FRAGMENTVC: ANY-TO-ANY VOICE CONVERSION BY END-TO-END EXTRACTING AND FUSING FINE-GRAINED VOICE FRAGMENTS WITH ATTENTION

*Yist Y. Lin** *Chung-Ming Chien** *Jheng-Hao Lin* *Hung-yi Lee* *Lin-shan Lee*

National Taiwan University
College of Electrical Engineering and Computer Science
{r08922048, r08922080, r08922049, hungyilee}@ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

Any-to-any voice conversion aims to convert the voice from and to any speakers even unseen during training, which is much more challenging compared to one-to-one or many-to-many tasks, but much more attractive in real-world scenarios. In this paper we proposed FragmentVC, in which the latent phonetic structure of the utterance from the source speaker is obtained from Wav2Vec 2.0, while the spectral features of the utterance(s) from the target speaker are obtained from log mel-spectrograms. By aligning the hidden structures of the two different feature spaces with a two-stage training process, FragmentVC is able to extract fine-grained voice fragments from the target speaker utterance(s) and fuse them into the desired utterance, all based on the attention mechanism of Transformer as verified with analysis on attention maps, and is accomplished end-to-end. This approach is trained with reconstruction loss only without any disentanglement considerations between content and speaker information and doesn't require parallel data. Objective evaluation based on speaker verification and subjective evaluation with MOS both showed that this approach outperformed SOTA approaches, such as AdaIN-VC and AUTOVC.

Index Terms— voice conversion, any-to-any, Transformer, concatenative, attention mechanism

1. INTRODUCTION

Voice conversion (VC) technologies are to convert the voice produced by a source speaker to sound like being produced by a target speaker. Conventional approaches using Gaussian Mixture Models [1] worked well, though were then outperformed by those based on Artificial Neural Networks (ANNs) [2]. Traditionally, aligned parallel data uttering the same text by different speakers were needed in training, but those data were difficult to obtain. In recent years many parallel-data-free ANN-based models were proposed [3], including CycleGAN-VC [4] and StarGAN-VC [5]. But these models can only perform conversion among a predefined set of speakers. On the other hand, any-to-any VC models aim to convert the voice from and to any speakers even unseen during training, which is much more challenging while much more attractive in real-world scenarios.

Here we present FragmentVC, a parallel-data-free ANN-based approach for any-to-any voice conversion with an encoder-decoder architecture. FragmentVC uses the latent phonetic structure of the utterance of the source speaker obtained with Wav2Vec 2.0 [6] as the query to extract the fine-grained voice fragments in the utterance(s) of the target speaker and fuse them into the desired utterance, all

---

* These authors contributed equally.

based on the attention mechanism of Transformer [7] and achieved end-to-end.

As shown in the lower part of Fig. 1, FragmentVC consists of a source encoder (left in red), a target encoder (middle in blue) and a decoder (right in purple) and is trained with a two-stage training process. The source encoder relies on Wav2Vec 2.0 to obtain the latent phonetic structure of the utterance from the source speaker, and the target encoder extracts spectral features from the log mel-spectrograms of utterances from the target speaker. Given the former as queries, the decoder learns to utilize the Transformer cross-attention to extract voice fragments from the utterance(s) of the target speaker and fuse them into the converted utterance. FragmentVC is directly trained with reconstruction loss only without considering speaker-content disentanglement, but has been shown to generalize well on unseen speakers. Further analysis showed that phonetically similar fragments of the utterances from the source speaker and the target speaker were implicitly aligned in the attention maps, implying that the attention mechanism actually achieved fine-grained unit-selection and concatenation, the relatively coarse versions of which have long been used in text-to-speech [8, 9] and VC [10, 11].

2. RELATED WORKS

Concatenative text-to-speech [8, 9] is to select voice segments from a corpus, concatenate them and smooth the boundaries. A large set of transcribed utterances from a specific speaker is usually needed. Some concatenative methods of VC were proposed earlier [10, 11]. They required parallel data and did one-to-one VC only. But FragmentVC used the Transformer attention to achieve the purpose with voice fragments and perform any-to-any VC without parallel data.

Attention mechanism was used for VC in sequence-to-sequence (seq2seq) models [12] or Transformer networks [13, 14] but both for one-to-one VC. FragmentVC is not only any-to-any, not a seq2seq method, but also not using the attention to learn a monotonic alignment between the source and the converted utterances. FragmentVC implicitly learns to use the latent phonetic information from the source utterance to extract and fuse the fine-grained voice fragments of the target utterances.

Any-to-any VC was achieved earlier in AdaIN-VC [15] and AUTOVC [16]. They both relied on the disentanglement of content and speaker information within an utterance with a content encoder and a speaker encoder. AdaIN-VC adopted instance normalization to convert from the source speaker to the target speaker, while AUTOVC used a pretrained speaker encoder to obtain speaker information and used an information bottleneck to limit the leakage of the source speaker information. Instead, FragmentVC directly attends on the utterances from the target speaker, extracting and fusing the proper
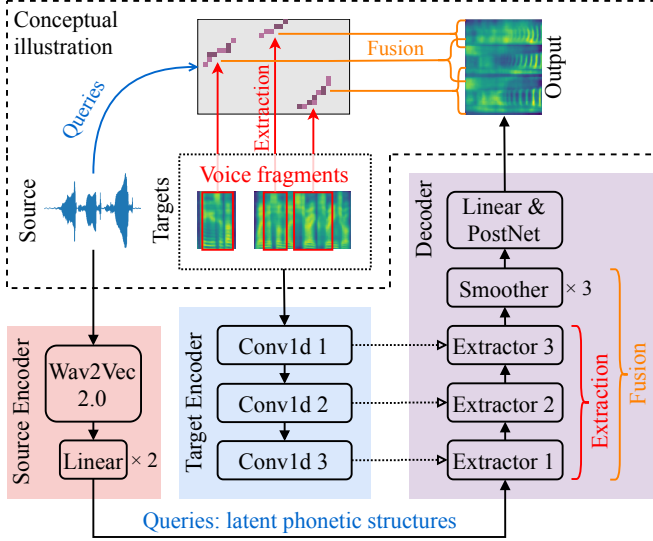
Fig. 1: The overall model architecture (lower half) and the concept of the process (higher half) of FragmentVC. The dotted arrows between the target encoder and the extractors indicate the attention.
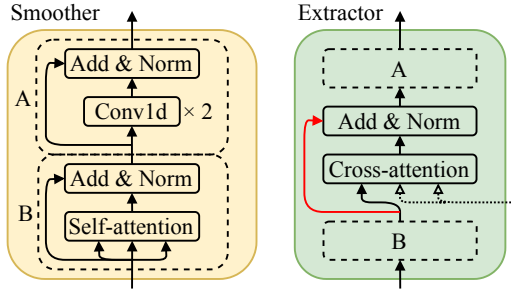

Fig. 2: The model architecture of the extractor and the smoother. Extractor 1 does not have the residual connection (red arrow).

fine-grained voice fragments, which was shown to perform comparably or better than AdaIN-VC and AUTOVC.

3. METHODOLOGY

The overall model as shown in the lower part of Fig. 1 is composed of a source encoder, a target encoder, and a decoder. The concept of the process is illustrated in the upper half of the figure.

3.1. Source encoder

Wav2Vec 2.0 [1] [6] is used as a pretrained feature extractor to extract 768-dimensional speech representations of the source utterance, with model weights fixed during training. The 768-dimensional features are then converted to 512-dimension by two linear layers with ReLU activation, to be used as the input to the decoder.

---

[1] The pretrained model of Wav2Vec 2.0 Base trained on LibriSpeech [17] without finetuning on transcribed speech was used throughout this paper.

3.2. Target encoder

The log mel-spectrograms of utterance(s) from the target speaker are concatenated and fed into the target encoder, which is composed of three ReLU-activated 1d-convolution layers, for extracting the voice fragments to be used below.

3.3. Decoder

The decoder is composed of a stack of extractors and smoothers, followed by a linear projection and a Tacotron-2-styled PostNet [18], to predict the log mel-spectrogram for the desired output voice in a non-autoregressive manner. Both the extractors and smoothers are Transformer [7] layers with two attention heads and hidden size being 512. The extractors are equipped with both self-attention and cross-attention that attend on the output of the target encoder, while the smoother contains self-attention only. Considering the high correlation among adjacent features in speech, the feed-forward layers in each Transformer layer are replaced by a convolutional network [19], with the detailed architecture shown in Fig. 2.

The extractors are based on the latent phonetic structure of the source speaker utterance by cross-attention to extract fine-grained voice fragments from target speaker utterances, and then fuse them up to produce the output voice. The cross-attention is purposely designed to have a U-Net-like [20] architecture as in Fig. 1 as explained below. Because Extractor 1 (lowest in Fig. 1) is to construct the highest-level phonetic structure of the output utterance based on the source representations, while Conv1d 3 (also lowest in Fig. 1) of the target encoder is supposed to produce the most abstractive spectral information too, so Extractor 1 attends on the output of Conv1d 3. On the other hand, Extractor 3 (highest in Fig. 1) is to offer only slight modifications or minor adjustments in the spectrogram, so it attends on the features obtained from Conv1d 1 (highest in Fig. 1) of the target encoder. The smoothers finally take the output of the extractor stack to further smooth the output utterances.

Since some residual speaker information is inevitably carried by the Wav2Vec features, we remove the residual connection over the cross-attention module in Extractor 1, as shown in red in Fig. 2, to restrict the information flow from the source encoder to the decoder. This was verified to be useful in the results in Sec. 5.

3.4. Loss function

Only L1 loss between the predicted and the ground-truth log mel-spectrograms is used to train the entire network, including the process of extracting and fusing the voice fragments, in an end-to-end manner, except for the fixed pretrained Wav2Vec model. No additional loss term is needed as done previously [3, 4, 5]. FragmentVC was shown to obtain the target speaker characteristics with the target encoder.

3.5. Two-stage training

We adopted a two-stage training scheme for the proposed model. In the first stage, the same single utterance from a training speaker is used as the input of both the source encoder and the target encoder, and the training goal is to reconstruct the log mel-spectrogram of the utterance. Although the spectrogram input of the target encoder is identical to the reconstruction target, there is no way for the attention mechanism to obtain the important absolute position information of the various acoustic events in the utterance from the spectral features. So the model has to learn end-to-end to align the hidden

structures between the Wav2Vec feature space from the source encoder and the spectral feature space from the target encoder by extracting and fusing the voice fragments. Though the Wav2Vec features are not only very abstractive but also very different from the spectral features, we believe the linear layers in the source encoder provides some of the conversion between the two very different feature spaces. Preliminary experiments showed the model was able to almost perfectly reconstruct the log mel-spectrogram in this way. However, if a target utterance different from the source utterance was given at this stage, we found that the model was able to extract voice fragments properly and the output sounded like spoken by the target speaker, but the converted result sounded rather discontinuous because the model didn't learn to perform such a task. This is why the second training stage as explained below is needed.

In the second stage, we concatenate the spectrograms of 10 utterances and feed them into the target encoder, while feeding a single utterance to the source encoder, all produced by the same speaker, with the training goal being reconstructing the spectrogram of the source utterance. In the beginning, the source utterance (also the reconstruction target) is always included in the 10 target utterances, but the probability that it is included then linearly decays to zero as the training proceeds, so as to learn incrementally the scenario that the source and targets are getting more and more different. In order to preserve the already well-trained attention modules, we reduce the learning rates of the source encoder, the target encoder and the extractors by 100 times in this stage, while the learning rates of the other components are left unchanged.

4. EXPERIMENTAL SETUP

4.1. Training Setup

The whole CSTR VCTK Corpus (VCTK) [21] with 44 hours of speech produced by 109 speakers was used to train the FragmentVC model. All the utterances were resampled to 16k Hz before fed into the Wav2Vec model. The hop and window sizes for mel-spectrogram computation were chosen to ensure the length of the mel-spectrogram was identical to that of the Wav2Vec features.

FragmentVC was optimized with the AdamW optimizer [22] (with learning rate 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and $weight\_decay = 0.01$) for 250k steps with batch size being 16, the first 50k steps for first stage while the rest for second stage training. In the second stage, the probability that the source utterance was included in the target utterances linearly decayed from one to zero from the 50k-th step to the 150k-th step and remained zero until the end. We also used cosine annealing for learning rate scheduling, with 500 steps of linear warmup. The code will be released online [2].

4.2. Other SOTA approaches

For performance comparison, we took AdaIN-VC [15] and AUTOVC [16] as the SOTA any-to-any VC approaches, with officially released pretrained models also trained over VCTK, adopted in experiments.

4.3. Vocoder

For each model, we trained a WaveRNN-based speaker-independent vocoder [23] to convert the log mel-spectrograms to waveforms. A subset of LibriTTS [24] (*train-clean-100*) and the CMU Arctic databases (CMU) [25] were used to train the vocoders for 150k steps.

4.4. Test scenarios

Two voice conversion scenarios were evaluated: (1) seen-to-seen (s2s) for the conversion between speakers in the VCTK training dataset and (2) unseen-to-unseen (u2u) for the conversion between unseen speakers from the CMU dataset. In both cases, we randomly sampled 1000 testing pairs within VCTK (s2s) and CMU (u2u), each including 1 utterance from a source speaker and 10 utterances from a target speaker for all models considered. Considering that both VCTK and CMU are parallel datasets, to match the real-world scenario, the utterances with the same transcription as the source utterance were not sampled as the targets.

4.5. Evaluation metrics

A speaker verification (SV) system [3] was adopted for objective evaluation of the converted speaker characteristics, as done in a previous work [26]. The SV system took a converted utterance as input and generated a fix-dimensional embedding. The conversion was considered successful if the cosine similarity between the embeddings of the target utterance and the converted utterance exceeded a predefined threshold. The threshold was determined based on the equal error rate (EER) of this SV system over the considered dataset. The SV Accuracy was then the percentage of successful conversion.

For subjective evaluation of the perceptual quality, we conducted two Mean Opinion Score (MOS) tests. In the first test, each subject was asked to listen to an authentic utterance from the target speaker and a converted result, and then to score from 1 to 5 regarding how confident they would consider these two utterances to be produced by the same speaker (5 being absolutely same and 1 absolutely different) [26]. In the second test, the subjects were given a converted utterance or a vocoder-reconstructed authentic utterance and asked to score from 1 to 5 how natural the utterance sounded. For every model considered, the converted results of the same 50 randomly sampled testing pairs out of the previously used 1000 pairs were evaluated, each by at least 5 subjects. The scores were then averaged and reported with the 95% confidence intervals for each model. Such subjective evaluation was conducted with the u2u scenario only, which is considered much more important than s2s for any-to-any VC.

5. EXPERIMENTAL RESULTS

5.1. Performance analysis

The results of the speaker verification accuracy are listed in Table 1, where columns (a) (b) are respectively for the proposed FragmentVC and that without the second stage training. With the s2s scenario (first row), FragmentVC (with second stage training or not) achieved comparable performance in speaker characteristics conversion with AdaIN-VC (column (c)), while leaving AUTOVC (column (d)) far behind. As for the u2u scenario (second row), which is much more important for any-to-any VC considered, FragmentVC clearly outperformed other models. On the other hand, column (e) (f) (g) are results of using 1, 5 or 20 utterances of the target speaker (10 used in column (a)). It can be found the performance was clearly improved or better voice fragments can be extracted with more utterances, but even only 1 target utterance worked very well.

Table 2 lists the two MOS scores for exactly the same columns (a) (b) (c) (d) as in Table 1. It can be found the proposed FragmentVC achieved significantly higher scores than the others (column

---

Table 1: The SV accuracy (%) for seen-to-seen (with EER being 5.6%) and unseen-to-unseen (with EER being 2.6%) scenarios. 10 target utterances were used except for columns (e) (f) (g).

Scenarios	Comparison with other SOTAs				Different # of target utterances			Ablation studies			
	(a) *Proposed	(b) *-ss	(c) AdaIN-VC	(d) AUTOVC	(e) 1 tgt	(f) 5 tgt	(g) 20 tgt	(h) *-ca	(i) *-lrr	(j) *-rrc	(k) *-unet
s2s	94.8	94.7	97.8	39.3	83.1	91.4	94.0	75.0	74.3	78.6	90.8
u2u	92.5	99.8	87.1	19.0	86.5	92.7	93.7	36.5	74.8	67.9	83.2

*-ss: w/o second stage training, k tgt: k target utterance(s), *-ca: w/o cross-attention, *-lrr: w/o learning rate reduction in the second stage training, *-rrc: w/o removal of the residual connection in Extractor 1, *-unet: w/o U-net-like attention.
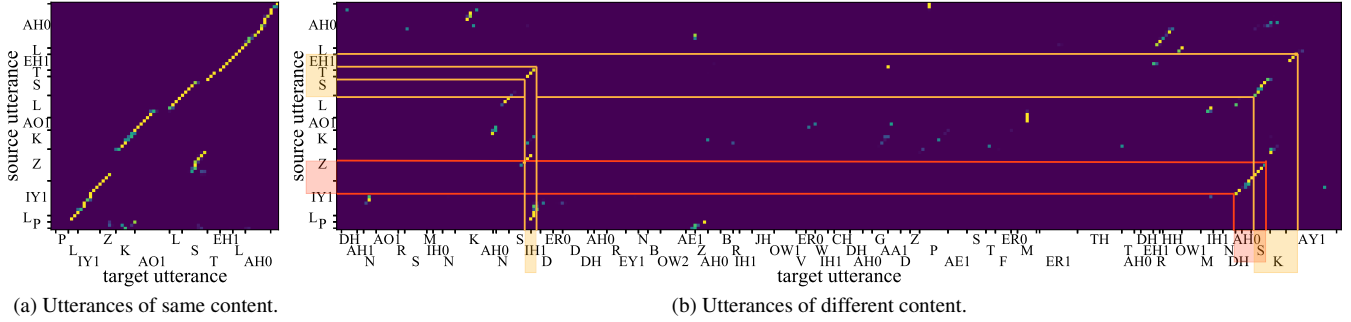

(a) Utterances of same content. (b) Utterances of different content.

Fig. 3: Attention maps between Extractor 3 of the decoder and Conv1d 1 of the target encoder. From VCTK, utterance *p225_001* of speaker *p225* as the source utterance, utterance *p227_001* (in (a)) and utterance *p227_016* (in (b)) of speaker *p227* as the target utterances.

Table 2: The MOS on unseen-to-unseen conversion. *Auth.* stands for vocoder-reconstructed authentic utterances.

MOS	(a) *Proposed	(b) *-ss	(c) AdaIN-VC	(d) AUTOVC	(e) Auth.
Sim.	3.32±0.15	3.81±0.15	2.75±0.15	2.12±0.14	–
Nat.	3.26±0.12	2.73±0.11	2.52±0.12	2.31±0.12	4.09±0.12

(a) v.s. (c) (d)), while the vocoder-reconstructed authentic utterances (column (e)) served as the upper bound. Column (b) of Table 2 is for the proposed FragmentVC but without the second stage training. The second row and columns (a) (b) in Table 2 clearly verified the second stage training is very important to achieve more natural converted utterances, although at the price of slightly degrading the target speaker characteristics (column (a) v.s. (b) in the first row of Table 2 and the second row of Table 1). With careful listening, we found that without the second stage training, the converted utterances sometimes sounded unsmooth, or even discontinuous, which is why the naturalness score was much lower.

5.2. Ablation studies

In column (h) of Table 1, we removed the cross-attention in the decoder and replaced the target encoder with a speaker encoder trained with GE2E loss [27] over LibriSpeech, VCTK, and LibriTTS. This speaker encoder generated a fixed-size speaker embedding for every target utterance, which was then averaged, linearly projected, and fed into the decoder. The obviously inferior performance (column (h) v.s. (a)) showed the importance of the attention mechanism.

Columns (i) (j) (k) are respectively for the cases without the learning rate reduction in the second stage training, without the removal of the residual connection in Extractor 1, and without the U-Net-like architecture, namely all extractors attending on the output of Conv1d 3. The results verified these approaches are important.

5.3. Attention analysis

Here we plotted and tried to analyze the attention maps between the extractors and the target encoder. Two example attention maps between Extractor 3 of the decoder and Conv1d 1 of the target encoder are shown in Fig. 3. Since there are two attention heads in the extractor, we showed in these maps the root-mean-square of the attention values at each position for easier visualization. For space limitation, only one utterance was used as the target here. More attention plots are available on our demo page [4].

In Fig. 3a, for the source and target utterance having the same content but being spoken by different speakers, a diagonal pattern can be easily seen in the attention plot, or the alignment between the two utterances was properly achieved considering their phonetic structure. In Fig. 3b, for the same source utterance but a different longer target utterance, it can be found the extractor is able to attend on phonetically similar frames in the target utterance, including finding acoustically similar voice fragments (e.g. within /IY1 Z/ and /AH0 S/ in the lower right corner, and /S T EH1/ and /S IH1/ & /S K AY1/ in the upper left & right corner) for constructing the converted utterance.

6. CONCLUSION

Here we propose to achieve any-to-any voice conversion by extracting and fusing voice fragments to construct the desired utterances with attention. The objective and subjective evaluations verified the proposed FragmentVC achieved comparable or even better performance than other SOTA approaches. How the Wav2Vec representations actually contributed to the model, if it can be jointly learned, or if it is possible to find some other pretrained representations for the purpose are yet to be investigated. However, we believe such attention-based approaches will be very useful for VC because of their easy implementation, flexibility and explainability.

---

[4] https://yistLin.github.io/FragmentVC

7. REFERENCES

[1] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[2] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3893–3896.

[3] 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," in *Proc. Interspeech 2018*, 2018, pp. 501–505.

[4] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100–2104.

[5] 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.

[6] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017.

[8] Robert E Donovan and Ellen M Eide, "The ibm trainable speech synthesis system," in *Fifth International Conference on Spoken Language Processing*, 1998.

[9] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1996, vol. 1, pp. 373–376 vol. 1.

[10] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 313–317.

[11] Z. Jin, A. Finkelstein, S. DiVerdi, J. Lu, and G. J. Mysore, "Cute: A concatenative method for voice conversion using exemplar-based unit selection," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5660–5664.

[12] 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)*, 2019, pp. 6805–6809.

[13] Wen-Chin Huang, Tomoki Hayashi, Yi-Chiao Wu, Hirokazu Kameoka, and Tomoki Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," 2019.

[14] R. Liu, X. Chen, and X. Wen, "Voice conversion with transformer network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7759–7759.

[15] Ju chieh Chou and Hung-Yi Lee, "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," in *Proc. Interspeech 2019*, 2019, pp. 664–668.

[16] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 5210–5219.

[17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[18] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[19] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," 2020.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. 10 2015, vol. 9351, pp. 234–241, Springer International Publishing.

[21] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017.

[22] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[23] Jaime Lorenzo-Trueba, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, Roberto Barra-Chicote, Alexis Moinet, and Vatsal Aggarwal, "Towards Achieving Robust Universal Neural Vocoding," in *Proc. Interspeech 2019*, 2019, pp. 181–185.

[24] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," 2019.

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

[26] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in Neural Information Processing Systems 31*, pp. 4480–4490. Curran Associates, Inc., 2018.

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