ONE-SHOT VOICE CONVERSION USING STAR-GAN

Ruobai Wang, *Yu Ding,* *Lincheng Li,* *Changjie Fan*

Netease FuXi AI Lab, Hangzhou, China
{wangruobai, dingyu01, lilincheng, fanchangjie}@corp.netease.com

ABSTRACT

Our efforts are made on one-shot voice conversion where the target speaker is unseen in training dataset or both source and target speakers are unseen in the training dataset. In our work, StarGAN is employed to carry out voice conversion between speakers. An embedding vector is used to represent speaker ID. This work relies on two datasets in English and one dataset in Chinese, involving 38 speakers. A user study is conducted to validate our framework in terms of reconstruction quality and conversion quality. The results show that our framework is able to perform one-shot voice conversion and also outperforms state-of-the-art methods when the speaker in the test is seen in the training dataset. The exploration experiment demonstrates that our framework can be updated with incremental training when the data from new speakers is available.

***Index Terms*—** voice conversion, generative adversarial networks, StarGAN, speech, embedding, neural network

1. INTRODUCTION

Voice conversion (VC) is a technique to convert a source speech to a target one by modifying the timbre in speech but keeping the spoken content (transcripts) and prosodic features. In the target speech, timbre is associated with the target speaker but prosody is consistent with the source speech. VC is able to generate an utterance colored with controllable prosodies in input speech. It has been well-documented by numerous researchers [1, 2, 3]. VC Challenge (VCC) has been held in the past several years [4, 5] and attracted a lot of attention from academic and industrial participants. Additionally, VC has shown its potentials in industry. For example, with the help of VC, speech can be generated with specific timbres for various human-like characters [6].

Similar to VC, text-to-speech synthesis (TTS) is also used to generate speech. Differing from VC, TTS outputs speech from input transcript, instead of a source speech. The prosody in TTS speech depends on the speech dataset that is used to build TTS. Once TTS is built, the prosody is unique for an input sentence and it cannot be manipulated in the step of synthesis. As known, prosody contribute to communication and intelligent comprehension [7]. To manipulate prosody in artificial speech, VC is an appropriate solution. Under VC, the timbre is controllable through the source speech.

Recently, VC systems have been constructed with deep learning (DL) networks such as auto-encoders (AE) [8] or generative adversarial nets (GAN) [2, 9]. Viewing the previous works, we classify DL-based approaches into two groups. The one named ASR-VC is based on an automatic speech recognition (ASR) module where phoneme alignment is explicitly recognized first and then used to carry out VC at the level of phoneme, as done in [3, 10]. The other group named non-ASR VC, does not require ASR to explicitly calculate the phoneme posteriorgram.

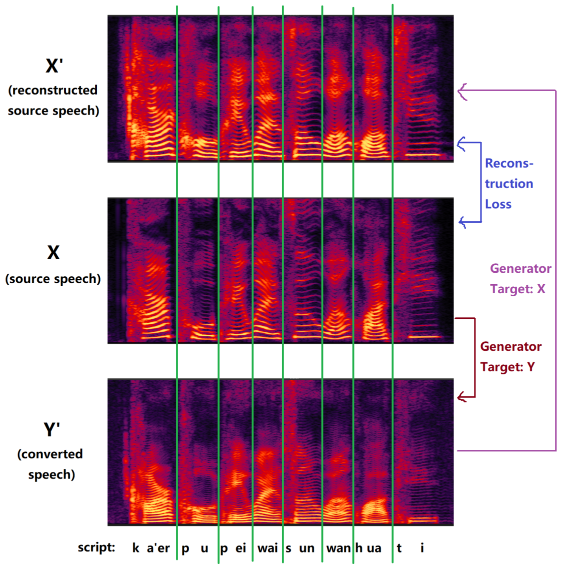


Fig. 1: Spectrograms before and after voice conversion. The prosody does not change, while F0 and spectral envelopes are converted to the target.

The group of ASR-VC methods mainly consist of two steps. The first step is to recognize the time-aligned phoneme sequence from the input speech through ASR; then the second step is to reconstruct speech by taking the time-aligned phoneme sequence as input to a specific TTS built on a target speaker/timbre. Therefore, the performance of ASR-VC methods highly depends on the accuracy of ASR. Training an ASR requires a huge training dataset with phoneme labeling and manual alignment. On the other hand, ASR may often be unavailable for many languages and dialects. To avoid the problem of requiring dataset at phoneme level, Non-ASR VC, which uses AE or GAN as network structure without explicit ASR or TTS modules, may be a potential solution.

The group of Non-ASR VC aim at reconstructing a target speech directly from a source speech without time-aligned phonemes as explicit features. The existing works employ a learning framework of DL-based GAN [11] consisting of a generator and a discriminator. The generator is learned from datasets to model real signals and generate plausible signals in reference. The discriminator is trained to classify signals from the generator as fake or real, while the generator is trained to fool the discriminator so that its outputs cannot be distinguished from real signals by the adversarially trained discriminator that is trained to do as well as possible at detecting the generator's "fakes" . GAN has been successfully used in image-to-image translation tasks [12, 13]. Classical GAN and its variants

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

have been proposed to perform VC, such as RNN-based GAN [1], CycleGAN [9] and StarGAN [2].

Chou et al. [1] take the discriminator in GAN to classify whether acoustic features from the generator are consistent with the source speaker or the target one. Under the supervision of the discriminator, the generator tends to output signals close to the target speaker. Differing from Chou et al. [1], Kaneko et al. [9] propose an existing variant of GAN, named CycleGAN [12], to perform voice conversion. Comparing to classical GAN, CycleGAN takes into account whether the output signals are translated back to the source input, which improves the quality of output signals. Classical GAN and cycleGAN perform one-to-one VC between two specific speakers. Later on, the authors propose StarGAN that allows for many-to-many VC [2]. The above works can only perform on specific speakers who have been seen in training dataset.

Our contribution is to perform VC between timbres not only included in training dataset but also from the speakers who are never seen in the step of training, which is viewed as one-shot VC. To avoid building dataset at phoneme level, we made efforts to improve state-of-the-art of Non-ASR VC. To achieve our aim, we develop an extension of StarGAN, called one-shot StarGAN. Our experiments show that one-shot StarGAN is able to achieve promising VC between any timbres and overcome the performance of StarGAN between timbres in training dataset.

2. METHODOLOGY

As mentioned before, one-shot StarGAN is developed to perform VC in our work. This section will introduce the algorithm pipeline and implementation details.

2.1. One-shot Voice Conversion

Taking one-hot speaker embedding, the original StarGAN fails when the source or target timbre is not in the training dataset. To deal with the timbres unseen in the training dataset, an embedding vector, named ID embedding, is taken into account and it represents the timbre of a speaker who may be or not seen in training dataset.

A straightforward idea is to apply a linear mapping on the acoustic features extracted from the input audio, and take the frame-wise mean as the embedding vector. However, since the uttered phonemes of each speech may be different, different speech segments of the same person can have different acoustic features and therefore be embedded to different vectors, which leads to unsatisfying performance.

We seek a neural network to generate ID embedding instead. In our work, Global Style Token (GST) speaker classification network [14] is taken to compute ID embedding according to the acoustic features in source or target speech.

In the previous works [2, 9], the acoustic features are normalized by speaker. Since samples of one-shot speech are often too few for the system to generalize meaningful statistical data such as mean and standard deviation of acoustic features, normalization by speaker is not able to deal with unseen speakers. Therefore, we normalize the acoustic features over all the speakers in the training dataset, and use the same mean and standard deviation to normalize the features of new speech in the step of inference.

GST is trained with a dataset consisting of normalized acoustic features from multiple speakers whose ID is represented/embedded by one-hot vector. Then a trained GST is capable of representing any arbitrary speaker ID, who is even unseen in the training dataset, with an embedding vector. Such an embedding vector refers to the "similarity" of timbre with those of each speaker in the training dataset. Thus, the resulting timbre would be a weighted combination of the timbres the speakers seen in the dataset. Embedding ID enables to project speaker timbre into a high-dimensional representation space.

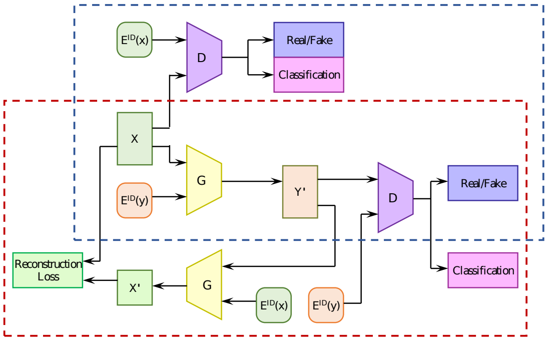


Fig. 2: Illustration of our proposed StarGAN framework. *G* is the generator and *D* is the discriminator. *X* denotes the acoustic features of source speech, $E^{ID}(\cdot)$ represents the embedding ID of a speaker, and *X'* and *Y'* refer to the features reconstructed through *G* with embedded ID $E^{ID}(X)$ and $E^{ID}(Y)$ respectively. The upper part framed with the blue block illustrates the training of *D* with the fixed *G*, and the lower part framed with the red block represents the process of training *G* with the fixed *D*.

Under the supervision of embedding ID from GST, StarGAN[1] is proposed to perform many-to-many VC, which relies on a discriminator, denoted as *D*, supervising the quality of output speech from a generator, denoted as *G*. The trained *D* is able to distinguish whether the output speech is real or fake and whether the output timbre is consistent with the target timbre. In the step of training, *G* is updated by adversarially updating *D*. In the step of inference, we first use GST to provide an embedding ID for the speaker, then we use *G* to generate the converted speech. Figure 1 shows an example of the spectrograms involved in this VC network.

Figure 2 illustrates the training process. The upper part framed with the blue block illustrates the training of *D* with the fixed generator, which expects to classify *X* as real and also corresponding to $E^{ID}(X)$, but *Y'* as fake. This process is carried out through minimizing two losses of distinguishing real vs. fake and embedding ID. The lower part framed with the red block represents the process of training *G* with the fixed discriminator, which is expected to fool the discriminator to classify *Y'* as real and corresponds to $E^{ID}(Y)$, and also to minimize the reconstruction loss between *X* and *X'*.

2.2. Improvements and implementation details

The bottleneck layer of the generator in StarGAN-VC is a 5-channel Conv2D layer [2]. In our experiments we found that it is too narrow and negatively affects the reconstruction quality. On the other hand, when we increase the number of channels, the encoder does not filter out all the source speaker's timbre information, which causes the converted speech to be affected by the source speaker's timbre, and the conversion is incomplete. This phenomenon is named information leakage in our work.

To avoid information leakage, we propose to use a Softmax layer as bottleneck instead, which would simulate the calculation of

[1] Readers can refer to [13] for more details about StarGAN.

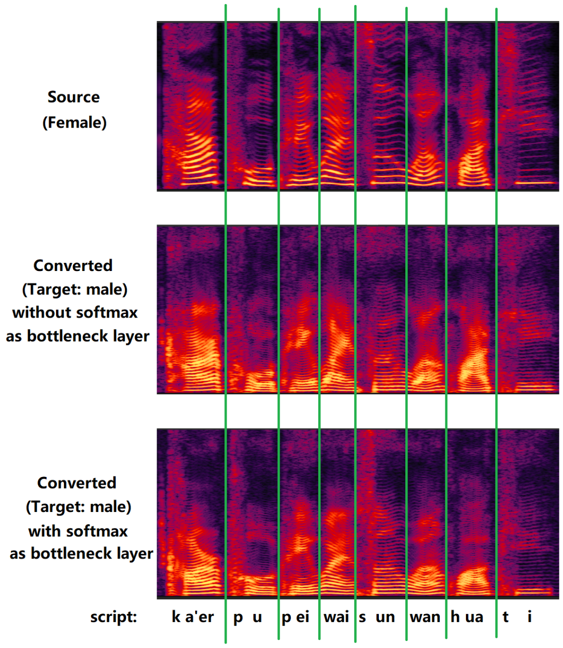


Fig. 3: The effect of softmax bottleneck layer on the spectrograms. Without the Softmax layer, the conversion does not have obvious effect on formants (peaks in the spectral envelope). With Softmax, the frequencies of formants become lower (especially for "wai" and "hua"), as expected in female-to-male conversion.

phoneme posteriorgram in ASR. The effect of softmax bottleneck can be observed in Figure 3.

Table 1 lists the details of the improved StarGAN. In even-numbered steps we update D and then G, while in odd-numbered steps we update D only. According to our experiments, 96-bin Mel spectral envelope as acoustic features outperform the 36-dimensional Mel cepstral coefficients (MFCC) which used in [2, 9]. The initial experiments show that MFCC incorporates more significant noise than Mel spectral envelope.

3. EXPERIMENTS

3.1. Datasets

In our experiments, three datasets are taken into account, including two in English and one in Chinese. The datasets in English are VCTK [15] and CMU-Arctic [16]. VCTK dataset involves 109 speakers and contains in average 20 minutes of random speech per speaker, containing untrimmed silence with occasional noise. CMU-Arctic dataset involves 18 speakers and contains about 1 hour of speech per speaker. Our dataset in Chinese is based on THCHS30 [17] involving 55 speakers and 30 minutes of speech in average from each speaker. Moreover, we record 3 female speakers, each of which utters 10 hours of high-quality speech, as a supplement to THCHS30.

3.2. Evaluation

Our proposed method is evaluated under three conditions (C), as follows.

Emb	Conv2d parameters				Add residual
	Transpose	Output channels	Kernel size	Stride size	
		32	(3,9)		
		64	(4,8)	2	✓
		128	(4,8)	2	✓
		256	(4,8)	2	✓
		256	(3,5)		✓
Reshape (B,C,H,W) to (B, C/8, 8H, W) Softmax along C, and reshape back					
✓		256	(3,5)		✓
✓	✓	128	4	2	✓
✓	✓	64	4	2	✓
✓	✓	32	4	2	✓
		1	7		

Table 1: Generator details of our proposed method. "Emb" concatenates the embedding vector to the channel dimension.

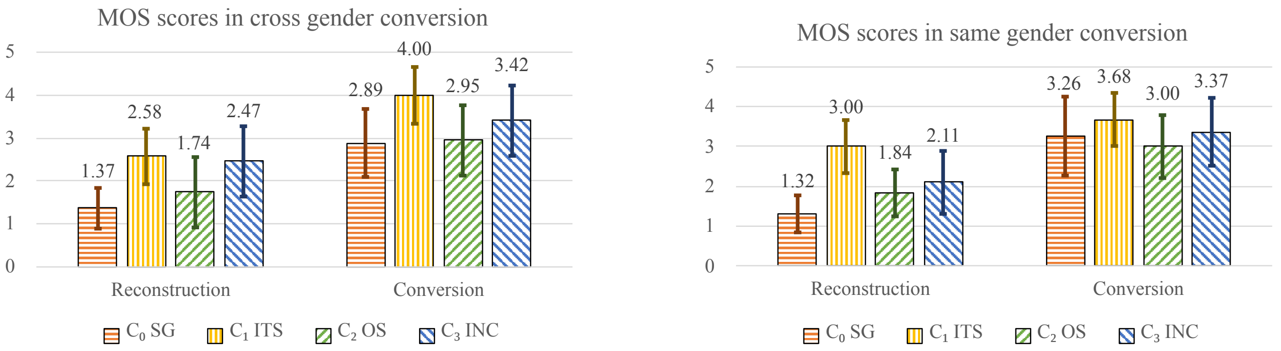
- C_1^{ITS}: the target speaker is seen in the training dataset. ITS refers to $in\ the\ training\ dataset$.
- C_2^{OS}: the target speaker is not seen during training. OS refers to $one\text{-}shot\ conversion$.
- C_3^{INC}: INC refers to $incremental\ training$. Incremental training [18] refers to that a trained framework is refined with data of a new speaker. In incremental training, we first train the proposed VC network with a slot in the one-hot embedding vector reserved for the new speaker, and store the trained model upon convergence. Then the data of a new speaker is added into the training dataset, and a biased training set is constructed by making the probability of data from the new speaker the same as from the existing speakers. The stored model is then refined with the biased data for a much shorter time than training from scratch. Incremental training can save training time while obtaining similar result as if the speaker has been in the dataset since the beginning.

In our earlier experiments, the StarGAN-VC framework proposed in [2] performs more robust than other existing methods including CycleGAN [9] and RNN-based VC [1]. Therefore, the StarGAN-VC framework [2] is taken as baseline to validate our framework. To facilitate reporting the comparison results, StarGAN-VC framework [2] is represented by C_0^{SG} where SG stands for "Star-GAN" .

Our training dataset contains CMU-Arctic[2] and a sub-dataset in Chinese consisting of 16 speakers picked up randomly from original THCHS30 and 1 hour of speech from each of the 3 supplemental speakers. In total, 37 speakers are involved in the training datasets, including all 18 English speakers from CMU-Arctic and all 19 Chinese speakers. The testing dataset involves an extra English speaker from VCTK dataset, and those from the training set. The speech samples in our test dataset are unseen in the training dataset. The speaker ID in the test dataset may be seen in the training dataset and may be unseen, depending on the conditions.

A user study is conducted to validate our framework. Four test samples are selected randomly from the test dataset. They are taken

[2] CMU-Arctic is taken as the training dataset in English as it is less noisy and easier to train on, and VCTK is used as testing dataset which provides unseen speakers.



(a) MOS scores in cross-gender conversion

(b) MOS scores in same-gender conversion

Fig. 4: Results of our experiments

	C_0^{SG}	C_1^{ITS}	C_2^{OS}	C_3^{INC}
C_0^{SG}	—	—	—	—
C_1^{ITS}	***	—	***	ns
C_2^{OS}	**	—	—	—
C_3^{INC}	***	—	***	—

(a) Cross-gender reconstruction quality.

	C_0^{SG}	C_1^{ITS}	C_2^{OS}	C_3^{INC}
C_0^{SG}	—	—	—	—
C_1^{ITS}	***	—	***	**
C_2^{OS}	ns	—	—	—
C_3^{INC}	*	—	**	—

(b) Cross-gender conversion quality.

	C_0^{SG}	C_1^{ITS}	C_2^{OS}	C_3^{INC}
C_0^{SG}	—	—	—	—
C_1^{ITS}	***	—	***	***
C_2^{OS}	***	—	—	—
C_3^{INC}	***	—	*	—

(c) Same-gender reconstruction quality.

	C_0^{SG}	C_1^{ITS}	C_2^{OS}	C_3^{INC}
C_0^{SG}	—	—	ns	—
C_1^{ITS}	*	—	**	ns
C_2^{OS}	—	—	—	—
C_3^{INC}	ns	—	ns	—

(d) Same-gender conversion quality.

Table 2: Statistical significance of the methods on the left being better than the ones on the top.
—: not applicable; ns: no statistically significant difference; *: $p<0.05$; **: $p<0.01$; ***: $p<0.001$

as input to the frameworks under the conditions of C_1^{ITS}, C_2^{OS} and C_3^{INC}, and the baseline of C_0^{SG}. In total, 16 samples of reconstructed speech are obtained and four ones for each condition. 19 participants are invited to evaluate the performance of four conditions. Each participant is asked to rate two terms of each speech, using mean opinion score (MOS) with 5-point Likert scale. One is reconstruction quality referring to the intelligibility, clarity and naturalness of reconstructed speech. In the 5-point Likert scale, *1* refers to the lowest quality, and *5* the highest. The other is conversion quality referring to how similar the timbre of the reconstructed speech is to the target speaker and to the source speaker. *1* means that the constructed timbre is perceived as being the source timbre, and *5* being the target timbre.

Intuitively, the timbre gap between cross-gender and same-gender is clear. To demonstrate more details, the experimental results are reported in two groups. One is based on cross-gender and the other is on same-gender. T-test is used to verify whether the preference between each pair of conditions is statistically significant. The T-test results are reported in Figure 4 and Table 2.

The **cross-gender results** rely on a female Chinese speaker as the source and a male English speaker as the target. As observed in Figure 4 and Table 2, C_1^{ITS}, C_2^{OS} and C_3^{INC} have statistically higher MOS scores than C_0^{SG} in reconstruction quality; C_1^{ITS} and C_3^{INC} also have better MOS scores than C_0^{SG} in conversion quality. This validates the performance of our method in the conditions of C_1^{ITS}, C_2^{OS} and C_3^{INC}. Additionally, C_1^{ITS} and C_3^{INC} outperform C_2^{OS} in both aspects. This result implies that the timbres used in the training affect the reconstructed timbre.

The **same-gender results** rely on a female Chinese source speaker and a female Chinese target speaker. As observed in Figure 4(b) and Table 2(c,d), C_1^{ITS} is valued higher than C_0^{SG} in reconstruction quality and conversion quality. This verifies that our method outperforms the StarGAN baseline when the speakers'data are available for training.

It is observed that both C_3^{INC} and C_2^{OS} are rated higher than C_0^{SG} in reconstruction quality, and there is no statistically significant difference between the conversion quality of C_3^{INC}, C_2^{OS} and C_0^{SG}. This result demonstrates that our framework is validated with incremental training, and also validates our one-shot VC method. Conversion samples are available at https://sniperwrb.github.io/icassp2020

4. CONCLUSION

This paper proposes to use embedding vector to represent speaker ID in the task of voice conversion. The proposed StarGAN framework has no requirement of parallel or labeled training data at all. It is able to perform voice conversion when the target and source speakers are unseen in the training dataset. The incremental training experiments show that our framework can be retrained to satisfy more new timbres. The current results rely on few test samples. In the future, we will conduct additional experiments with more test samples. Moreover, we will improve the proposed neural network to enhance the performance between two similar timbres.

5. REFERENCES

[1] Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, and Lin-shan Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," *Proc. Interspeech 2018*, pp. 501–505, 2018.

[2] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[3] Lifa Sun, Hao Wang, Shiyin Kang, Kun Li, and Helen Meng, "Personalized, cross-lingual tts using phonetic posteriorgrams," *Interspeech 2016*, pp. 322–326, 2016.

[4] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "The voice conversion challenge 2016," *Interspeech 2016*, pp. 1632–1636, 2016.

[5] 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 *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 195–202.

[6] Oytun Turk and Levent M. Arslan, "Subband based voice conversion," in *Seventh International Conference on Spoken Language Processing*, 2002.

[7] Anne Cutler, Delphine Dahan, and Wilma Van Donselaar, "Prosody in the comprehension of spoken language: A literature review," *Language and speech*, vol. 40, no. 2, pp. 141–201, 1997.

[8] 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," *Proc. Interspeech*, 2017.

[9] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[10] 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.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[12] 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.

[13] Yunjey Choi, 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[14] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*, 2018, pp. 5167–5176.

[15] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[16] John Kominek and Alan W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[17] Dong Wang and Xuewei Zhang, "Thchs-30: A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015.

[18] Robi Polikar, Lalita Upda, Satish S. Upda, and Vasant Honavar, "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)*, vol. 31, no. 4, pp. 497–508, 2001.