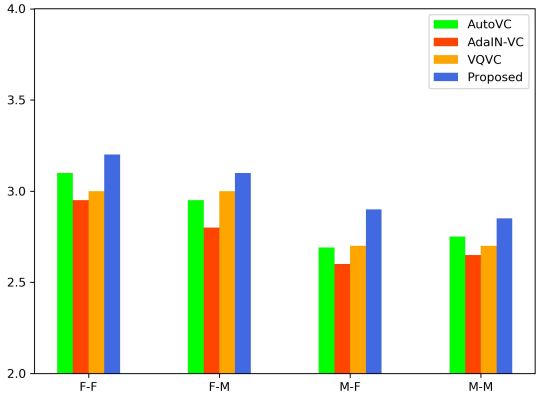

Fig. 3. MOS scores on similarity.

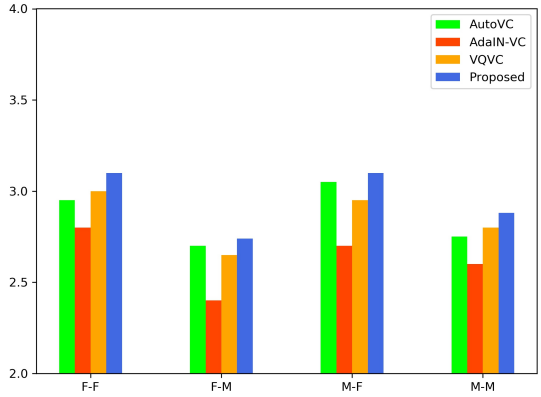

Fig. 4. MOS scores on audio quality.

A. DATASET

The evaluation is performed on the VCTK corpus, which contains 109 speakers with a total duration of about 44 hours and a sampling rate of 48kHz. For training, it is divided into training and test sets in a ratio of 9:1, and the data for each speaker is also divided by this ratio.

B. Model setup

Speaker Encoder E_T: The Speaker Encoder is used to extract speaker embedding, which is used to represent the unique features of each person. Considering the effect of whether to use pre-trained timbre encoders on the model effect, two approaches were used for our timbre encoders. The first one is a pre-trained model, similar to AutoVC [11], the speaker encoder is pre-trained on the GE2E loss [18]. It consists of two LSTM layers with a cell size of 768, using only the last linear output, and the generated speaker embedding is a 256×1 vector. The training set used is the VoxCeleb1 [19] and Librispeech [20] corpora. The second one will follow the whole framework along with the training, without using a pre-trained model, and tries to extract the speaker embedding. It starts with a ConvBank layer and subsequently consists of six residual blocks, each containing two convolutional layers, with an activation function using ReLU and a pooling operation after every two residual blocks.

Content Encoder E_C: The Content Encoder is used to extract the content and separate the identity information. In Content Encoder, the ConvBank layer is used first, followed by a 13-layer convolutional network with instance normalization, a pooling operation after every two convolutions after the first convolution, and an activation function of ReLU.

Generator G: The generator is used to synthesize content and timbre. In the generator, a convolutional network with instance normalization is used first, followed by a convolutional network residual block with two AdaIN layers, such that there are six residual blocks, using ReLU as activation.

Discriminator D: The discriminator is used to discriminate the authenticity of samples and timbre. In the first part of the structure of D_{sub1}, each layer consists of Conv, Instance Normalization, and GLU, four layers in total, and set the dropout of each layer to 0.1; in D_{sub2}, it consists of three layers of LSTM and one fully connected layer. Both output dimensions are 256×1.

For training, the input to the model is an 80-channel Mel-spectrogram, which is converted to a waveform by WaveNet vocoder. All input audio is downsampled to 16 kHz. λ_{rec}, λ_{kl}, λ_{mel} and λ_T in the parameters are 10, 0.5, 0.5, and 0.5, respectively. The ADAM optimizer has a learning rate of 0.0005, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set the batch size to 64.

C. Objective Evaluation

We use the Mel Cepstral Distortion (MCD) between the converted speech and the known target speech as an objective evaluation to assess the performance of the different models. The lower the MCD value, the better. Given that the MCD test requires both voices to say the same content, we chose 10 utterances in which the converted voice and the source speaker say the same phonetic content. The average of the test set is determined, and the particular scores are listed in Table. I.

TABLE I
MCD SCORES. COMPARISON OF DIFFERENT METHODS IN VC EVALUATION.

	AutoVC	AdaIN-VC	VQVC	Proposed
MCD	8.89	9.26	9.00	8.18

D. Subjective Evaluation

In this section, we use Mean-Opinion-Scores (MOS) to evaluate the quality and similarity of voice samples for the proposed model. The higher the MOS value, the better. We randomly select 10 speakers from the test set, including 5 males and 5 females. The speech is tested by pairwise conversions to produce $10 \times 9 = 90$ conversions, of which we selected 20 conversions as the evaluation set. The evaluation set contains four parts: Male-to-Male(M-M), Male-to-Female(M-F), Female-to-Male(F-M), and Female-to-Female(F-F), each with five samples. A total of 20 participants performed this assessment. In the quality test, participants rated the audio quality and similarity of an original and converted voice

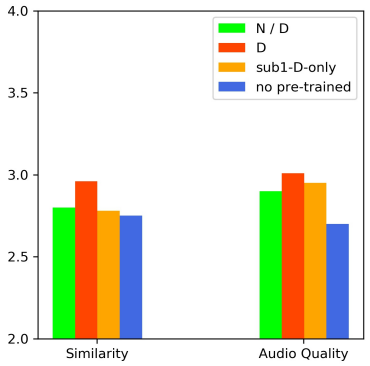

Fig. 5. MOS scores for different methods in terms of audio quality and similarity.

sample on a scale of 1 to 5. The MOS scores are shown in Fig. 3. and Fig. 4., respectively.

E. Analysis

We will examine the outcomes of the objective and subjective experiments in this section. According to the objective experiments, our MCD values are lowered by 0.71, 1.08, and 0.82 when compared to AutoVC, AdaIN-VC, and VQVC, respectively. While in the subjective evaluation, we increased the similarity by 5%, 10%, and 6%, respectively, compared to the three models. In terms of audio quality, our models improved by 3%, 10%, and 5%, respectively.

F. About Speaker Encoder

In this section, we will discuss the effect of the speaker encoder on the effect of model transformation. The two encoders we describe here correspond to the speaker encoders in Section III-B, while they are denoted by "D" and "no pre-trained" in Fig. 6 and Table. 5. We first explored the effectiveness of both in extracting speaker embedding from 10 unseen speakers, and both obtained greater than 97% accuracy, with the first at 98.5% and the second at 97.2%. Then, in the model, we conducted objective and subjective experiments on the conversion efficacy of both. As shown in Table. II, the MCD value of the model without the pre-trained encoder is 8.83, which is more than that of the model with the pre-trained encoder, which is 8.18. Meanwhile, we calculated the MCD values for 280 different audios, and their trends and ranges are shown in Fig. 6, with standard deviations of 0.78 and 0.82 for both, with a trivial difference. In the subjective experiments using MOS scores, as shown in Fig. 5, the model with the pre-trained speaker encoder surpasses the model without it in terms of the quality and similarity of the converted audio. Therefore, our model uses pre-trained speaker encoders, as discussed in Section III-G.

G. Discuss

In this section, the effect of no discriminator, single discriminator, and multi-discriminator on audio optimization will be discussed. We consider three methods: no discriminator, multiple discriminators, and just one discriminator (D_{sub1}). We conduct both subjective and objective experiments. As shown in Table. II, the difference in MCD values between the three situations in the objective experiment is minor, and it is difficult to say which model is superior. As shown in Fig. 5, we used MOS scores to rate audio quality and similarity in the subjective experiments, and the figure shows that the model with multiple discriminators has higher MOS scores in these two areas. As a result, the model with multiple discriminators is chosen as our final model.

TABLE II
MCD SCORES FOR DIFFERENT METHODS.

	N/D	D	D_{sub1}-only	no pre-trained
MCD	8.17	8.18	8.18	8.83

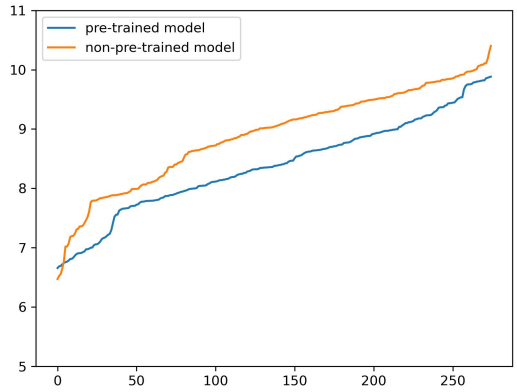

Fig. 6. Trend and range of MCD values for different types of speaker encoders.

IV. CONCLUSION

In this paper, we introduce a generative adversarial network with multiple discriminators based on an autoencoder to achieve zero-shot conversion. The model extracts speech content information through instance normalization and combines it with speaker embedding to generate new speech in the generator. Experimental results show that our proposed method further improves the audio quality under unseen speakers conversion in the VC tasks.

REFERENCES

[1] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100-2104.

[2] T. Kaneko, H. Kameoka, K. Tanaka and N. Hojo, "Cyclegan-VC2: Improved Cyclegan-based Non-parallel Voice Conversion," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6820-6824.

[3] Kaneko, Takuhiro, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. "CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-Spectrogram Conversion." *Proc. Interspeech 2020* (2020): 2017-2021.

[4] J. -Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242-2251.

[5] H. Kameoka, T. Kaneko, K. Tanaka and N. Hojo, "StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks," *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 266-273.

[6] Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2019). Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion. *arXiv preprint arXiv:1907.12279.*

[7] H. Kameoka, T. Kaneko, K. Tanaka and N. Hojo,"Nonparallel Voice Conversion With Augmented Classifier Star Generative Adversarial Networks," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2982-2995, 2020.

[8] Choi, Yunjey, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789-8797. 2018.

[9] 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," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1-6.

[10] Qian, Kaizhi, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. "Autovc: Zero-shot voice style transfer with only autoencoder loss." In *International Conference on Machine Learning*, pp. 5210-5219. PMLR, 2019.

[11] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401-4410. 2019.

[12] Chou, Ju-chieh, and Hung-Yi Lee. "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization." *Proc. Interspeech 2019* (2019): 664-668.

[13] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).

[14] Chen, Yen-Hao, Da-Yi Wu, Tsung-Han Wu, and Hung-yi Lee. "Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5954-5958. IEEE, 2021.

[15] Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae. "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis." *Advances in Neural Information Processing Systems* 33 (2020): 17022-17033.

[16] D. -Y. Wu and H. -y. Lee, "One-Shot Voice Conversion by Vector Quantization," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7734-7738.

[17] Wan, Li, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. "Generalized end-to-end loss for speaker verification." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4879-4883. IEEE, 2018.

[18] Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset." *arXiv preprint arXiv:1706.08612* (2017).

[19] Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. "Librispeech: an asr corpus based on public domain audio books." In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206-5210. IEEE, 2015.