StarGANv2-VC: A Diverse, Unsupervised, Non-parallel Framework for Natural-Sounding Voice Conversion

Yinghao Aaron Li, Ali Zare, Nima Mesgarani

Department of Electrical Engineering, Columbia University, USA

yl4579@columbia.edu, az2584@columbia.edu, nima@ee.columbia.edu

Abstract

We present an unsupervised non-parallel many-to-many voice conversion (VC) method using a generative adversarial network (GAN) called StarGAN v2. Using a combination of adversarial source classifier loss and perceptual loss, our model significantly outperforms previous VC models. Although our model is trained only with 20 English speakers, it generalizes to a variety of voice conversion tasks, such as any-to-many, cross-lingual, and singing conversion. Using a style encoder, our framework can also convert plain reading speech into stylistic speech, such as emotional and falsetto speech. Subjective and objective evaluation experiments on a non-parallel many-to-many voice conversion task revealed that our model produces natural sounding voices, close to the sound quality of state-of-the-art text-to-speech (TTS) based voice conversion methods without the need for text labels. Moreover, our model is completely convolutional and with a faster-than-real-time vocoder such as Parallel WaveGAN can perform real-time voice conversion.

1. Introduction

Voice conversion (VC) is a technique for converting one speaker's voice identity into another while preserving linguistic content. This technique has various applications, such as movie dubbing, language learning by cross-language conversion, speaking assistance, and singing conversion. However, most voice conversion methods require parallel utterances to achieve high-quality natural conversion results, which strongly limits the conditions where this technique can be applied.

Recent work in non-parallel voice conversion using deep neural network models can mainly be divided into three categories: auto-encoder-based approach, TTS-based approach, and GAN-based approach. Auto-encoder approach, such as in [1, 2, 3, 4], seeks to encode speaker-independent information from input audio by training models with proper constraints. This approach requires carefully designed constraints to remove speaker-dependent information, and the converted speech quality depends on how much linguistic information can be retrieved from the latent space. On the other hand, GAN-based approaches, such as CycleGAN-VC3[5] and StarGAN-VC2[6] do not constrain the encoder, instead, they use a discriminator that teaches the decoder to generate speech that sounds like the target speaker. Since there is no guarantee that the discriminator will learn meaningful features from the real data, this approach often suffers from problems such as dissimilarity between converted and target speech, or distortions in voices of the generated speech. Unlike the previous two methods, TTS-based approaches like Cotatron [7], AttS2S-VC [8] and VTN [9] take advantage of text labels and synthesizes speech directly by extracting aligned linguistic features from the input speech. This ensures that the converted speaker identity is the same as the target speaker identity. However, this approach requires text labels, which are not often available at hand.

Here, we present a new method for unsupervised non-parallel many-to-many voice conversion using recently proposed GAN architecture for image style transfer, StarGAN v2 [10]. Our framework produces natural-sounding speech and significantly outperforms the previous state-of-art method, AUTO-VC [1], in terms of both naturalness and speaker similarity, approaching the TTS based approaches such as VTN [9] as reported in Voice Conversion Challenge 2020 (VCC2020) [11] . Besides, our model can generalize to a variety of voice conversion tasks, including any-to-many conversion, cross-language conversion, and singing conversion, even though it was trained only on monolingual speech data with limited numbers of speakers. Furthermore, when trained on a corpus with diverse speech styles, our model shows the ability to convert into stylistic speech, such as converting a plain reading voice into an emotive acting voice and convert a chest voice into a falsetto voice.

We have multiple contributions in this work: (i) applying StarGAN v2 to voice conversion, which enables converting from plain speech into speech with a diversity of styles, (ii) introduce a novel adversarial source classifier loss that greatly improves the similarity in terms of speaker identity between the converted speech and target speech, and (iii) the first voice conversion framework, as far as we know, that employs perceptual losses using both automatic speech recognition (ASR) network and fundamental frequency (F0) extraction network.

2. Method

2.1. StarGANv2-VC

StarGAN v2 [10] uses a single discriminator and generator to generate diverse images in each domain with the domain-specific style vectors from either the style encoder or the mapping network. We have adopted the same architecture to voice conversion, treated each speaker as an individual domain, and added a pre-trained joint detection and classification (JDC) F0 extraction network [13] to achieve F0-consistent conversion. An overview of our framework is shown in Figure 1.

Generator. The generator G converts an input mel-spectrogram X_{src} into $G(X_{src}, h_{sty}, h_{f0})$ that reflects the style in h_{sty}, which is given either by the mapping network or the style encoder, and the fundamental frequency in h_{f0}, which is provided by the convolution layers in the F0 extraction network F.

F0 network. The F0 extraction network F is a pre-trained JDC network [13] that extracts the fundamental frequency from an input mel-spectrogram. The JDC network has convolutional layers followed by BLSTM units. We only use the convolutional output $F_{conv}(X)$ for $X \in \mathcal{X}$ as the input features.

Mapping network. The mapping network M generates a style vector $h_M = M(z, y)$ with a random latent code $z \in \mathcal{Z}$ in a domain $y \in \mathcal{Y}$. The latent code is sampled from a Gaussian distribution to provide diverse style representations in all domains.
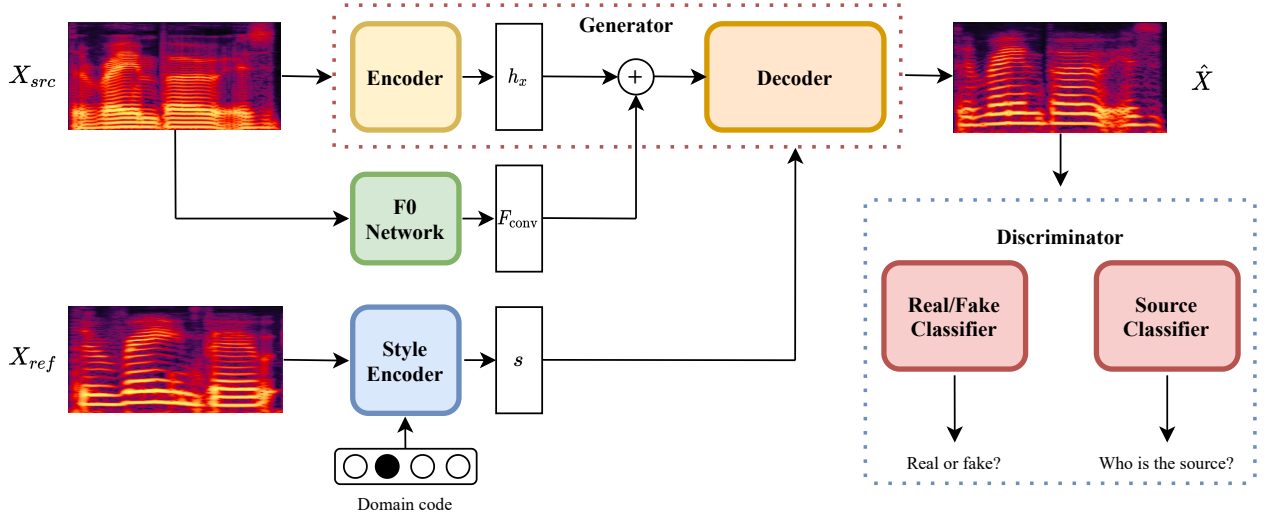
Figure 1: *StarGANv2-VC framework with style encoder.* $\boldsymbol{X_{src}}$ *is the source input,* $\boldsymbol{X_{ref}}$ *is the reference input that contains the style information, and* $\hat{\boldsymbol{X}}$ *represents the converted mel-spectrogram.* h_x, F_{conv} *and* s *denote the latent feature of the source, the F0 feature from convolutional layers of the source, and the style code of the reference in the target domain, respectively.* h_x *and* h_F0 *are concatenated by channel as the input to the decoder, and* h_{sty} *is injected into the decoder by the adaptive instance normalization (AdaIN) [12]. Two classifiers form the discriminators that determine whether a generated sample is real or fake and who the source speaker of* $\hat{\boldsymbol{X}}$ *is. In another scheme where the style encoder is replaced with the mapping network, the reference mel-spectrogram* $\boldsymbol{X_{ref}}$ *is not needed.*

The style vector representation is shared for all domains until the last layer, where a domain-specific projection is applied to the shared representation.

Style encoder. Given a reference mel-spectrogram $\boldsymbol{X_{ref}}$, the style encoder S extracts the style code $h_{sty} = S(\boldsymbol{X_{ref}}, y)$ in the domain $y \in \mathcal{Y}$. Similar to the mapping network M, S first processes an input through shared layers across all domains. A domain-specific projection then maps the shared features into a domain-specific style code.

Discriminators. The discriminator D in [10] has shared layers that learns the common features between real and fake samples in all domains, followed by a domain-specific binary classifier that classifies whether a sample is real in each domain $y \in \mathcal{Y}$. However, since the domain-specific classifier consists of only one convolutional layer, it may fail to capture important aspects of domain-specific features such as the pronunciations of a speaker. To address this problem, we introduce an additional classifier C with the same architecture as D that learns the original domain of converted samples. By learning what features elude the input domain even after conversion, the classifier can provide feedback about features invariant to the generator yet characteristic to the original domain, upon which the generator should improve to generate a more similar sample in the target domain. A more detailed illustration is given in Figure 2.

2.2. Training Objectives

The aim of StarGANv2-VC is to learn a mapping $G : \mathcal{X}_{y_{src}} \to \mathcal{X}_{y_{trg}}$ that converts a sample $\boldsymbol{X} \in \mathcal{X}_{y_{src}}$ from the source domain $y_{src} \in \mathcal{Y}$ to a sample $\hat{\boldsymbol{X}} \in \mathcal{X}_{y_{trg}}$ in the target domain $y_{trg} \in \mathcal{Y}$ without parallel data.

During training, we sample a target domain $y_{trg} \in \mathcal{Y}$ and a style code $s \in \mathcal{S}_{y_{trg}}$ randomly via either mapping network where $s = M(z, y_{trg})$ with a latent code $z \in \mathcal{Z}$, or style encoder where $s = S(\boldsymbol{X_{ref}}, y_{trg})$ with a reference input

$\boldsymbol{X_{ref}} \in \mathcal{X}$. Given a mel-spectrogram $\boldsymbol{X} \in \mathcal{X}_{y_{src}}$, the source domain $y_{src} \in \mathcal{Y}$ and the target domain $y_{trg} \in \mathcal{Y}$, we train our model with the following loss functions.

Adversarial loss. The generator takes an input mel-spectrogram $\boldsymbol{X}$ and a style vector s and learns to generate a new mel-spectrogram $G(\boldsymbol{X}, s)$ via the adversarial loss

$$\mathcal{L}_{adv} = \mathbb{E}_{\boldsymbol{X}, y_{src}} \left[\log D(\boldsymbol{X}, y_{src}) \right] + \\ \mathbb{E}_{\boldsymbol{X}, y_{trg}, s} \left[\log \left(1 - D(G(\boldsymbol{X}, s), y_{trg}) \right) \right] \quad (1)$$

where $D(\cdot, y)$ denotes the output of real/fake classifier for the domain $y \in \mathcal{Y}$.

Adversarial source classifier loss. We use an additional adversarial loss function with the source classifier C (see Figure 2)

$$\mathcal{L}_{advcls} = \mathbb{E}_{\boldsymbol{X}, y_{trg}, s} \left[\text{CE}(C(G(\boldsymbol{X}, s)), y_{trg}) \right] \quad (2)$$

where $\text{CE}(\cdot)$ denotes the cross-entropy loss function.

Style reconstruction loss. We use the style reconstruction loss to ensure that the style code can be reconstructed from the generated samples.

$$\mathcal{L}_{sty} = \mathbb{E}_{\boldsymbol{X}, y_{trg}, s} \left[\| s - S(G(\boldsymbol{X}, s), y_{trg}) \|_1 \right] \quad (3)$$

Style diversification loss. The style diversification loss is maximized to enforce the generator to generate different samples with different style codes. In addition to maximizing the mean absolute error (MAE) between generated samples, we also maximize MAE of the F0 features between samples generated with different style codes

$$\mathcal{L}_{ds} = \mathbb{E}_{\boldsymbol{X}, s_1, s_2, y_{trg}} \left[\| G(\boldsymbol{X}, s_1) - G(\boldsymbol{X}, s_2) \|_1 \right] + \\ \mathbb{E}_{\boldsymbol{X}, s_1, s_2, y_{trg}} \left[\| F_{conv}(G(\boldsymbol{X}, s_1)) - F_{conv}(G(\boldsymbol{X}, s_2)) \|_1 \right] \quad (4)$$

where $s_1, s_2 \in \mathcal{S}_{y_{trg}}$ are two randomly sampled style codes from domain $y_{trg} \in \mathcal{Y}$ and $F_{conv}(\cdot)$ is the output of convolutional layers of F0 network F.
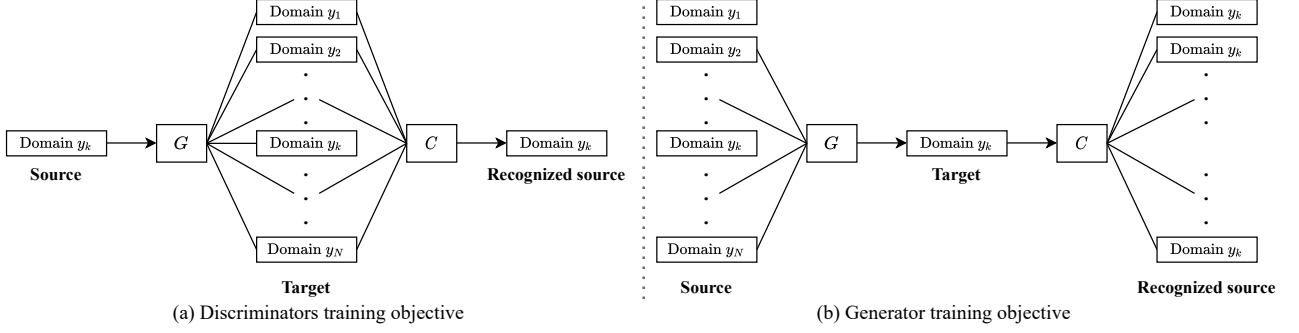
(a) Discriminators training objective

(b) Generator training objective

Figure 2: *Training schemes of adversarial source classifier for a domain y_k. The case $y_{trg} = y_{src}$ is omitted to prevent amplification of artifacts from the source classifier.* **(a)** *When training the discriminators, the weights of the generator G are fixed, and the source classifier C is trained to determine the original domain y_k of the converted samples, regardless of the target domains.* **(b)** *When training the generator, the weights of the source classifier C are fixed, and the generator G is trained to make C classify all generated samples as being converted from the target domain y_k, regardless of the actual domains of the source.*

F0 consistency loss. To produce F0-consistent results, we add an F0-consistent loss with the normalized F0 curve provided by F0 network F. For an input mel-spectrogram X, $F(X)$ provides the absolute F0 value in Hertz for each frame of X. Since male and female speakers have different average F0, we normalize the absolute F0 values $F(X)$ by its temporal mean, denoted by $\hat{F}(X) = \frac{F(X)}{\|F(X)\|_1}$. The F0 consistency loss is thus

$$\mathcal{L}_{f0} = \mathbb{E}_{X,s}\left[\left\|\hat{F}(X) - \hat{F}(G(X,s))\right\|_1\right] \quad (5)$$

Speech consistency loss. To ensure that the converted speech has the same linguistic content as the source, we employ a speech consistency loss using convolutional features from a pre-trained joint CTC-attention VGG-BLSTM network [14] given in Espnet toolkit [1][15]. Similar to [16], we use the output from the intermediate layer before the LSTM layers as the linguistic feature, denoted by $h_{asr}(\cdot)$. The speech consistency loss is defined as

$$\mathcal{L}_{asr} = \mathbb{E}_{X,s}\left[\|h_{asr}(X) - h_{asr}(G(X,s))\|_1\right] \quad (6)$$

Norm consistency loss. We use the norm consistency loss to preserve the speech/silence intervals of generated samples. We use the absolute column-sum norm for a mel-spectrogram X with N mels and T frames at the tth frame, defined as $\|X_{\cdot,t}\| = \sum_{n=1}^{N} |X_{n,t}|$, where $t \in \{1, \ldots, T\}$ is the frame index. The norm consistency loss is given by

$$\mathcal{L}_{norm} = \mathbb{E}_{X,s}\left[\frac{1}{T}\sum_{t=1}^{T}\|X_{\cdot,t}\| - \|G(X,s))_{\cdot,t}\|\right] \quad (7)$$

Cycle consistency loss. Lastly, we employ the cycle consistency loss [17] to preserve all other features of the input

$$\mathcal{L}_{cyc} = \mathbb{E}_{X,y_{src},y_{trg},s}\left[\|X - G(G(X,s),\tilde{s}))\|_1\right] \quad (8)$$

where $\tilde{s} = S(X, y_{src})$ is the estimated style code of the input in the source domain $y_{src} \in \mathcal{Y}$.

Full objective. Our full generator objective functions can be summarized as follows:

$$\min_{G,S,M} \mathcal{L}_{adv} + \lambda_{advcls}\mathcal{L}_{advcls} + \lambda_{sty}\mathcal{L}_{sty}$$
$$- \lambda_{ds}\mathcal{L}_{ds} + \lambda_{f0}\mathcal{L}_{f0} + \lambda_{asr}\mathcal{L}_{asr} \quad (9)$$
$$+ \lambda_{norm}\mathcal{L}_{norm} + \lambda_{cyc}\mathcal{L}_{cyc}$$

---

[1]https://github.com/espnet/espnet

where $\lambda_{advcls}, \lambda_{sty}, \lambda_{ds}, \lambda_{f0}, \lambda_{asr}, \lambda_{norm}$ and λ_{cyc} are hyper-parameters for each term.

Our full discriminators objective is given by:

$$\min_{C,D} -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls} \quad (10)$$

where λ_{cls} is the hyperparameter for source classifier loss $\mathcal{L}_{cls}$, which is give by

$$\mathcal{L}_{cls} = \mathbb{E}_{X,y_{src},s}\left[\mathbf{CE}(C(G(X,s)), y_{src})\right] \quad (11)$$

3. Experiments

3.1. Datasets

For a fair comparison, we train both the baseline model and our framework on the same 20 selected speakers reported in [18] from VCTK dataset [19]. To demonstrate the ability to convert to stylistic speech, we train our framework on 10 randomly selected speakers from the JVS dataset [20] with both regular and falsetto utterances. We also train our model on 10 English speakers from the emotional speech dataset (ESD) [21] with all five different emotions. We train our ASR and F0 model on the *train-clean-100* subset from the LibriSpeech dataset [22]. All datasets are resampled to 24 kHz and randomly split according to an $80\%/10\%/10\%$ of train/val/test partitions. For the baseline model, the datasets are downsampled to 16 kHz.

3.2. Training details

We train our model for 150 epochs, with a batch size of 10 two-second long audio segments. We use AdamW [23] optimizer with a learning rate of 0.0001 fixed throughout the training process. The source classifier joins the training process after 50 epochs. We set $\lambda_{cls} = 0.1, \lambda_{advcls} = 0.5, \lambda_{sty} = 1, \lambda_{ds} = 1, \lambda_{f0} = 5, \lambda_{asr} = 1, \lambda_{norm} = 1$ and $\lambda_{cyc} = 1$. The F0 model is trained with pitch contours given by World vocoder [24] for 100 epochs, and the ASR model is trained at phoneme level for 80 epochs with a character error rate (CER) of 8.53%. AUTO-VC is trained with one-hot embedding for 1M steps.

3.3. Evaluations

We evaluate our model with both subjective and objective metrics. The ablation study is conducted with only objective evaluations because the subjective evaluations are expensive and

Table 1: *Mean and standard error with subjective metrics.*

Method	Type	MOS	SIM
Ground truth	M	4.64 ± 0.14	—
	F	4.53 ± 0.15	
StarGANv2-VC	M2M	4.02 ± 0.22	3.86 ± 0.24
	F2M	4.06 ± 0.28	
	F2F	4.25 ± 0.26	
	M2F	4.03 ± 0.26	
AUTO-VC	M2M	2.70 ± 0.26	3.57 ± 0.27
	F2M	2.34 ± 0.20	
	F2F	2.86 ± 0.23	
	M2F	2.17 ± 0.23	

Table 2: *Results with objective metrics. Full StarGANv2-VC represents that all loss functions are included. CLS for ground truth is evaluated on the test set of AlexNet. All other metrics are evaluated on 1,000 samples with random source and target pairs. Mean and standard deviation of pMOS are reported.*

Method	pMOS	CLS	CER
Ground truth	4.02 ± 0.46	98.67%	11.47%
Full StarGANv2-VC	3.95 ± 0.50	96.20%	12.55%
$\lambda_{f0} = 0$	3.99 ± 0.54	96.50%	13.02%
$\lambda_{asr} = 0$	3.95 ± 0.53	96.90%	30.34%
$\lambda_{norm} = 0$	3.83 ± 0.47	96.30%	15.58%
$\lambda_{advcls} = 0$	3.98 ± 0.45	63.90%	12.33%
AUTO-VC	3.43 ± 0.51	50.30%	47.43%

time-consuming. We use a pre-trained Parallel WaveGAN[2] [25] to synthesize waveforms from mel-spectrogram. We downsampled our waveforms to 16 kHz to match the sample rate of [1].
Subjective metric. We randomly selected 5 male and 5 female speakers as our target speakers. The source speakers are randomly chosen from all 20 speakers to form 40 conversion pairs. Both source and ground truth samples were chosen to be at least 5-second long to ensure that there is enough information to judge the naturalness and similarity. We asked 46 online subjects to rate the naturalness of each audio clip on Amazon Mechanical Turk [3] on a scale of 1 to 5, where 1 indicates completely distorted and unnatural and 5 indicates no distortion and completely natural. Furthermore, we asked the subjects to rate from 1 to 5 whether the speakers of each pair of the audio clips could have been the same person, disregarding the distortion, speed, and tone of speech, where 1 indicates completely different speakers and 5 indicates exactly the same speaker. The subjects were not told whether an audio clip is the ground truth or converted. To make sure that the subjects did not complete the survey with random selections, we included 6 completely distorted and unintelligible audios as attention checks. The raters are excluded from the analysis if more than two of these samples were rated above 2. After excluding bad raters, we ended up with 43 subjects for our analysis. All raters are self-identified as native English speakers and are located in the United States.
Objective metric. We use predicted mean opinion score (pMOS) from MOSNet [26] to conduct the ablation study. Das et. al. [27] suggest automatic speaker verification (ASV) can be used as an objective metric for speaker similarity. For simplicity, similar to [16], we only train an AlexNet [28] for speaker recognition on the 20 selected speakers and report the classification accuracy (CLS). We also report the character error rate (CER) using the aforementioned ASR model for intelligibility.

4. Results

The results of the survey show that our method significantly outperforms the AUTO-VC model in terms of rated naturalness in all four conversion types (MOS column in Table 1, $p < 0.001$ for randomization test, $p < 0.001$ for t-test). The survey results also indicate that converted samples using our framework are significantly more similar to the ground truth in terms of speakers' identity than the baseline model (SIM column in Table 1). The accuracy of the speaker recognition model on samples con-

verted using our framework is much higher than those converted using AUTO-VC , and the predicted MOS of samples converted using our model is also significantly higher than AUTO-VC (see Table 2). Lastly, CER on audio clips converted using our model is significantly lower than those converted using AUTO-VC.

The ablation study shows that λ_{asr} is crucial for converting linguistic content, and removing λ_{advcls} significantly decreases the speaker recognition accuracy. We notice that removing λ_{norm} introduces noises to the converted audios during the silent intervals of the source audios. We also note that removing λ_{f0} does not necessarily lower pMOS, CLS, or CER because all these three metrics are not sensitive to F0. However, we find that the converted samples have unnatural intonations by examining the converted samples without λ_{f0}.

Lastly, we show that our framework is generalizable to a variety of voice conversion tasks with audio demos. Demo audio samples can be found online[4].

5. Conclusion

We present an unsupervised framework with StarGAN v2 for voice conversion with novel adversarial and perceptual losses to achieve state-of-art performance in terms of both naturalness and similarity. In VCC2020, only PGG-VC based methods [29] could achieve MOS higher than 4.0 with autoregressive vocoders, and the model with highest MOS was based on TTS using a transformer architecture [9]. Our model achieves similar MOS but without the need of text labels during training and the use of autoregressive vocoders. Our framework can also learn to convert to stylistic speech such as emotional acting or falsetto speech from plain reading source speech. Furthermore, our model generalizes to various voice conversion tasks, such as any-to-many, cross-lingual and singing conversion, without the need for explicit training. Using Parallel WaveGAN vocoder, our model can convert an audio clip hundreds of times faster than real time on Tesla P100, which makes it suitable for real-time voice conversion applications. Future work includes improving the quality of any-to-many, cross-lingual, and singing conversion schemes with our framework. We would also like to develop a real-time voice conversion system with our models.

6. Acknowledgements

---

[2]Available at `https://github.com/kan-bayashi/ParallelWaveGAN`

[3]The survey can be found at `https://survey.alchemer.com/s3/6266556/SoundQuality2`

---

[4]Available at `https://starganv2-vc.github.io/`

7. References

[1] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[2] S. Ding and R. Gutierrez-Osuna, "Group latent embedding for vector quantized variational autoencoder in non-parallel voice conversion." in *INTERSPEECH*, 2019, pp. 724–728.

[3] W.-C. Huang, H. Luo, H.-T. Hwang, C.-C. Lo, Y.-H. Peng, Y. Tsao, and H.-M. Wang, "Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 4, pp. 468–479, 2020.

[4] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6284–6288.

[5] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion," *arXiv preprint arXiv:2010.11672*, 2020.

[6] ——, "Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion," *arXiv preprint arXiv:1907.12279*, 2019.

[7] S.-w. Park, D.-y. Kim, and M.-c. Joe, "Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data," *arXiv preprint arXiv:2005.03295*, 2020.

[8] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6805–6809.

[9] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," *arXiv preprint arXiv:1912.06813*, 2019.

[10] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8188–8197.

[11] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," *arXiv preprint arXiv:2008.12527*, 2020.

[12] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[13] S. Kum and J. Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," *Applied Sciences*, vol. 9, no. 7, p. 1324, 2019.

[14] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[15] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[16] A. Polyak, L. Wolf, Y. Adi, and Y. Taigman, "Unsupervised cross-domain singing voice conversion," *arXiv preprint arXiv:2008.02830*, 2020.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[18] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," *arXiv preprint arXiv:1804.02812*, 2018.

[19] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[20] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "Jvs corpus: free japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[21] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," *arXiv preprint arXiv:2010.14794*, 2020.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[24] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[25] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[26] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning based objective assessment for voice conversion," *arXiv preprint arXiv:1904.08352*, 2019.

[27] R. K. Das, T. Kinnunen, W.-C. Huang, Z. Ling, J. Yamagishi, Y. Zhao, X. Tian, and T. Toda, "Predictions of subjective ratings and spoofing assessments of voice conversion challenge 2020 submissions," *arXiv preprint arXiv:2009.03554*, 2020.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[29] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.