Voice Reenactment with F0 and timing constraints and adversarial learning of conversions

Frederik Bous, Laurent Benaroya, Nicolas Obin, Axel Roebel

HAL Id: hal-03677462

https://hal.archives-ouvertes.fr/hal-03677462

Submitted on 24 May 2022

Voice Reenactment with F0 and timing constraints and adversarial learning of conversions

Frederik Bous, Laurent Benaroya, Nicolas Obin, Axel Roebel
STMS Lab
IRCAM, CNRS, Sorbonne Université
Paris, France

Abstract—This paper introduces *voice reenactement* as the task of voice conversion (VC) in which the expressivity of the source speaker is preserved during conversion while the identity of a target speaker is transferred. To do so, an original neural-VC architecture is proposed based on sequence-to-sequence voice conversion (S2S-VC) in which the speech prosody of the source speaker is preserved during conversion. First, the S2S-VC architecture is modified so as to synchronize the converted speech with the source speech by mean of phonetic duration encoding; second, the decoder is conditioned on the desired sequence of F0-values and an explicit F0-loss is formulated between the F0 of the source speaker and the one of the converted speech. Besides, an adversarial learning of conversions is integrated within the S2S-VC architecture so as to exploit both advantages of reconstruction of original speech and converted speech with manipulated attributes during training and then reducing the inconsistency between training and conversion. An experimental evaluation on the VCTK speech database shows that the speech prosody can be efficiently preserved during conversion, and that the proposed adversarial learning consistently improves the conversion and the naturalness of the reenacted speech.

Index Terms—Voice conversion, voice reenactement, prosody preservation

I. INTRODUCTION

A. Context and Related works

Voice conversion (VC) consists of digitally altering the voice of an individual - e.g., its identity, accent, or emotion - while maintaining its linguistic content unchanged. Primarily applied to identity conversion, VC has considerably gained in popularity and in quality thanks to the advances accomplished with neural VC (see for instance, [1]–[3]). Neural VC is now widely considered as a standard in VC and has reached a highly-realistic rendering of voice identity conversion from a small amount of data of a target voice. Similarly to face manipulation, voice conversion has a wide range of potential applications, such as voice cloning and deep fake in the fields of entertainment and fraud [4], anonymization of voice identity in the field of security and data privacy [5], [6], or digital voice prosthesis of impaired speech in the field of digital healthcare [7].

Introduced in the late 2000's, neural VC has gradually moved towards many-to-many and non-parallel datasets allowing the scalability of neural VC to large and multiple speakers datasets with the assumption that the increase of data will induce a substantial increase in terms of quality and naturalness of the VC. In particular, starGAN-VC [8], [9] has been proposed to extend the paradigm of cycle-GAN to many-to-many and non-parallel VC by proposing a conditional encoder-decoder architecture. As opposed to the cycleGAN-VC, starGAN-VC is composed of a single encoder-decoder in which the decoder is conditioned on the speaker identity to be converted. In the starGAN-VC, a discriminator is employed to distinguish between real original speech and fake converted speech together with a speaker classifier is employed to determine whether the converted speech has been produced by the target speaker. As in the cycleGAN, a cycle-consistency loss attempts to preserve the linguistic content during conversion. To overcome the issue of linguistic preservation trough conversion, phonetic posterior-grams (PPGs) [10], [11] have been proposed to integrate explicitly linguistic content information by mean of time-aligned phonetic posterior probabilities that are used to condition the conversion.

More recently, neural VC architectures has reformulated the VC problem as a simple auto-encoder [12]–[14] (further referred to as AE-VC). The underlying idea is to provide a more structured representation of the information in the process of VC, in particular by explicitly encoding and disentangling linguistic content. VC is then achieved by manipulating the speaker identity by conditioning the decoder on the desired speaker identity. This has been achieved either by carefully tuning a bottleneck to encode only the linguistic content information as in [13] or through adversarial learning of linguistic content and speaker identity codes as in [14]. The advantage of those architectures is that the ground truth is known during training, so that a simple reconstruction loss can be efficiently applied (as opposed to the starGAN-VC). Moreover, linguistic and speaker information are explicitly represented and learned through disentanglement, which should idealistically provide a more accurate control during conversion. Those architectures have opened the possibility of processing VC from a very small number of examples of the target speaker (at the extreme from one-shot [12] or zero-shot [13]). One important assumption of these architectures relies on the fact that though a simple reconstruction is used during training (thus no conversion is actually learned), the manipulation of the speaker identity is still effective during conversion. However, this might seem somehow inconsistent since the actual conversion is not used during training (as opposed to the starGAN-VC).

B. Limitations and Contributions

Though VC systems can now achieve highly-realistic voice identity conversion [2], [3], the only information which is assumed to be preserved from the source speaker through conversion is the actual linguistic content of the spoken utterance (i.e., the text transcript). This constitutes an important limitation of current VC systems since one could desire to preserve some aspects of the source speaker through conversion, e.g. the prosody and the expressivity of the source speaker. By direct analogy with the image domain, face reenactment is defined by preserving the expression of the source face while converting the identity to the one of a target face [15]. This provides a more flexible and controllable transformation as it is commonly known as deep fakes, in which the source face provides the desired pose and expression to be rendered with the target identity.

The same definition can be applied in voice conversion, which will be further referred to as *voice reenactment*. Voice reenactment aims at preserving some aspects of the prosody of the source speaker during conversion, namely its timing - by means of phoneme durations - and its fundamental frequency (F0) both being commonly considered as the most important aspects of speech prosody. Though timing is ineherently preserved in some VC architectures such as cycleGAN-VC or starGAN-VC, this is not the case for all architectures including some of the aforementioned VC architectures based on auto-encoders. Besides, the only research to date to the knowledge of the authors on the F0-preservation of the source speaker in VC has been proposed by [16], which is solely obtained by conditioning the decoder of Auto-VC on the desired sequence of F0-values. Moreover, AE-based VC suffers from the fact that the conversion is actually not learned during training while the GAN-based VC suffers from the fact that no ground truth is available to learn the conversion (and instead applying the adversarial loss). However, research on face manipulation including the original starGAN [17] have presented strategies that can make both advantage of reconstruction of the original real data and adversarial learning of the manipulated data [18]. The main contributions of this paper can be listed as follows:

1) The task of *voice reenactment* is introduced with the objective of preserving the expressivity of the source speaker during conversion while transferring the identity of the target speaker;
2) An original solution to this task is presented by preserving the timing and the F0 of the source speaker during conversion within a S2S AE-VC architecture [14]. In particular, a F0-loss is explicitly formulated and exploited during training;
3) Inspired by the original formulation of the starGAN, an adversarial learning of identity conversion is integrated within the AE-VC architecture so as to both take advantage of reconstruction and conversion during training and re-ducing the inconsistency between training and conversion.

II. PROPOSED METHOD

A. Original S2S Neural VC

The VC framework used in this paper is rooted on a sequence-to-sequence (S2S) auto-encoder as proposed in [14], in which disentangled linguistic and speaker representation are encoded through dedicated encoders as illustrated in Figure 1. The inputs of the system are the speech signal matrix $\mathbf{A}$ represented by the Mel-spectrogram computed on T time frames, and the sequence of T phonemes $\mathbf{p}$ corresponding to the phonetic transcription of the input text aligned to the corresponding speech signal. Dual encoders E^c and E^s are employed to encode linguistic content and speaker information. The speaker encoder E^s converts the speech signal $\mathbf{A}$ into a time-independent vector $\mathbf{h}^s$, since it is assumed that the identity of a speaker does not vary within an utterance. A speaker classification loss L_{SE} is defined between the speaker identity predicted from $\mathbf{h}^s$ and the true speaker identity $\mathbf{s}$. The linguistic encoder E^c converts either the phoneme sequence $\mathbf{p}$ or the speech signal $\mathbf{A}$ into a shared linguistic embedding $\mathbf{H}^c$ through a contrastive loss. Contrary to [14], the linguistic embedding has the same length T as the aligned phoneme sequence (and the Mel-spectrogram), so that the time information is preserved during encoding. A S2S decoder G^a conditioned on the content embedding $\mathbf{H}^c$ and the speaker embedding $\mathbf{h}^s$ is employed to reconstruct an approximation $\widehat{\mathbf{A}}$ of the original speech signal $\mathbf{A}$. A reconstruction loss L_{RC} is defined between the reconstructed speech signal $\widehat{\mathbf{A}}$ and the original speech signal $\mathbf{A}$. During training, the S2S-VC neural network is pre-trained on a multiple-speakers dataset, and then fine-tuned with respect to a given pair of source and target speakers. During conversion, the recognition encoder E^c computes the content embedding $\mathbf{H}^c_{src}$ corresponding to one utterance $\mathbf{A}_{src}$ of the source speaker; and the speaker encoder E^s computes the speaker embedding $\mathbf{h}^s_{tgt}$ corresponding the one utterance $\mathbf{A}_{tgt}$ of the target speaker. Then, decoder G_a is conditioned on the linguistic embedding $\mathbf{H}^c_{src}$ and the speaker embedding $\mathbf{h}^s_{tgt}$ to generate the utterance $\mathbf{B}_{tgt}$ with the identity of the target speaker. Without explicit notification, the architecture and parameters of the S2S-VC are the same as those described in [14].

B. Pre-processing and Post-Net

The S2S-VC operates on a Mel-spectrogram representation of the speech signal. For the signal analysis we follow the parameterization proposed in [19], that is the input signal is down-sampled to 16kHz, converted into an STFT using an Hanning window of 50ms with hop size of 12.5ms and an FFT size of 2048. We then use 80 Mel bins covering the frequency band from 0 to 8Khz and convert the result into log amplitude domain. A standardization of the log Mel-spectrogram is applied at the input of the VC system, i.e. on each Mel bin, removing the mean and diving by the standard deviation, which are pre-computed on the entire training dataset. For rendering the speech signal from a generated Mel-spectrogram, a multi-speaker waveglow neural decoder following loosely [20] was employed. This decoder was trained over $900,000$ iterations
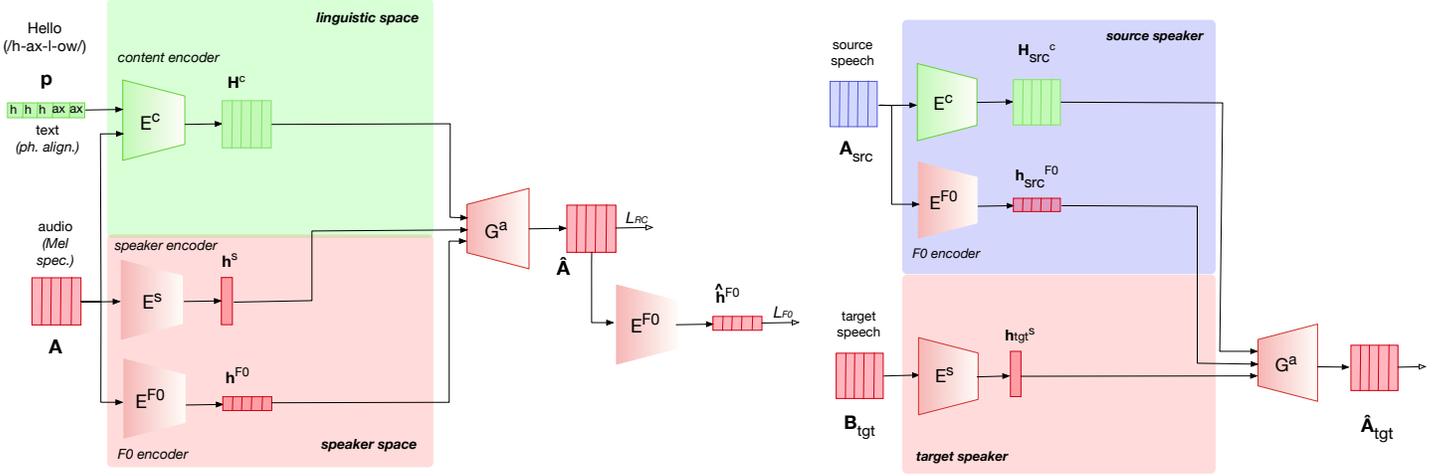
Fig. 1. Architecture of the proposed S2S-VC system. On left: The original VC comprises linguistic and speaker encoder E^c and E^s, and a decoder G^a which reconstructs the speech signal $\widehat{\mathbf{A}}$ conditioned on content embedding $\mathbf{H}^c$ and speaker embedding $\mathbf{h}^s$. A F0 encoder E^{F_0} is added to encode the sequence of F0 values from the original Mel-spectrogram, and a F0 loss is computed between the F0 of the reconstructed speech signal and the one of the original speech signal. On right: During conversion, content embedding $\mathbf{H}^c_{src}$ and F0 $\mathbf{h}^{F_0}_{src}$ is extracted from speech signal of the source speaker and the speaker embedding $\mathbf{h}^s_{tgt}$ of the target speaker.

using all samples of the VCTK database with batch size 50, and segment length 375ms using the Adam optimizer with learning rate $1e^{-4}$.

C. Proposed S2S Neural Voice Reenactemen

1) Contribution 1: S2S-VC with time-synchronization and F0-conditioning: S2S-VC architecture does not ensure time-synchronicity between the original speech and the converted speech, by definition of the sequence-to-sequence architecture. The S2S-VS is basically composed on a speech-to-text encoding time-compression and a text-to-speech decoding time-decompression with a similar architecture as the Tacotron ([21], [22]). In order to preserve time-synchronization between the original and the reconstructed speech signals, the time dimension of length T is preserved all through the network, from the original speech signal $\mathbf{A}$ to the linguistic embedding $\mathbf{H}^t$, and to the reconstructed speech signal $\widehat{\mathbf{A}}$. To do so, the auto-regressive S2S part of the recognition encoder E^t and the decoder G^a are modified accordingly by employing simple recurrent architectures. The recognition encoder E^t is composed of two bidirectional LSTM layers of dimension 128 followed by a fully connected layer (FC) of dimension 128, resulting in a linguistic embedding of dimension $(128 \times T)$. The decoder G^a is using two bidirectional LSTMs of dimension 128 each and a Fully Connected layer of dimension 80 which outputs an approximated Mel-spectrogram with the same dimensions as the input Mel spectrogram, i.e., $(80 \times T)$. These simplifications enable time-synchronous conversions and a consequent saving in computational time: approximately 33% of the computational time for training on our server with a single GPU. In order to preserve the F0 of the original speech signal during conversion, a F0 encoder E^{F_0} is added and a F0-loss is explicitly formulated (as opposed to [16]). The F0 encoder converts the input Mel-spectrogram into a corresponding sequence of F0 values. This value corresponds to the estimated $F0$ value for voiced frames and to 0 for unvoiced

frames. This encoding is passed to condition the decoder G_a in addition to the linguistic and speaker embeddings to create the Mel-spectrogram. A F0 loss is defined as the mean square error between the F0 values of the generated speech and the F0 values of the original speech, as

$$\mathcal{L}_{F_0}(\mathbf{h}^{F_0}, \mathbf{h}^{F_0}) = \frac{1}{T} \sum_{t=1}^{T} (\mathbf{h}^{F_0}(t) - \widehat{\mathbf{h}}^{F_0}(t))^2 \qquad (1)$$

The F0 encoder E^{F_0} is described in [23], pre-trained on the training set of the VCTK speech database, and then fixed during VC training, i.e. is only used to compute the F0 loss $\mathcal{L}_{F_0}$. Contrary to [16] this loss explicitly constrains the $F0$ to the desired $F0$ by defining a dedicated loss. This ensures that the $F0$ of the converted speech to be effectively preserved during conversion. This loss is added to the reconstruction loss with a weight λ_{F0} varying linearly from $10e^{-6}$ to $10e^{-2}$ with the effect of increasing gradually the importance of the F0 preservation versus the reconstruction loss during training. During conversion, the $F0$ can be transferred from an utterance of a source speaker or fixed arbitrarily (e.g., by applying transposition or setting any arbitrary values). In this paper, the $F0$ used for conditioning was adapted to the range of the target speaker in order to prevent unnatural converted speech that would be caused by some important difference between the respective ranges of the source and target speakers (typically when converting a male to a female or conversely). This was accomplished by normalizing the $F0$ values corresponding to the sentence of the source speaker with respect to the log(F0) mean and standard deviation of the target speaker.

2) Contribution 2: Adversarial Training of S2S-VC: S2S-VC basically relies on an auto-encoder optimizing a reconstruction loss between the original speech signal and the reconstructed one. This is somehow inconsistent with the conversion in which

TABLE I
ROOT MEAN SQUARE ERROR OF THE RECONSTRUCTED F0 (IN HZ.).

VC system	M-to-M	F-to-F	M-to-F	F-to-M
F0 cond w/ adv. same id	2.712	5.686	2.970	5.090
F0 cond w/ adv. diff id	2.574	6.246	4.171	5.252

the identity of the speaker is manipulated during conversion (eventually, the F0 and timing) . In the case of a conversion, one does not have access to the ground truth speech signal and thus one cannot apply the reconstruction loss of the auto-encoder. To overcome this limitation and construct a VC system whose training is more consistent with conversion, we propose to split the training process into two modes, as inspired by the original starGAN [17]: a *reconstruction* loss corresponding to the classical auto-encoder in which the reconstruction loss can be computed; an *adversarial* loss, in which we assume that the true speech signal may not be available. This is typically the case in which least one of the codes conditioning the decoder is manipulated. For this mode, we introduce an adversarial module which is similar to the one used in a GAN. A discriminator D^{adv} is optimized to distinguish between the real speech samples and the converted ones, while the decoder G^a is optimized to fool the discriminator. During training, each samples contained in a batch is both passed to the decoder in the reconstruction mode with original unchanged codes and in the conversion mode with unchanged or manipulated codes. In this paper, only the speaker identity is manipulated so that the reconstruction has the right identity and the conversion mode a randomly picked identity. The total loss $\mathcal{L}_{GEN}$ including reconstruction and conversion losses can then be expressed as,

$$\mathcal{L}_{GEN} = \mathcal{L}_{RC} + \lambda_{adv}\mathcal{L}_{ADV} \qquad (2)$$

For clarity, the F0 loss is not specified in the above loss but is still used as previously described. In this paper, the discriminator D^{adv} is composed of 4 convolution layers with 128 filters of size (3×3) with a stride of (2×2) and $\lambda_{adv} = 1$ was found to provide a good trade-off between the effect of the reconstruction and the conversion on the decoder G^a.

III. EXPERIMENTS

A. Speech dataset

The English multi-speaker corpus VCTK [24] is used for training and testing. The VCTK dataset contains speech data uttered by 110 speakers and the corresponding text transcripts. Each speaker read about 400 sentences selected from English newspaper, which represents a total of about 44 hours of speech. Four speakers All speakers are included into the training and validation sets. For each speaker, we split the database in a training set with 90% of the sentences and a validation set with 10% of them. The total duration of the database is around 27 hours after removing silences at the beginning and at the end of each sentence.

B. Subjective Experiment Setup

The experiment consisted into the judgment by listeners of: the similarity of the converted speech to the target speaker

and the naturalness of the converted speech, using 5-degree MOS scale as commonly used for the experimental evaluation of VC algorithms. Each participant had to judge 15 speech samples which were randomly selected among the total number of speech samples produced for the subjective experiments. The experiment was conducted on-line. Participants were encourage to perform the experiment in a quiet environment and with a headphone. Four speakers were used for the experiment: two males (p232 and p274) and two females (p253 and p300) with eight randomly chosen sentences per speaker in the validation set. Conversion were computed between male speakers and between females speakers, resulting in two male and two female conversion setups. Four configurations were compared: 1) the original audio signal and the converted speech with : 2) time-synchronous and F0-preserved VC system (referred to as F0 cond.); 3) time-synchronous and F0-preserved VC system with discriminator trained only with the true speaker identity (same identity as the source speech utterance) , referred to as F0 cond. w/adv same id; 4) time-synchronous and F0-preserved VC system with discriminator trained only with varying speaker identities (different identities from the one of the source speech utterance), referred to as F0 cond. w/adv diff id [1].

C. Results and Discussion

Table I presents the root mean squared error of the F0 of the converted speech with the proposed configurations for the 4 speakers used in the experiment. The error is around 5 Hz. in average which is not audible in most cases. This shows the efficiency of the proposed F0 preservation strategy, even in combination with the adversarial loss. Table II presents the mean MOS scores of the compared system configurations obtained with 25 participants. On the one hand, the baseline VC system using time-synchronization and F0 preservation presents fair to good similarity to the target speaker (MOS=3.92 for similarity), even though the timing and the F0 contours is inherited from the source speaker. Also, the naturalness is also rated to the fair range (MOS=3.09). One can observe that the naturalness is degraded with the female speakers (MOS=2.85) as compared to the male speakers (MOS=3.38). These results are consistent with the ones obtained in the literature about S2S-VC, as in [14]. This indicates that the application of a timing and a F0 contour from a different speaker does not degrade consistently neither the similarity nor the naturalness of the converted speech. On the other hand, the proposed VC system with time and F0 preservation together with adversarial loss on varying speaker identities improves the scores in almost all cases compared to the baseline. The overall similarity to the target speaker is good (MOS=4.06) and the naturalness of the conversion is fair (3.18). The improvement in similarity is particularly substantial for the female speakers (MOS=4.23) while in the same time the difference in naturalness between the male and female conversions is much less pronounced (MOS=3.16 for male speakers and MOS=3.21 for female

---

[1]Converted speech samples are available at: http://recherche.ircam.fr/anasyn/obin/VC_EUSIPCO22.html

TABLE II
MOS OBTAINED FOR THE DIFFERENT VC SYSTEMS.

VC system	Male-to-Male		Female-to-Female		TOTAL	
	Similarity	Naturalness	Similarity	Naturalness	Similarity	Naturalness
Orig: target speaker	4.92	4.94	4.98	4.97	4.98	4.96
F0 cond.	3.90	**3.38**	3.93	2.85	3.92	3.09
F0 cond w/ adv. same id	3.90	3.15	3.94	2.91	3.96	3.14
F0 cond w/ adv. diff id	**3.91**	3.16	**4.23**	**3.21**	**4.06**	**3.18**

speakers). This indicates that the add of the discriminator not only helps to improve the naturalness of the converted speech (by suppressing perceptible artifacts) but also increases the similarity to the target speaker. Also, the use of varying speaker identities with the adversarial loss improves the scores in all cases compared to using only the true identity of the speaker. This is probably mainly due to the fact that the discriminator is being more efficient when subject to a larger variety of sentence and speaker identities.

IV. CONCLUSION

This paper presented a S2S-VC algorithm which allows to preserve the timing and the F0 of the source speaker during conversion. Moreover, an adversarial module is added so that the S2S-VC can learn both from real speech samples as well as manipulated ones. Experimental evaluation on the VCTK speech database shown that the F0 is effectively preserved during conversion and that the adversarial module clearly helps to reduce the audible artifacts of the conversion as well as improve the perceived identity of the converted speech. Further research will investigate the manipulation of timing and F0 during training, in addition to the speaker identity.

V. ACKNOWLEDGEMENTS

REFERENCES

[1] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "The Voice Conversion Challenge 2016," in *ISCA Interspeech*, 2016, pp. 1632–1636.

[2] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Speaker Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 195–202.

[3] Yi Zhao, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, and Tomoki Toda, "Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," in *ISCA Interspeech*, 2020, pp. 80–98.

[4] Jaime Lorenzo-Trueba, Fuming Fang, Xin Wang, Isao Echizen, Junichi Yamagishi, and Tomi Kinnunen, "Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data," in *Speaker Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 240–247.

[5] B. M. Lal Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2802–2806.

[6] David Ericsson, Adam Östberg, Edvin Listo Zec, John Martinsson, and Olof Mogren, "Adversarial representation learning for private speech generation," in *Workshop on Self-supervision in Audio and Speech at the International Conference on Machine Learning (ICML)*, 2020.

[7] D. Wang, J. Yu, X. Wu, S. Liu, L. Sun, X. Liu, , and H. Meng, "End-To-End Voice Conversion Via Cross-Modal Knowledge Distillation for Dysarthric Speech Reconstruction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, p. 7744–7748.

[8] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 266–273.

[9] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion," in *ISCA Interspeech*, 2019, pp. 679–683.

[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 *IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[11] Seyed Hamidreza Mohammadi and Taehwan Kim, "One-Shot Voice Conversion with Disentangled Representations by Leveraging Phonetic Posteriorgrams," in *ISCA Interspeech*, 2019, pp. 704–708.

[12] Hui Lu, Zhiyong Wu, Dongyang Dai, Runnan Li, Shiyin Kang, Jia Jia, and Helen Meng, "One-Shot Voice Conversion with Global Speaker Embeddings," in *ISCA Interspeech*, 2019, pp. 669–673.

[13] Kaizhi Qian, 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 (ICML)*, 2019, pp. 5210–5219.

[14] Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai, "Non-Parallel Sequence-to-Sequence Voice Conversion With Disentangled Linguistic and Speaker Representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 540–552, 2020.

[15] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2387–2395.

[16] Kaizhi Qian, Zeyu Jin, Mark Hasegawa-Johnson, and Gautham J. Mysore, "F0-Consistent Many-To-Many Non-Parallel Voice Conversion Via Conditional Autoencoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6284–6288.

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

[18] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen, "AttGAN: Facial Attribute Editing by Only Changing What You Want," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[19] Li-Juan Liu, Zhen-Hua Ling, Yuan Jiang, Ming Zhou, and Li-Rong Dai, "WaveNet Vocoder with Limited Training Data for Voice Conversion," in *Interspeech 2018*, 2018, pp. 1983–1987.

[20] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.

[21] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," in *ISCA Interspeech*, 2017, pp. 4006–4010.

[22] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[23] L. Ardaillon and A. Roebel, "Fully-Convolutional Network for Pitch Estimation of Speech Signals," in *Interspeech*, 2019.

[24] J. Yamagishi, C. Veaux, and K. Macdonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," 2019, University of Edinburgh. The Centre for Speech Technology Research (CSTR).