TRAINING ROBUST ZERO-SHOT VOICE CONVERSION MODELS WITH SELF-SUPERVISED FEATURES

Trung Dang[⋆], *Dung Tran*[†], *Peter Chin*[⋆], *Kazuhito Koishida*[†]

[⋆] Boston University   [†] Microsoft Corporation

ABSTRACT

Unsupervised Zero-Shot Voice Conversion (VC) aims to modify the speaker characteristic of an utterance to match an unseen target speaker without relying on parallel training data. Recently, self-supervised learning of speech representation has been shown to produce useful linguistic units without using transcripts, which can be directly passed to a VC model. In this paper, we showed that high-quality audio samples can be achieved by using a length resampling decoder, which enables the VC model to work in conjunction with different linguistic feature extractors and vocoders without requiring them to operate on the same sequence length. We showed that our method can outperform many baselines on the VCTK dataset. Without modifying the architecture, we further demonstrated that a) using pairs of different audio segments from the same speaker, b) adding a cycle consistency loss, and c) adding a speaker classification loss can help to learn a better speaker embedding. Our model trained on LibriTTS using these techniques achieves the best performance, producing audio samples transferred well to the target speaker's voice, while preserving the linguistic content that is comparable with actual human utterances in terms of Character Error Rate.

Index Terms— voice conversion, unsupervised, zero-shot

1. INTRODUCTION

The task of alternating the voice characteristic of someone's speech as if spoken by another target speaker has always been an interesting branch of style transfer learning, with applications ranging from entertainment to protection of the speaker's identity. Traditional voice conversion systems require training on parallel data and/or audio transcription, and only target one or several voices of seen speakers. However, recent architectures have been able to perform unseen-to-unseen, or zero-shot, voice conversion in a completely unsupervised manner. Some previously proposed architectures include conditional variational autoencoders (CVAE) [1, 2], generative adversarial networks (GAN) [3, 4, 5], or autoencoders with feature disentanglement [6, 7].

Recently, works on self-supervised representation learning (SSL) in the speech domain have been able to produce speech units by solving a contrastive task [8, 9]. These speech features can be obtained without training on labeled data, which perfectly suits the voice conversion task, where it is important to learn phonetic representations, but not necessary to map them to actual phonemes. Despite producing competitive results compared to traditional approaches, two challenges remain in the SSL approach. First, with the content embedding fixed, VC models with a frame-by-frame decoder become less flexible. While the speech units can be efficiently obtained from speech features of a lower resolution, vocoder benefits from a smaller window size and a higher sampling rate. Moreover, the speaker embedding needs to be more feature-rich, both well representing the speaker's voice characteristic and distinguishing itself from embeddings by other speakers.

To address the aforementioned problems, in this paper, we propose a VC model with a length resampling decoder that works with length-incompatible input and output features, allowing the self-supervised acoustic model and the vocoder to be trained on their task-optimal speech features. We also demonstrate that training on pairs of segments from the same utterance and adding a cycle-consistency loss can help to learn a lower variant speaker embedding, which leads to higher audio quality. Moreover, we show that training a speaker encoder jointly with a speaker classification loss produces utterances with better-transferred voice characteristics. We combine these training techniques and demonstrate performance improvement in terms of objective and subjective evaluation, on the VCTK dataset and the LibriTTS dataset.

2. RELATED WORKS

In this section, we revisit recent works on unsupervised models that can perform zero-shot voice conversion.

CVAE and GAN-based   Conditional Variational Autoencoders (CVAEs) have been successfully applied to the VC task [1, 2]. While it allows models to be trained without parallel data, its quality is usually poor due to over smoothed decoder outputs. On the other hand, Generative Adversarial Networks (GANs), with their distribution matching property,

---

are shown to produce better audio quality [3, 4, 5], though they are notoriously hard to train.

Autoencoders with Feature Disentanglement   Voice conversion can ideally be achieved by pulling out speaker-related information from an utterance and plugging in information from another speaker. For this to be feasible, we desire to disentangle speaker embedding (global, time-invariant) with content embedding (local, time-variant). This can be achieved by precise embedding size tuning [6], adaptive instance normalization [7, 10], or vector quantization [11]. These frameworks have been shown to have better performance compared to CVAE and GAN-based counterparts. The disentanglement can also be specifically achieved with β-VAE-style loss [12] or minimizing mutual information [13]. However, some architectures rely on an information bottleneck that can affect the audio quality. These methods also do not explicitly leverage linguistic content in the utterance, thus sometimes produce unclear or distorted samples.

Using Self-supervised Acoustic Features   Early works have utilized speaker-independent acoustic features from a pre-trained ASR for the VC task [14]; however, this requires an extensive amount of labeled data for training the ASR. Recent works show that self-supervised speech representations can be used instead [15, 16, 17]. Since the content embedding is fixed as an input to the VC model, it is important to learn a good speaker embedding. Our work aims to design an architecture that can be efficiently used in conjunction with existing acoustic feature extractors and vocoders. We also propose techniques to produce speaker embeddings that are more useful to the task.

3. METHODOLOGY

First, let us introduce the conventional framework. Let $\boldsymbol{X}_u$ and $\boldsymbol{X}_v$ be the source and target audio samples. Let $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$ denote the speech features (e.g. mel filter banks or MFCCs). We desire to produce speech features of the same length as $\boldsymbol{x}_u$ (we denote the length as d_{out}). The content embedding of the source utterance $\boldsymbol{c}_u$ is obtained from the raw audio: $\boldsymbol{c}_u = E_{\text{c}}(\boldsymbol{X}_u) \in \mathbb{R}^{d_{\text{inp}} \times d_c}$, while the speaker embedding of the target utterance $\boldsymbol{s}_v$ is obtained from the speech features: $\boldsymbol{s}_v = E_{\text{s}}(\boldsymbol{x}_v) \in \mathbb{R}^{d_s}$, where E_c and E_s are the content and the speaker encoder, respectively, d_{inp} is the length of the content embedding, and d_s, d_c are the embedding sizes. A decoder $D(\boldsymbol{c}, \boldsymbol{s}) : \mathbb{R}^{d_s} \times \mathbb{R}^{d_{\text{inp}} \times d_c} \mapsto \mathbb{R}^{d_{\text{out}} \times d}$ produces the d-dimensional speech features for audio generation from $\boldsymbol{c}_u$ and $\boldsymbol{s}_v$. During training, since parallel data is not available, the samples used for the source and the target utterance are identical, thus $\boldsymbol{X}_u = \boldsymbol{X}_v = \boldsymbol{X}$, $\boldsymbol{c}_u = \boldsymbol{c}_v = \boldsymbol{c}$, and $\boldsymbol{s}_u = \boldsymbol{s}_v = \boldsymbol{s}$. The self-reconstruction loss is defined as $\mathcal{L}_{\text{self-same}} = \|\boldsymbol{x} - D(\boldsymbol{c}, \boldsymbol{s})\|_2$

Within the above framework, in subsection 3.1, we propose a length resampling decoder, which enables the VC model to operate on inputs and outputs of different (but proportional) length (i.e., $d_{\text{in}} \neq d_{\text{out}}$). We also cover, in the 3.2 and 3.3 subsections, three proposed techniques aiming to train a better speaker encoder.

3.1. Length resampling decoder

In the conventional architecture, the decoder receives the content embedding and produces speech features, which are used to generate audio, either through a reconstruction algorithm or a neural-based vocoder. A problem arising when generating speech features for audio generation from self-supervised acoustic features is the length mismatch. While vocoder may benefit from a higher sampling rate and lower window size, speech representation can be more effectively obtained from features with a longer window size, or directly from raw audio. Though the sequence length of the input d_{inp} and the output d_{out} of the decoder are different, they are proportional. To map sequences of fixed length ratio, we propose using a sequence of upsampling layers to produce high-resolution features and perform a pooling average at the end. For example, if the content encoder receives 16kHz audio samples and produces features with 320x reduction, and the vocoder produces 24kHz audio samples from speech features with a window size of 300, the $d_{\text{inp}} : d_{\text{out}}$ ratio is 5:8. We can upsample the decoder input by a factor of 8, and perform an average pooling of stride 5. This allows the self-supervised acoustic model and the vocoder to be trained independently on their task-optimal preprocessing configurations.

3.2. Self-reconstructing from pairs of utterances

A majority of prior works train the model with $\boldsymbol{x}_u = \boldsymbol{x}_v = \boldsymbol{x}$. While $\boldsymbol{x}$ should be self-reconstructed from its own features, i.e. $\boldsymbol{x} = D(E_c(\boldsymbol{x}), E_s(\boldsymbol{x}))$, for feature disentanglement, we find this training schema different from the inference setting when $\boldsymbol{x}_u \neq \boldsymbol{x}_v$. Moreover, speaker embeddings could have access to content information used for self-reconstruction, which is undesirable. We modify the training process by allowing the model to be trained on pairs of audio segments from the same speaker. Specifically, the loss becomes

$$\mathcal{L}_{\text{self-diff}} = \|\boldsymbol{x}_u - D(\boldsymbol{c}_u, \boldsymbol{s}_v))\|_2 + \|\boldsymbol{x}_v - D(\boldsymbol{c}_v, \boldsymbol{s}_u))\|_2 \quad (1)$$

Since for the same speaker, the voice characteristic also varies with the spoken context, we only sample segments from the same utterance.

3.3. Enforcing consistency and separability for speaker embeddings

While the speaker embedding can be obtained from a speaker verification/classification model pre-trained on a large corpus with speaker ids [6], it has been shown that a jointly trained
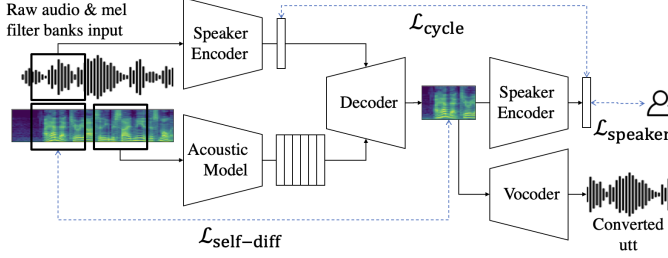
Fig. 1. Our proposed methods during the training stage. Two segments sampled from the same utterance are used to extract the speaker embedding and the acoustic features. The decoder fuses the speaker embedding into each decoding frame, also resamples it to match the length of features required for the vocoder. A cycle-consistency loss $\mathcal{L}_{\text{cycle}}$ and a speaker classification loss $\mathcal{L}_{\text{speaker}}$ are optionally used.

speaker embedding without using the speaker identity is more suitable for VC tasks [18].

To encourage the speaker embedding to be invariant for the same speaker, we add a loss term for cycle-consistency. Let x'_u and x'_v be the speech features produced by the decoder (i.e. $x'_u = D(c_u, s_v)$ and $x'_v = D(c_v, s_u)$). The cycle-consistency loss is defined as

$$\mathcal{L}_{\text{cycle}} = \|s_u - s_v\|_2 + \|s_u - E_s(x'_u)\|_2 + \|s_v - E_s(x'_v)\|_2 \quad (2)$$

The cycle-consistency loss pushes speaker embeddings of segments of the same speaker closer. Speaker embeddings of utterances by different speakers should remain distinguishable since they must share mutual information with the corresponding utterances for the self-reconstruction to be feasible. However, minimizing this loss may also make the speaker embeddings of different speakers less separable. We resolve this by additionally utilizing *speaker id* information. A projection head is added on top of the speaker embedding for classification. Let y represent the one-hot embedding of the speaker of sampled utterances, and $\mathcal{L}_{\text{xent}}$ be the cross-entropy loss. The speaker id loss is as follows

$$\mathcal{L}_{\text{speaker}} = \mathcal{L}_{\text{xent}}(\text{Proj}(s_u), y) + \mathcal{L}_{\text{xent}}(\text{Proj}(s_v), y) \quad (3)$$

Figure 1 provides an illustration of our method.

4. EXPERIMENTS

4.1. Experimental Setup

Baselines We compare our method with other self-supervised architectures that can perform zero-shot voice conversion. We use publicly available checkpoints of AutoVC[1] and FragmentVC[2] trained on VCTK. We also follow instructions to train the AdaIN-VC baseline.

Datasets We adopt the VCTK dataset [19], which is commonly used in training VC models. Since the train/test splits of VCTK are inconsistent for baselines, we need to resample testing data. Specifically, we randomly sample 10 speakers from the whole dataset, from which we further sample 100 pairs of utterances for testing. We train our model only on the remaining 100 speakers[3]. Note that some baselines use additional datasets, for example, Auto-VC uses a speaker embedding model pre-trained on VoxCeleb, or FragmentVC uses a wav2vec feature extractor pre-trained on LibriSpeech. Our results also use LibriSpeech/LibriTTS for the acoustic feature extractor and the vocoder. For a more integrated setting, we also set up a LibriTTS benchmark in which the whole VC pipeline is trained on LibriTTS[4]. Using LibriTTS enables us to test models on utterances recorded in a more diversified and realistic setting. We sample 1,000 pairs of source/target utterances with different speakers from the test-clean set of LibriTTS for evaluation. We also set up a validation set from the dev-clean subset of LibriTTS and report the results with the lowest self-reconstruction loss.

Model Architecture & Training Details We use the wav2vec2 checkpoint released by the authors for the self-supervised acoustic model, which is trained on 960 hours of audio from the LibriSpeech dataset. For the vocoder, we use the official checkpoint released with ParallelWaveGAN [20], both available for the VCTK and the LibriTTS dataset. We do not fine-tune these models, but only train our VC models with input and output following these pre-trained models' data preprocessing recipes. Specifically, the 256-dimensional projected states of wav2vec2 are given to the models (this is different from FragmentVC where the last hidden states of 768-dimensional vectors are used). The vocoder receives decoded speech features with a window size of 300 and generates 24kHz audio. Our speaker encoder consists of 12 stacked 1D convolutional neural network (CNN) layers with and kernel size 5, followed by 12 fully-connected layers, all with a hidden size of 128 and residual connections. Our decoder consists of 12 CNN layers with PixelShuffle layers [21] for 8x upsampling. The output is then down-sampled by a factor of 5. Our VC models are trained on (1) 100 speakers from the VCTK dataset (2) 100 hours and (3) 100+360 hours of clean speech from the LibriTTS dataset [22]. All utterances are randomly cropped or padded to 2-second audio segments. We train our models with 100k steps for VCTK and LibriTTS-100hr, and 500k steps for LibriTTS-460hr, with the batch size of 64 and a constant learning rate of 5×10^{-4}.

4.2. Evaluation

Objective Evaluation We evaluate converted utterances of each model on three tasks: (1) Speaker Verification Accuracy (SVA): We use Resemble.ai, a third-party speaker verification

---

[1]Checkpoints available at https://github.com/auspicious3000/autovc
[2]Checkpoints available at https://github.com/yistLin/FragmentVC

[3]Test utterances could have been used for the baseline checkpoints
[4]LibriSpeech can be alternatively used.

Table 1. Objective (SVA, o-MOS , CER) and subjective scores (NAT, SIM) of the AdaIN-VC baseline and our methods on LibriTTS dataset. For reference, we also report scores from human recorded utterances when available.

	LibriTTS 100hr					LibriTTS 460hr				
	SVA	o-MOS	CER	NAT	SIM	SVA	o-MOS	CER	NAT	SIM
Original	N/A	3.26	4.15	4.56	N/A	N/A	3.26	4.15	4.56	N/A
AdaIN-VC [7]	0.71	2.79	14.87	2.50	1.85	0.86	2.92	15.48	2.71	1.88
$\mathcal{L}_{\text{self-same}}$	**0.91**	2.84	8.08	3.02	**2.11**	**0.97**	**2.94**	7.79	3.39	**2.34**
$\mathcal{L}_{\text{self-diff}}$	0.82	2.76	7.20	2.71	1.85	0.92	2.93	6.09	3.31	2.01
$+\mathcal{L}_{\text{cycle}}$	0.83	2.76	**6.94**	2.91	1.86	0.92	2.90	5.90	3.51	2.08
$+\mathcal{L}_{\text{speaker}}$	0.88	2.87	7.55	3.16	1.95	0.95	2.90	**4.15**	**3.53**	2.15
+both	**0.91**	**2.90**	6.95	**3.41**	1.97	0.95	2.91	4.25	3.31	2.29

Table 2. Objective (SVA, o-MOS , CER) and subjective scores (NAT, SIM) on the VCTK dataset. Utterances produced by our method appears to preserve well the linguistic content, have good perceptual quality, and be similar to the target voice

	SVA	o-MOS	CER	NAT	SIM
AutoVC [6]	0.21	**3.08**	68.25	2.31	1.68
AdaIN-VC [7]	0.46	2.80	24.61	2.58	1.89
FragVC[5] [15]	**0.71**	2.83	32.71	3.24	2.08
$\mathcal{L}_{\text{self-same}}$	0.64	2.88	9.79	3.53	**2.23**
$\mathcal{L}_{\text{self-diff}}$	0.54	2.91	**6.15**	**3.66**	2.12

service to compute the similarity scores between the target and converted utterances. The conversion is considered successful if the score is greater than a threshold, which is decided by the equal error rate (EER) over 100/1000 samples from VCTK/LibriTTS (2) Character Error Rate (CER): We use Speech to Text in Microsoft Azure to obtain the transcript of converted utterances. We compute CER scores with regards to the original transcripts to evaluate the intelligibility of converted utterances. (3) Objective Mean Opinion Score (o-MOS): We train a MOSNet [23] on the human evaluation results of the Voice Conversion Challenge (VCC) 2018, to provide an objective metric that correlates to human's evaluation.

Subjective Evaluation Following prior works, we provide the naturalness score based on the Mean Opinion Score (MOS) and the speaker similarity score. (1) Speaker Similarity (SIM): Each subject is asked to listen to the target and converted utterances, and give to score from 1 to 3, which corresponds to *not similar / unsure / similar*, regarding how confident they would consider these two utterances to be spoken by the same speaker. (2) Naturalness (NAT): Each subject is asked to listen to the converted utterance and give a score from 1 to 5 regarding how natural it sounds. Subjective evaluation is performed on 100 utterances for each

model. Each audio sample is rated by 7 subjects from a pool of 1000+ raters.

4.3. Results

Using VCTK Table 2 shows evaluation results on baselines and our proposed methods using only the self-reconstruction loss. For baseline methods, the CER scores are substantially higher than the actual human speech, and far above the scores of utterances generated with our models. On subjective evaluation, our models outperform the best baseline FragmentVC on NAT score by 0.42 (relatively 13%) for $\mathcal{L}_{\text{self-diff}}$, and on NAT score by 0.15 (relatively 7.2%) for $\mathcal{L}_{\text{self-same}}$.

Using LibriTTS The results for LibriTTS are presented in table 1. Since pre-trained checkpoints are not available, we only train and compare our results with AdaIN-VC. It can be observed that training with $\mathcal{L}_{\text{self-diff}}$ and training with $\mathcal{L}_{\text{cycle}}$ sometimes improve the audio quality, but degrade the speaker similarity score. This suggests that though these techniques can reduce the variance for speaker embeddings from the same speaker, they also make speaker embeddings from different speakers less separable. Our model trained with the speaker classification loss produces a more balanced result between audio quality and speaker similarity. When training on LibriTTS-460hr, our models achieve comparable CER scores with human recorded voices. We encourage our readers to listen to these audio samples[6].

5. CONCLUSION

We demonstrated that high-quality voice conversion systems can be achieved by training on self-supervised speech units with our proposed architectures and training strategies. Our models produce good perceptual quality utterances with similar voice characteristic as the target speaker, while maintaining the linguistic content of the source utterance.

---

[6]https://trungd.github.io/ssl_vc/index.html

6. REFERENCES

[1] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–6, IEEE, 2016.

[2] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *ICASSP*, pp. 5274–5278, IEEE, 2018.

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

[4] Z. Zhang, B. He, and Z. Zhang, "Gazev: Gan-based zero-shot voice conversion over non-parallel speech corpus," *arXiv preprint arXiv:2010.12788*, 2020.

[5] B. Nguyen and F. Cardinaux, "Nvc-net: End-to-end adversarial voice conversion," *arXiv preprint arXiv:2106.00992*, 2021.

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

[7] J.-c. Chou, C.-c. Yeh, and H.-y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *arXiv preprint arXiv:1904.05742*, 2019.

[8] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[9] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[10] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H.-y. Lee, "Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization," in *ICASSP*, pp. 5954–5958, IEEE, 2021.

[11] D.-Y. Wu and H.-y. Lee, "One-shot voice conversion by vector quantization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7734–7738, IEEE, 2020.

[12] M. Luong and V. A. Tran, "Many-to-many voice conversion based feature disentanglement using variational autoencoder," *arXiv preprint arXiv:2107.06642*, 2021.

[13] S. Yuan, P. Cheng, R. Zhang, W. Hao, Z. Gan, and L. Carin, "Improving zero-shot voice style transfer via disentangled representation learning," *arXiv preprint arXiv:2103.09420*, 2021.

[14] H. Lu, Z. Wu, D. Dai, R. Li, S. Kang, J. Jia, and H. Meng, "One-Shot Voice Conversion with Global Speaker Embeddings," in *Proc. Interspeech 2019*, pp. 669–673, 2019.

[15] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H.-y. Lee, and L.-s. Lee, "Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," in *ICASSP*, pp. 5939–5943, IEEE, 2021.

[16] J.-h. Lin, Y. Y. Lin, C.-M. Chien, and H.-y. Lee, "S2vc: A framework for any-to-any voice conversion with self-supervised pretrained representations," *arXiv preprint arXiv:2104.02901*, 2021.

[17] H.-S. Choi, J. Lee, W. Kim, J. Lee, H. Heo, and K. Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[18] T.-h. Huang, J.-h. Lin, and H.-y. Lee, "How far are we from robust voice conversion: A survey," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 514–521, IEEE, 2021.

[19] C. Veaux, J. Yamagishi, and K. MacDonald, "Superseded - cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[20] 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)*, pp. 6199–6203, IEEE, 2020.

[21] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

[22] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[23] 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.