Towards Unseen Speakers Zero-Shot Voice Conversion with Generative Adversarial Networks

Weirui Lu*, Xiaofen Xing†, Xiangmin Xu‡ and Weibin Zhang§
* South China University of Technology, College of Electronics and Information. E-mail: luweirui1022@gmail.com
† South China University of Technology, College of Electronics and Information. E-mail: xfxing@scut.edu.cn
‡ South China University of Technology, College of Electronics and Information. E-mail: xmxu@scut.edu.cn
§ Shenzhen VoiceAI Technology Co. Ltd. E-mail: eeweibin@gmail.com

***Abstract*—Zero-shot many-to-many voice conversion receives wide-spread attention but remains a challenging task. Recently, AUTOVC, which is based on conditional autoencoders, achieves state-of-the-art results in zero-shot voice conversion. However, the carefully designed bottleneck and the autoencoders based framework of AUTOVC limit further improvement of zero-shot voice conversion. In this paper, we propose a Generative Adversarial Network(GAN) based framework to disentangle the timbre and content of speech, and to synthesize new speech given unseen speakers and corpora. Towards unseen speaker, our framework extracts timbre embedding from an input speech with timbre encoder and produces content distribution embedding from any other speech with content encoder. Our framework learns to synthesize new speeches via conversion-reconstruction cycle training and to enhance the quality of conversion with adversarial training. Our experiments demonstrate that the proposed framework can generate outputs with comparable quality even for speakers that are not seen in the training dataset.**

I. Introduction

Voice conversion (VC) aims to convert the speech of one speaker to sound like another speaker without changing the linguistic content. This technique can solve a wide variety of tasks such as personalized text-to-speech system[1], speaking-aid device support[2] and accent conversion[3]. However, many-to-many zero-shot conversion tasks, *i.e*. converting from multiple source speakers to multiple target speakers where all the source and target speakers are not seen in the training data, are still challenging.

Previous works have attempted to convert speech only on seen speakers in the training corpus. Inspired by the success of image style transfer in computer vision, generative adversarial networks (GANs) have been proposed to achieve comparable performance with supervised training. For example, CycleGAN-Based methods[4], [5] perform comparably to a parallel VC method by incorporating the CycleGAN with a gated CNN and identity mapping loss. Further, StarGAN-based methods[6], [7] use a single generator to build non-parallel multi-speaker VC system. Other classical approaches to VC are based on variational autoencoders (VAEs)[8], [9], [10] and conditional variational autoencoders (C-VAEs)[11]. However, these methods can only deal with speakers that are seen in the training dataset. Given an unseen speaker, these VC systems require the collection of new training data in order to train new models and get fine results.

Recently, AUTOVC[12] and its F0-consistent version F0-AUTOVC[13], that are based on conditional autoencoders, achieved state-of-the-art results by disentangling the speaker timbre and speech content. They achieve zero-shot conversion by synthesizing a new voice with a new timbre embedding and the same speech content information from the source speakers. The disentanglement of the speech timbre and content is based on a carefully designed bottleneck. In addition, the secret of the success of AUTOVC lies in the dimension of the bottleneck. Wider bottleneck results bad disentanglement between the timbre and content of the speech from the source. Narrower bottleneck results in the loss of the content information. Although AUTOVC learns to match the distribution with the carefully designed bottleneck, it is hard to find a ideal dimention for the bottleneck. Furthermore, the carefully designed structure and AEs-based framework are not generative enough and limited to the further improvement of AUTOVC on zero-shot VC. Given new corpus with different domain, the structure requires to be carefully designed again. Otherwise, the mismatched structure may leads to a gap between the real target and the converted speech. In general, learning to match distribution directly with GANs is obviously a better choice if we set the sophisticated and difficult training of GANs aside.

More recent works of conditional GAN such as CVAE-GAN[14] combines a VAE with a GAN for fine-grained category image generation through asymmetric training. Further, the improvement of CVAE-GAN, IPGAN[16] changed the conditional input of CVAE-GAN to an identity embedding extractor to that, IPGAN can synthesis new samples with identities outside the training dataset. Based on the idea of CVAE-GAN and IPGAN, we attempt to extract timbre embedding from any specific speaker no matter the speaker is presented in the training corpus or not. The content distribution embedding extracted from a given source speech, together with the target speaker's timbre embedding, are utilized to synthesize new speech. The new speech has the same linguistic content as source speech but sounds like target speaker.

The proposed framework extracts timbre embedding with a LSTM-Based encoder and produces content distribution with a VAE-Based encoder. The proposed framework disentangle the timbre and content information through a GAN-Based conversion-reconstruction training. In the training process, adversarial loss are utilized to enhance the quality of the syn-

thesized result and two feature matching losses are utilized to make the training of GAN more stable. It achieves remarkable performance on many-to-many zero-shot conversion task.

The rest of the paper is organized as follows. In Section 2, we present the framework architecture and the training process of our system. Experimental setups and results are elaborated and discussed and presented in Section 3. Section 4 concludes the paper.

II. Methods

As shown in Fig 1, our framework takes mel-spectrogram as input and outputs converted mel-spectrogram. Our framework consists of five parts: **Timbre encoder** E_T, which extracts the speakers' timbre embedding of arbitrary speakers; **Content distribution encoder** E_C, which produces linguistic content distribution of any given speech; **Generator** G, which synthesizes a new speech with the combination of timbre and content distribution; **Timbre preserve network** TPN, which preserves the timbre of the generated speech; and **Discriminator** D, which distinguishes real and generated speech. These five parts are trained end-to-end.

In the subsequent sections, we first introduce the method to disentangle timbre and content components via encoder E_T and encoder E_C. In section 2.2, we discuss the loss functions used for GAN training in our framework. In section 2.3, we introduce the implementation of each part in details.

A. Disentanglement of Timbre and Content

1) Timbre Encoder E_T: The goal of the timbre encoder is to extract the a unique representation for an input speaker, regardless of the content of the speech. The one-hot encoding of speakers is convenient and efficient, and widely used in many-to-many voice conversion. However, the one-hot encoding does not contain the encoding of unseen speakers and thus is inapplicable to zero-shot conversion. In previous works, extracting the speakers timbre embedding is relatively straight-forward and the encoder can be implemented with speakers category annotations and classification training. Our speaker encoder is pre-trained by using the corpora of both VoxCeleb1[17] and Librispeech[18] with GE2E loss[15]. GE2E loss pushes the embedding towards the centroid of the true speaker, and away from the centroid if the speech comes from other speakers.

2) Content Encoder E_C: In order to train the content encoder E_c to encode content information only, we use conversion-reconstruction cycle training to extract the content distribution embedding. In the training process, different input data pairs produce two training situations: whether the content speech S_c and the timbre speech S_t comes from the same speaker or not. We attempt to balance two situations by using two loss function: reconstruction loss and KL divergence loss.

KL divergence loss: encoder E_C outputs the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\sigma}$ and the content distribution Z is sampled with reparameterization $Z = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathcal{I})$ and $\odot$ represents the element-wise multiplication. In order to train encoder E_C in a fully unsupervised manner to extract

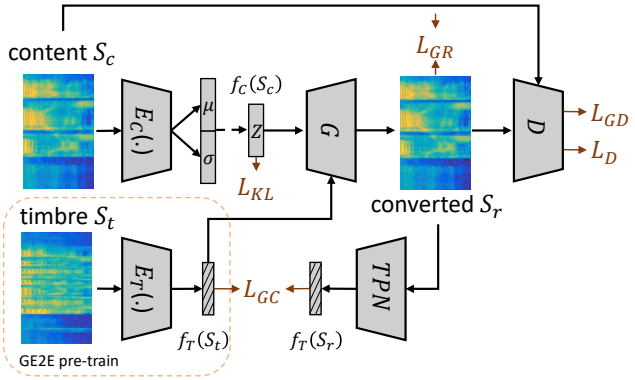

Fig. 1. The overview of framework. E_C extracts the content distribution f_C via VAE-Based framework. E_T extracts timbre embedding f_T as the conditional input of G. TPN shares the parameters with E_T and preserve timbre information by encouraging target S_c and converted S_r to have similar timbre representations. Discriminator D is used to enhance the quality of generated speech.

the content distribution embedding, KL loss is used to reduce the gap between the prior Gaussian distribution $\mathcal{N}(0, \mathcal{I})$ and the proposal distribution. The KL divergence loss will limit the distribution range of the content distribution embedding such that it does not contain much timbre information. As proved in [20], minimizing the KL divergence is equivalent to minimizing the following loss:

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

where N is the dimension of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$.

Reconstruction loss: After extracting the timbre embedding $f_T(S_t)$ of arbitrary speaker and content distribution embedding $f_C(S_c)$, a generator G take the combined embedding $\left[f_T(S_t)^T, f_C(S_c)^T\right]^T$ as input and synthesize a converted mel-spectrogram S_r. The synthesis process is supervised as follow:

$$L_{GR} = \begin{cases} \frac{1}{2} \left\| \boldsymbol{S_r} - \boldsymbol{S_c} \right\|_2^2, & \text{if reconstruction} \\ \frac{\lambda}{2} \left\| \boldsymbol{S_r} - \boldsymbol{S_c} \right\|_2^2, & \text{if conversion} \end{cases} \tag{2}$$

When the input timbre speech S_t is from the same speaker as the content speech S_c, the converted result is expected to be the same as speech S_c. In each reconstruction training batch, two input speeches have the same timbre embedding. As the content distribution are different in different reconstruction training batch, the reconstruction loss will force the content encoder E_c to learn different content representation $f_c(S_c)$.

When input timbre speech S_t and content speech S_c are from different speakers, we expect the converted result has the same content as S_c but the timbre of S_t. In other words, the converted result and content speech S_c are very similar, but not the same. Therefore, the pixel-wise reconstruction loss must have a relatively small weight λ to maintain the content attributes and leave space for speaker's embedding to change. In our work, λ is 0.1. In the training process, we define n

iterations as a conversion-reconstruction cycle, where each cycle contains 1 conversion and $n-1$ reconstruction. As the training continues, n was decreased from 5 down to 2.

B. Quality Enhancement with GAN

1) Adversarial Loss: The mel-spectrogram generated from VAE framework often suffers from over-smoothing. In order to make the conversion result sound more similar to real speech, we propose to use the adversarial loss with discriminator D on training. We define the adversarial loss as:

$$L_D = -\mathbb{E}_{S_c \sim P_c}[logD(S_c)] - \mathbb{E}_{z \sim P_z}[log(1 - D(G(z)))] \quad (3)$$

However, the training of traditional GAN is sophisticated and difficult as the distributions of real and fake may not overlap with each other. According to the theoretical analysis of recent works[21], distributing non-overlappings causes the gradient vanishing problem in the training of GANs. To address this problem, recent works[21] proposes a pairwise feature matching objective for the generator.

2) Feature Matching Loss: Inspired by [16], we use two feature matching loss to solve this problem and guarantee convergence of training. To generate realistic mel-spectrogram quality, we match the feature of the discriminator D of real and fake mel-spectrogram. We denote the feature on last Fully Connect layer of discriminator D as $f_D(S)$. The Euclidean distance between the feature representations is used to calculate loss L_{GD}, *i.e.* ,

$$L_{GD} = \frac{1}{2} \left\| \boldsymbol{f_D(S_r)} - \boldsymbol{f_D(S_c)} \right\|_2^2, \quad (4)$$

Meanwhile, timbre preserve network TPN shares parameters with encoder E_T and extracts the timbre of converted result S_r. In order to generate timbre-preserving mel-spectrogram, the timbre embedding of converted S_r must be similar to target input S_t. Therefore, we propose to minimize the Euclidean distance between two timbre embedding, *i.e.* ,

$$L_{GT} = \frac{1}{2} \left\| \boldsymbol{f_T(S_r)} - \boldsymbol{f_T(S_t)} \right\|_2^2, \quad (5)$$

Since the parameters of the timbre encoder E_T is freezed after the pre-training, L_{GT} force the generator G to generate timbre-preserved mel-spectrograms

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

$$L_{full} = \lambda_{kl} L_{KL} + \lambda_{gr} L_{GR} + \lambda_{gt} L_{GT} + \lambda_{gd} L_{GD} + L_D \quad (6)$$

,where λ_{kl}, λ_{gr}, λ_{gt}, and λ_{gd} are trade-off parameters, and L_{KL} and L_{GR} are determined by an additional parameter λ in Equation 2. In our work, λ_{kl} is 0.00005, λ_{gd} is 0.005, λ_{gr} and λ_{gt} are 1. Larger L_{KL} results in the over matching of the content distribution and explicit Gaussian distribution. Larger L_{GD} results in the loss of the content information as the importance of reconstruction loss L_{GR} is relatively small in training. Although λ_{kl} and λ_{gd} are small, these parameters balance the loss. Finally, the value of each loss converges to the same scale, about 10^{-4} to 10^{-3}. The training process with different trade-off parameters is shown in Fig 2.

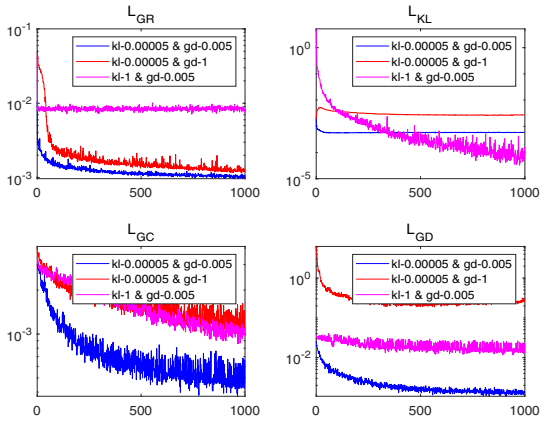

Fig. 2. Training on VCTK with different trade-off parameters. L_{GR} and L_{KL} are determined by parameter $\lambda = 0.1$ and only the conversion loss is shown on the figures. When $\lambda_{kl} = 0.00005$ and $\lambda_{gd} = 0.005$(blue), all losses are balanced and the value of each loss converges to the same scale, about 10^{-4} to 10^{-3}.

III. Experiments

To evaluate the effectiveness of the proposed framework for zero-shot conversion, we conducted objective and subjective evaluation experiments. The evaluation was performed on a public corpus VCTK, which contains 109 speakers with 200 to 500 pieces of utterances per speaker. We randomly selected ten speakers, 5 females and 5 males, as unseen speakers and train the system with the rest speakers. For each utterances, we randomly sampled 120 frames of spectrogram with overlap. All experiment conversion results are generated with unseen speakers. We compared the result with the state-of-the-art zero-shot baseline AUTOVC[12].

A. Experimental Setup

The input of the system is 80-channels mel-spectrogram with a 1024 window size and a 256 hop size and the output is converted to waveform by using WaveNet[19] vocoder. The configuration of WaveNet is the same as [12]. All the input speech is uniformly downsampled to 16kHz and thus the frame rate of the mel-spectrogram is 62.5Hz. The WaveNet vocoder is pre-trained on the VCTK corpus.

Similar to [12], the speaker encoder E_T consists of a stack of two LSTM layers with cell size 768 and the outputs of the last layer are projected down to 256 with a linear layer. The output f_T is normalized with L2-norm. The encoder E_C has three 5×1 1D-convolutional layers and a stack of two bidirectional LSTM layers with cell size 256. We sample all outputs of LSTM layers uniformly in temporal dimension. The sampling interval is 12. Then all the outputs feature are concatenated and fed into two sibling fully connected layers to generate the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\sigma}$. The dimensions of $\boldsymbol{\mu}$ and σ are 64. For the generator G, the content embedding vector and timbre embedding vector are concatenated and upsampled by a fully connnected layer. The features with original temporal resolution are fed to three 5×1 convolutional

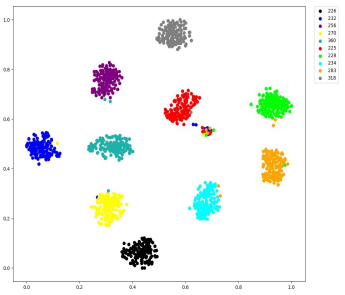

Fig. 3. The timbre embeddings are projected into two dimension with t-SNE for visualization. Each color represents one speaker from VCTK and the timbre encoder was not finetuned with the VCTK dataset.

	$E_T + n$	$n + E_C$	$E_T + E_C$
Class. Acc.	58.71	0.011	60.10
Recon. Err.	0.162	0.061	0.014

TABLE I
RECONSTRUCTION ERROR AND SPEAKER CLASSIFICATION ACCURACY ON VCTK DATASET. n DENOTES NOISE INPUT.

layers and a stack of three LSTM layers. Also similiar to [12], a post network is used to construct the generated details of the spectrogram better on top of the initial estimate. For the discriminator D, 2D-convolution are used to fully extract information. The converted results S_r are fed into five 5×2 2D-convolutional layers and a fully connected layer classify whether input the mel-spectrogram is real or not.

We trained the timbre encoder E_T for 80k iterations using the SGD optimizer with a momentum of 0.9 and an initial-learning rate of 0.001 that decays by 0.1 for iterations 60k and 70k. In each training iteration, we sampled 32 speakers, each with 25 utterances as a training batch. Then, we freezed the parameters of the encoder E_T and trained the rest of our framework for 20k iterations using ADAM optimizer with fixed a learning rate of 0.00005, β_1 of 0.5, β_2 of 0.999 and a batch-size of 64.

B. Objective Analysis

Firstly, we use t-SNE to visualize the timbre embedding and the 2-D features is plotted in Fig.3. It is worth noting that the timbre encoder was not finetuned with the VCTK dataset. We can observe that the timbre embedding can be easily separated although speakers are never seen in training set.

To evaluate the disentanglement of the timbre and content information, we evaluated the reconstruction error and speakers classification accuracy with different timbre and content input from the trained system. We use pixel-wise mean squared error (MSE) as evaluated metric of reconstruction error. As the timbre encoder of the proposed framework directly outputs the timbre embedding instead of classification category, we use K-Nearest Neighbor (KNN, $k = 100$) to classify the embedding. As shown in Table I, with the timbre embedding from real speech and noise content embedding, the proposed framework preserves enough information for accurate speaker classification. However, generates the worst reconstruction error in this case as content information is missing. With the content embedding from real speech and noise timbre embedding, our framework results in the worst speaker classification accuracy as the timbre information is lost. With the real timbre and real content from the same speech, we achieved the best classification accuracy and the minimum reconstruction error. The above experiments show that the encoder E_T and E_C effectively extract the timbre-independent and content-independent embedding separately.

	AUTOVC	w/o D	with D
MCD	5.73	5.22	5.61

TABLE II
COMPARISON OF MCD ON VCTK DATASET WITH/WITHOUT THE DISCRIMINATOR D.

To assess the quality of synthesized speech from our voice conversion system, we used the mel-cepstral distortion (MCD) to evaluate the outputs of the systems. MCD is a measure of how different two sequences of mel-cepstra are, with the smaller MCD being preferred. Given two mel-cepstra, $[x_1, ..., x_{24}]^T$ and $[y_1, ..., y_{24}]^T$, MCD is calculated as:

$$MCD[dB] = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(x_d - y_d)^2} \tag{7}$$

where 24 is the order of mel-cepstra. To evaluate effectiveness of the discriminator in our framework, we measured the MCDs of the converted speech of VC systems with/without the discriminator D. The results are shown in Table II. It can be seen that the proposed framework shows our framework with discriminator D achieves lower MCD scores than AUTOVC.

C. Qualitative Analysis

In this section, we perform two qualitative tests with 20 participants. We randomly choose 10 speakers from unseen data-set, 5 male and 5 female. Then, $10 \times 9 = 90$ conversions are produced by converting a test speech in pairs.We randomly choose 20 conversions as the evaluation set, including 5 male to male, 5 male to female, 5 female to male and 5 female to female. In conversions, timbre, content and converted speech in similarity and quality test, those three say the same text. We conduct Mean-Opinion-Score (MOS) evaluation on quality and similarity test of synthesized voice samples. In the quality test, participants listen to a original content speech and a conversion speech, and assign a score of 1 to 5 to evaluate the synthesized speech quality, where 5 indicates excellent, 4 indicates good, 3 indicates fair, 2 indicates poor and 1 indicates bad. In the similarity test, participants first listen to two original speech, one from timbre speaker and the other from content speaker. Then these participants listen to the converted speech and are asked to assign a score of 1 to 5.Following the design of [12], the similarity score of 5 corresponds to the same speaker with high confidence, and 1 corresponds to content speaker with high confidence.

As shown in Table III, the proposed framework outperforms the AUTOVC in similarity and quality. Especially in quality

MOS	Similarity				
	M2M	M2F	F2M	F2F	Avg
OURS	**2.96**	**3.72**	**4.01**	**2.90**	**3.39**
AUTOVC	2.81	3.64	3.93	2.75	3.27
MOS	Quality				
	M2M	M2F	F2M	F2F	Avg
OURS	**4.31**	**3.83**	**4.06**	**4.28**	**4.12**
AUTOVC	3.39	3.32	3.35	3.75	3.46

TABLE III
MOS BETWEEN AUTOVC AND OUR FRAMEWORK. QUALITY AND SIMILARITY TEST ARE PERFORMED FOR 4 GENDER CONVERSIONS: M2M, M2F, F2M, F2F, WHERE F INDICATES FEMALE SPEAKER AND M INDICATES MALE.

test, ours is significantly higher than AUTOVC. For similarity, although slightly improvement happened, ours avoids the need of careful bottleneck design. These results indicate that, GAN and conversion-reconstruction cycle training leads to a comparable improvement of similarity and conversion quality.

IV. CONCLUSION

In this paper, we propose an GAN-based framework for zero-shot many-to-many voice conversion. We use conversion-reconstruction cycle training to disentangle the timbre and content of speech. In order to enhance the conversion quality with GAN, we utilize two addition feature matching loss to relieve the difficulty in the training of GAN and to also guarantee the convergence of training. The objective experimental results show the effectiveness of conversion-reconstruction cycle training and the introduction of GAN. The MOS experimental results show that the proposed framework obtains higher sound quality and speaker similarity than the baseline method.

ACKNOWLEDGMENT

The work is supported by Key-Area Research and Development Program of Guangdong 2019B010154003, the National Natural Science Foundation of China U1801262, and the Science and Technology Project of Guangzhou 202103010002, Fundamental Research Funds for the Central Universities(2019PY21), Science and Technology Project of Zhongshan (2019AG024).

REFERENCES

[1] Kain, A. and Macon, M. W, "Spectral voice conversion for text-to-speech synthesis," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing,* ICASSP '98 1998.

[2] Nakamura, Keigo and Toda, Tomoki and Saruwatari, Hiroshi and Shikano, Kiyohiro. "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication,* vol. 54, pp. 134-146, 2012.

[3] Kaneko, Takuhiro and Kameoka, Hirokazu and Hiramatsu, Kaoru and Kashino, Kunio, "Sequence-to-Sequence Voice Conversion with Similarity Metric Learned Using Generative Adversarial Networks," *Interspeech,* 2017.

[4] Kaneko, Takuhiro and Kameoka, Hirokazu, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," *2018 26th European Signal Processing Conference (EUSIPCO),* 2018.

[5] Kaneko, Takuhiro and Kameoka, Hirokazu and Tanaka, Kou and Hojo, Nobukatsu, "CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* 2019.

[6] Kameoka, Hirokazu, Kaneko, Takuhiro, Tanaka, Kou, and Hojo, Nobukatsu, "StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks," *Spoken Language Technology Workshop,* 2018.

[7] Kaneko, Takuhiro, Kameoka, Hirokazu, Tanaka, Kou and Hojo, Nobukatsu, "StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion," *Interspeech,* 2019.

[8] Hsu, Chin Cheng, Hwang, Hsin Te, Wu, Yi Chiao, Tsao, Yu and Wang, Hsin Min, "Voice conversion from non-parallel corpora using variational auto-encoder," *Signal S Information Processing Association Summit S Conference,* 2016.

[9] Saito, Yuki, Ijima, Yusuke, Nishida, Kyosuke and Takamichi, Shinnosuke, "Non-Parallel Voice Conversion Using Variational Autoencoders Conditioned by Phonetic Posteriorgrams and D-Vectors," *ICASSP 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2018.

[10] Tobing, Patrick Lumban, Wu, Yi Chiao, Hayashi, Tomoki, Kobayashi, Kazuhiro and Toda, Tomoki, "Non-Parallel Voice Conversion with Cyclic Variational Autoencoder," *Interspeech,* 2019.

[11] Hsu, Chin-Cheng, Hwang, Hsin-Te, Wu, Yi-Chiao, Tsao, Yu and Wang, Hsin-Min, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849,* 2017.

[12] Qian, Kaizhi, Zhang, Yang, Chang, Shiyu, Yang, Xuesong, Hasegawa-Johnson, Mark, "AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss," *Proceedings of the 36th International Conference on Machine Learning, ICML ,* June, 2019.

[13] Qian, Kaizhi, Jin, Zeyu, Hasegawa-Johnson, Mark and Mysore, Gautham J, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2020.

[14] Bao, Jianmin, Chen, Dong, Wen, Fang, Li, Houqiang and Hua, Gang, "CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training," *IEEE International Conference on Computer Vision,* 2017.

[15] Wan, Li, Wang, Quan, Papir, Alan and Moreno, Ignacio Lopez, "Generalized End-to-End Loss for Speaker Verification," *ICASSP 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2018.

[16] Bao, Jianmin and Chen, Dong and Wen, Fang and Li, Houqiang and G.Hua, "Towards Open-Set Identity Preserving Face Synthesis," *In Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 6713-6722, June, 2018.

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

[18] Panayotov, Vassil, Chen, Guoguo, Povey, Daniel and Khudanpur, Sanjeev, "Librispeech: An ASR corpus based on public domain audio books," *ICASSP 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2015.

[19] Oord, Aaron Van Den, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, et.al, "WaveNet: A Generative Model for Raw Audio," *arXiv preprint arXiv:1609.03499,* 2016.

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

[21] Arjovsky, Martin and Bottou, Lon, "Towards Principled Methods for Training Generative Adversarial Networks," *Stat,* vol. 1050, 2017.