



ALO-VC: Any-to-any Low-latency One-shot Voice Conversion

Bohan Wang¹, Damien Ronssin², Milos Cernak²

¹École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

²Logitech Europe S.A., Lausanne, Switzerland

bohan.wang@epfl.ch, damien.ronssin@gmail.com, milos.cernak@ieee.org

Abstract

This paper presents ALO-VC, a non-parallel low-latency one-shot phonetic posteriorgrams (PPGs) based voice conversion method. ALO-VC enables any-to-any voice conversion using only one utterance from the target speaker, with only 47.5 ms future look-ahead. The proposed hybrid signal processing and machine learning pipeline combines a pre-trained speaker encoder, a pitch predictor to predict the converted speech's prosody, and positional encoding to convey the phoneme's location information. We introduce two system versions: ALO-VC-R, which uses a pre-trained d-vector speaker encoder, and ALO-VC-E, which improves performance using the ECAPA-TDNN speaker encoder. The experimental results demonstrate both ALO-VC-R and ALO-VC-E can achieve comparable performance to non-causal baseline systems on the VCTK dataset and two out-of-domain datasets. Furthermore, both proposed systems can be deployed on a single CPU core with 55 ms latency and 0.78 real-time factor. Our demo is available online¹.

Index Terms: Voice conversion, Real-time, Phonetic Posteriorgrams (PPGs), LPCNet

1. Introduction

The objective of voice conversion (VC) is to convert an utterance from a source speaker into a target speaker's voice while preserving the linguistic content. An any-to-any one-shot VC system has the potential to allow any source speaker to clone the voice of any target speaker using only one reference utterance. Various one-shot VC systems, including VQMIVC [1] and DiffVC [2], have demonstrated promising results in terms of speech quality and speaker similarity. However, when deploying the VC systems on actual hardware, it is critical to consider the low-latency implementation of the systems.

So far, numerous deep learning-based one-shot VC systems that can disentangle the source speaker's identity and content representation and clone the target speaker's voice characteristics have been developed. FragmentVC [3] employs Wav2Vec 2.0 to capture the latent phonetic structure of source speech, a convolution-based speaker encoder to extract speaker embedding from target speech, and the Transformer attention mechanism to predict the log mel-spectrogram of the converted speech. To reduce leakage of the source speaker's content information into the target speaker's representation, VQMIVC [1] introduces vector quantization (VQ) for content representation and mutual information (MI) for optimization during training. Recently, diffusion-based one-shot VC system so-called DiffVC [2] applies a Transformer-based content encoder, speaker encoder, and diffusion-based decoder to achieve guiding results

in speech quality and speaker similarity even on out-of-domain unseen speakers. The aforementioned non-causal systems contain a large number of parameters and use a huge number of future frames look-ahead. Hence, the latency exceeds the limited algorithmic latency of real-time applications (e.g., 60 ms).

FastVC [4] can perform many-to-many voice conversion four times faster than real-time on a single CPU core. However, it relies on future frames and cannot convert source speech in streaming. For practical real-time applications, FastS2S-VC [5], and Streamable VQMIVC [6] are proposed to employ a streaming version of voice conversion. While FastS2S-VC can only convert the source speaker's voice into a limited set of target voices, Streamable VQMIVC can perform real-time streaming any-to-any voice conversion. However, Streamable VQMIVC adds two complex modules to the content encoder and pitch extractor of VQMIVC: (1) a content augments using two 2-D convolution layers and 12 Conformer blocks; (2) a pitch post-processing network consisting of 4 1-D convolution layers used to smooth the extracted streaming pitch. Therefore, Streamable VQMIVC contains many parameters (much more than VQMIVC) and a high level of complexity, making it difficult to deploy on current hardware.

In this work, we propose the any-to-any low-latency one-shot voice conversion (ALO-VC) system based on the assumption that PPGs representation can help develop a high-quality causal system. To achieve better performance of low-latency voice conversion, we incorporate a pitch predictor, as well as causal positional encoding, to provide location information of a phoneme in PPGs. We introduce two versions of ALO-VC systems that can perform low-latency one-shot voice conversion for any-to-any speaker pairs using a look-ahead of 47.5 ms. ALO-VC-R uses a pre-trained d-vector speaker encoder. We further improve the performance by using ECAPA-TDNN [7] speaker encoder, which results in the second version of ALO-VC, ALO-VC-E. These systems contain 8.3 million parameters, which is 3.2 times smaller than VQMIVC. The parameter count of the speaker encoder is not included since it is only used once before streaming inference. Our experimental results demonstrate that ALO-VC-R and ALO-VC-E can achieve higher speech naturalness and speaker similarity than VQMIVC on VCTK [8] dataset and even on two out-of-domain unseen speech datasets. Moreover, they narrow the performance gap between the low-latency one-shot VC system and the diffusion-based one-shot VC system. To our knowledge, they are the first any-to-any low-latency one-shot VC systems capable of achieving comparable performance to non-causal systems on datasets other than VCTK [8].

This paper is structured as follows. We summarise the state-of-the-art low-latency VC systems in Section 2. Then, we introduce the architecture of our proposed ALO-VC systems in Section 3. The experimental results are demonstrated in Section 4.

¹<https://bohan7.github.io/ALO-VC-demo/>

2. Related work

2.1. Existing low-latency VC systems

Low-latency real-time non-parallel voice conversion (LLRT-VC) framework [9] can perform many-to-many voice conversion with 2 lookup frames. It can be executed faster than real-time with a 0.87–0.95 real-time factor (RTF) when deployed on a single CPU core. This LLRT-VC framework is designed based on a cyclic variational autoencoder (CycleVAE) [10, 11] spectral model and multiband WaveRNN with data-driven linear prediction (MWDLP) [12]. In addition, the CycleVAE module is fine-tuned with waveform domain loss from the pre-trained MWDLP. The fine-tuning approach can help LLRT-VC achieve similar performance to non-causal CycleVAE with Parallel WaveGAN. However, for intra-lingual conversion, the LLRT-VC framework is outperformed by cascaded automatic speech recognition (ASR) with TTS (ASR-TTS) [13] and Nagoya University (NU) T23 system [14]. Furthermore, when applying the LLRT-VC framework to new source or target speakers, fine-tuning is required.

Almost causal low latency VC (AC-VC) [15] can achieve comparable speech quality to ASR-TTS [13] with only 57.5 ms future look-ahead and approximately 2 million parameters. AC-VC has been successfully deployed as a desktop application performing real-time any-to-many voice conversion using the JUCE framework². However, the speaker similarity score of AC-VC is lower than that of ASR-TTS [13] and CASIA [16] systems due to the absence of future context. It is hypothesized that future frames are necessary to convert source speech prosody according to the target speaker’s prosody, and future information can help perform the consistent conversion [15]. AC-VC is an any-to-many voice conversion method and thus can only convert the voice of the source speaker to a limited number of target speakers. Therefore, to add a new target speaker, the entire system needs to be retrained.

In this work, we propose two versions of any-to-any low-latency one-shot VC system, which can perform voice conversion using a single utterance from any target speaker. Inspired by AC-VC, we also choose PPGs as the causal content representation. To convert the source speaker’s prosody, a pitch predictor is used to predict the converted pitch trajectory, and pitch variance information is added to the PPGs representations. Moreover, we try to improve speaker similarity over AC-VC by using pre-trained speaker encoders from state-of-the-art speaker verification works.

2.2. LPCNet

LPCNet [17] is a state-of-the-art low-latency neural vocoder. It is composed of a frame rate network and a sample rate network that can synthesize highly quality waveform using only 30 ms future look-ahead. The frame rate network processes each 10 ms frame from the input 18 Bark-scale cepstral coefficients (BSCCs), and 2 pitch parameters (period, correlation). Then, the sample rate network operated at 16kHz generates speech waveform using the output conditioning vector from the frame network, the linear predicted sample, the previously generated excitation sample, and the previous output sample. Despite LPCNet being an auto-regressive model, LPCNet can be implemented in real-time on a single-core CPU and even mobile phones.

Our systems employ LPCNet to generate the converted

speech. Therefore, the converted BSCCs, pitch and pitch correlation for every 10 ms are the output from the conversion model.

3. Proposed methods

Our proposed system, as indicated in Figure 1 (a), consists of five components, including an acoustic model, a pitch extractor based on speech signal processing, a speaker encoder, a conversion model, and LPCNet. In the following sections, we provide a detailed description of each component in our system. Additionally, we present a method to fine-tune the LPCNet model.

3.1. Acoustic model

The purpose of the acoustic model is to extract a speaker independent linguistic content representation of the source utterance. In this work, phonetic posteriorgrams are used. We design our acoustic model based on Conformer [18] architecture, but without multi-head attention layers. Multi-head attention layers can increase the number of parameters and inference latency. As shown in Figure 1 (b), our acoustic model consists of a convolution module, a causal convolution module, causal conformer blocks, and uni-directional LSTM layers. The convolution module, which comprises a single 1-D convolution layer with a 3×3 kernel, detects the local correlation between adjacent frames. The module is followed by a causal convolution layer with ReLU activation, 4 causal Conformer blocks, two uni-directional LSTMs and a time-distributed fully connected layer with softmax activation. To ensure causality in the causal conformer blocks, we perform layer normalization for each time frame independently. Finally, 512-D PPGs representation is extracted by the last LSTM layer.

The acoustic model has a total of 2.7 million parameters and uses a 10 ms future look-ahead frame. The model is trained using cross-entropy loss between the ground-truth phoneme alignments [19, 20] and the output from the acoustic model. Moreover, we trained the acoustic model separately from the conversion model to prevent the source speaker’s identity from being exposed in the PPGs representation.

3.2. Speaker encoder

To obtain speaker embeddings, we use the pre-trained d-vector system [21] from Resemblyzer³ and the pre-trained ECAPA-TDNN [7] model from SpeechBrain toolkit [22] respectively. The d-vector speaker encoder has been trained on VoxCeleb2 and contains 1.4 M parameters. The 14.7 M-parameter ECAPA-TDNN is trained on VoxCeleb1 [23] and VoxCeleb2 [24] datasets. The pre-trained models can reliably separate the identity and content representations of the target speaker. Furthermore, they are trained independently of the acoustic and conversion models, which prevents the leakage of content information. We propose two versions of ALO-VC system based on the two speaker encoders: ALO-VC-R with a d-vector speaker encoder and ALO-VC-E with ECAPA-TDNN.

3.3. Conversion model

The conversion model shown in Figure 1 (c) is responsible for generating converted 18-dimensional Bark-scale cepstral coefficients (BSCCs), as well as converted pitch and voiced/unvoiced flag (VUF) for LPCNet to synthesize the converted speech.

The designed model includes a pitch predictor, two positional encoding modules, time-distributed fully connected lay-

²<https://juce.com/>

³<https://github.com/resemble-ai/Resemblyzer>

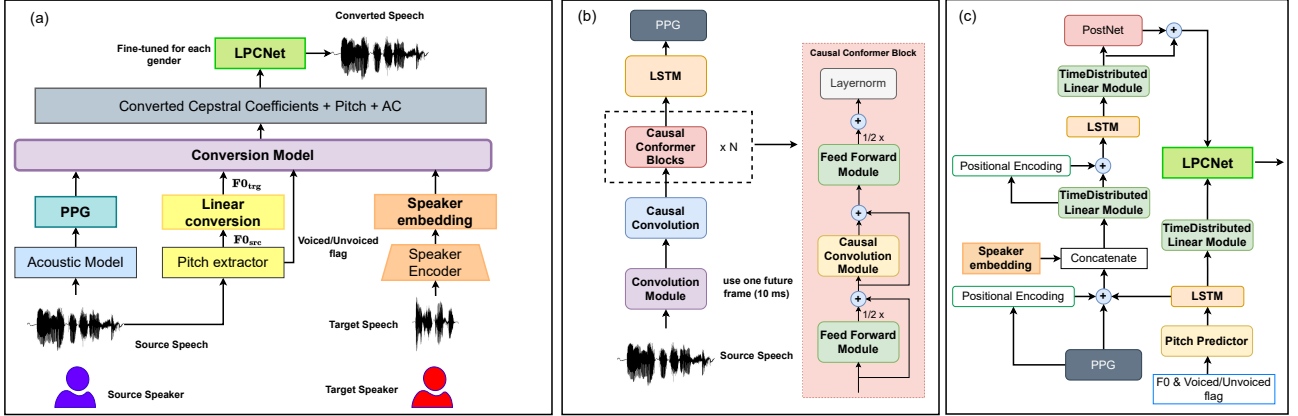


Figure 1: Schematic diagram of low latency PPGs based one-shot VC system. (a) the overview of the low latency one-shot VC system. (b) the schematic of Conformer based acoustic model (c) the schematic of the conversion model

ers, and uni-directional LSTM layers. The pitch predictor employs four 1-D causal convolution layers and one uni-directional LSTM, with converted pitch $F0_{trg}$ and voiced/unvoiced flag (VUF) as inputs. The built-in pitch detector of LPCNet is used to extract F0 and VUF from the source speech. $F0_{trg}$ can be calculated with log-scaled F0 statistics of the target speech as follows:

$$\log F0_{trg} = (\log F0_{src} - \mu_{src}) * \left(\frac{\sigma_{trg}}{\sigma_{src}} \right) + \mu_{trg} \quad (1)$$

where μ_{src} and μ_{trg} are the mean values of the source speaker's and the target speaker's log F0. σ_{src} and σ_{trg} are the variance values of the source speaker's and the target speaker's log F0. We convert the $\log F0_{trg}$ back to $F0_{trg}$ after the pitch conversion. Finally, one time-distributed fully connected layer projects the representations of $F0_{trg}$ and VUF into 2 dimensions. The pitch predictor is utilized to predict both the VUF and target speech prosody. The VUF and pitch variance information are also added to PPGs.

The positional encoding is used to provide location information of a phoneme in PPGs. It can be implemented in streaming inference as follows:

$$\begin{aligned} P(L+1, 2i) &= \sin\left(\frac{L+1}{10000^{2i/d}}\right) \\ P(L+1, 2i+1) &= \cos\left(\frac{L+1}{10000^{2i/d}}\right) \end{aligned} \quad (2)$$

where L is the length of previous input representation. d is the dimension of the output embedding space. i ($0 \leq i < d/2$) is the index to map the position function for a dimension in the output embedding space. We can use L for the positional encoding in streaming inference, which ensures that no future information is used.

Finally, we use a PostNet consisting of three 1-D causal convolution layers with Tanh activation to post-process the output representation. During training, the conversion model is optimized using mean absolute error (MAE) between the ground-truth BSCCs and the output representation from the PostNet. Additionally, MAE between the ground-truth pitch and predicted pitch from the pitch predictor and MAE between ground-truth VUF and predicted VUF are used. The conversion model contains 5.6 M parameters.

3.4. Fine-tuned LPCNet for each gender

To further improve the VC system's performance, the average LPCNet is fine-tuned for VCTK [8] male speakers and female speakers respectively. This involves training the base LPCNet with a 0.0001 learning rate, using speech from only one gender.

4. Experiments

4.1. Experimental setup

Dataset The experiments were conducted on LibriSpeech [25] and VCTK [8] datasets. The acoustic model was trained on the "train-clean-100" and "train-clean-360" subsets of LibriSpeech, which contained a total of 1172 speakers. 40 speakers (20 females and 20 males) from the "dev-clean" subset were not seen during training. The conversion model was trained on 104 speakers from the VCTK corpus. The test set includes six unseen speakers, consisting of three females and three males.

Model configurations We resample each speech into 16kHz sampling rate, and randomly crop or pad a speech into a fixed length (5 seconds). Each audio is also chunked into overlapping frames of 25 ms with a hop size of 10 ms. Finally, for each frame, 13-D Mel Frequency Cepstral Coefficients (MFCCs) are extracted along with their delta values, giving a sequence of 39-D acoustic features.

Training details We trained the acoustic and conversion models for 300 epochs with batch size 32, using the Transformer learning rate scheduler [26]. The Adam optimizer was used with an initial learning rate of $1e-7$. The learning rate increases by $1e-7$ per step for the initial 10,000 training steps and then decreases proportionally to the inverse square root of the step number.

Baseline systems VQMIVC is the main baseline we use in our comparisons. VQMIVC has been shown to outperform other state-of-the-art any-to-any one-shot VC systems. We also report the performance of DiffVC, which has been shown to outperform the PPG-based VC systems on a diverse group of unseen speakers, as indicated in [2]. Our primary aim is not to achieve better performance than DiffVC but rather to reduce the performance gap between low-latency VC systems and state-of-the-art non-causal VC systems.

Metrics In Section 4.2.1, we conducted subjective tests on the converted audio samples of VCTK [8]. The details of the subjective tests are in Section 4.2.1. In Section 4.2.2, we employed widely-used deep learning models to evaluate speech

quality and speaker similarity. Pre-trained MOSNet [27] and MOSRA [28] are used to predict the mean opinion score (MOS) for speech quality. For speaker similarity, we use pre-trained ECAPA-TDNN or d-vector speaker encoder to capture the representations from target speech and converted speech. Cosine similarity is computed for each pair of representations. The average and the standard deviation values of each experiment are reported in Section 4.2.

4.2. Experimental results and analysis

4.2.1. Subjective test on VCTK

We designed the subjective tests following the Voice Conversion Challenge 2020 [29]. 10 participants undertook the subjective tests to rate the speech naturalness and speaker similarity. The evaluated systems are blind to each participant. The listeners initially rated the voice naturalness on a scale from 1 to 5 (Bad, Poor, Fair, Good, Excellent). Then, the participants listened to the converted audio sample and the audio sample from the target speaker and rated the speaker similarity of the pair on a 4-point scale: different speakers (absolutely sure), different speakers (not sure), same speaker (not sure), same speaker (absolutely sure). 2 unseen female speakers (p268 and p301) and 2 unseen male speakers (p252 and p256) are selected for the subjective tests. 30 audio samples converted by each VC system are evaluated. Each sample was assessed 8 times on average for speech naturalness and speaker similarity. The results of the subjective test are reported in Table 1. The results indicate that ALO-VC-R and ALO-VC-E can achieve better speech naturalness and speaker similarity than VQMIVC and reduce the performance gap between low-latency VC systems and DiffVC.

Table 1: *MOS and speaker similarity of subjective test on VCTK.*

Systems	MOS	Similarity
Source speech	4.24 ± 0.85	-
VQMIVC [1]	2.36 ± 1.01	2.13 ± 1.02
DiffVC [2]	3.50 ± 1.10	2.23 ± 1.11
ALO-VC-R	2.60 ± 0.98	2.18 ± 1.00
ALO-VC-E	3.04 ± 0.92	2.35 ± 1.06

4.2.2. Objective tests

We perform the unseen-to-unseen experiment on the LibriSpeech dataset to evaluate our systems on a distribution that differs from that of VCTK. We consider 6 unseen speakers including 3 females (1988, 2277, and 2412) and 3 males (251, 652, and 777). There are 30 pairs and 150 converted audio samples. The performance for each system is reported in Table 2.

Table 2: *Objective tests of the unseen-to-unseen VC. ALO-VC systems outperform non-causal VQMIVC in quality and speaker similarity.*

Systems	Naturalness		Cosine similarity	
	MOSNet [27]	MOSRA [28]	Speech brain [22]	Resemblyzer
LibriSpeech unseen speakers cross-dataset test				
VQMIVC [1]	3.28 ± 0.51	3.52 ± 0.62	0.12 ± 0.10	0.60 ± 0.00
DiffVC [2]	3.73 ± 0.47	4.79 ± 0.28	0.28 ± 0.17	0.76 ± 0.00
ALO-VC-R	3.94 ± 0.40	4.49 ± 0.30	0.18 ± 0.14	0.66 ± 0.00
ALO-VC-E	3.85 ± 0.40	4.59 ± 0.17	0.27 ± 0.14	0.64 ± 0.10
Internal unseen speakers cross-dataset speakers test				
VQMIVC [1]	3.21 ± 0.28	2.78 ± 0.70	0.15 ± 0.10	0.64 ± 0.00
DiffVC [2]	3.61 ± 0.35	4.58 ± 0.26	0.30 ± 0.10	0.84 ± 0.00
ALO-VC-R	4.10 ± 0.31	4.34 ± 0.30	0.18 ± 0.10	0.74 ± 0.00
ALO-VC-E	3.97 ± 0.37	4.45 ± 0.24	0.28 ± 0.10	0.73 ± 0.00

We conduct one last experiment on an internal speech dataset to evaluate the proposed systems for real-world application. This experiment involved 6 unseen speakers (3 females and 3 males) who are not native English speakers. There are 30 pairs and 150 converted audio samples. Table 2 exhibits the performance of each system. Overall, the results demonstrate our proposed systems can outperform VQMIVC in speech naturalness and speaker similarity, and narrow the performance gap between low-latency VC systems and DiffVC.

4.2.3. Real-time factors and latency

We measure the runtime performance of our proposed systems on a single core of an Intel(R) Core(TM) i7-6850K CPU clocked at 3.60 GHz. The latency for each frame and real-time factor including input/output, PPG extraction, pitch conversion, speech conversion and waveform generation are reported in Table 3. All the results are for converting 5 seconds source speech. As shown in Table 3, our proposed systems could be used in real-time applications.

Table 3: *Latency and real-time factor of the proposed systems.*

Systems	Latency	Real-time factor
VQMIVC [1]	Not causal	1.08
DiffVC [2]	Not causal	63.15
ALO-VC-R	55.4 ms	0.78
ALO-VC-E	55.3 ms	0.78

5. Discussion

In our one-shot VC experiments detailed in Section 4.2, the duration of the target speaker data varies from 1 to 20 seconds. Our experimental results in Section 4.2 showcase the effectiveness of our pitch conversion method across varying target audio lengths. These results further indicate that ALO-VC-R and ALO-VC-E can outperform VQMIVC on all three datasets. This is probably attributed to the pitch variance information incorporated by the pitch predictor, and the location information conveyed by the causal positional encoding. Moreover, the use of PPGs representation can help develop a causal system. Our work intends to encourage further research in real-time voice conversion tasks, which could facilitate the creation of voice avatars and recovery of impaired voice. Furthermore, the combination of PPGs representation and diffusion model for a causal system remains an unexplored area for future research.

6. Conclusion

This work proposes two versions of the any-to-any low-latency one-shot voice conversion (VC) system, so called ALO-VC-R and ALO-VC-E, which use only 47.5 ms future lookahead. We also introduce a pitch predictor to predict the converted speech prosody, which helps the low-latency systems convert source speech pitch without future information. The pre-trained ECAPA-TDNN speaker encoders from state-of-the-art speaker verification works further improve speaker similarity. Our thorough evaluation with three datasets demonstrate ALO-VC-R and ALO-VC-E can outperform VQMIVC (a non-causal VC system), and close the performance gap to DiffVC (a non-causal diffusion based VC system). The proposed systems can perform voice conversion with 55 ms latency, and faster than real-time on a single CPU core.

7. References

- [1] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-Shot Voice Conversion," in *Proc. Interspeech 2021*, 2021, pp. 1344–1348.
- [2] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," *arXiv preprint arXiv:2109.13821*, 2021.
- [3] 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 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5939–5943.
- [4] O. Barbany and M. Cernak, "Fastvc: Fast voice conversion with non-parallel data," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 145–149.
- [5] H. Kameoka, K. Tanaka, and T. Kaneko, "Fasts2s-vc: Streaming non-autoregressive sequence-to-sequence voice conversion," *arXiv preprint arXiv:2104.06900*, 2021.
- [6] H. Yang, L. Deng, Y. T. Yeung, N. Zheng, and Y. Xu, "Streamable speech representation disentanglement and multi-level prosody modeling for live one-shot voice conversion," *Proc. Interspeech 2022*, pp. 2578–2582, 2022.
- [7] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, 2020, pp. 3830–3834.
- [8] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vtck corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.
- [9] P. L. Tobing and T. Toda, "Low-latency real-time non-parallel voice conversion based on cyclic variational autoencoder and multiband wavernn with data-driven linear prediction," *arXiv preprint arXiv:2105.09858*, 2021.
- [10] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-Parallel Voice Conversion with Cyclic Variational Autoencoder," in *Proc. Interspeech 2019*, 2019, pp. 674–678.
- [11] P. L. Tobing, Y.-C. Wu, and T. Toda, "Baseline system of voice conversion challenge 2020 with cyclic variational autoencoder and parallel wavegan," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, Oct. 2020.
- [12] P. L. Tobing and T. Toda, "High-Fidelity and Low-Latency Universal Neural Vocoder Based on Multiband WaveRNN with Data-Driven Linear Prediction for Discrete Waveform Modeling," in *Proc. Interspeech 2021*, 2021, pp. 2217–2221.
- [13] W.-C. Huang, T. Hayashi, S. Watanabe, and T. Toda, "The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 160–164. [Online]. Available: <http://dx.doi.org/10.21437/VCC.BC.2020-24>
- [14] W.-C. Huang, P. L. Tobing, Y.-C. Wu, K. Kobayashi, and T. Toda, "The NU Voice Conversion System for the Voice Conversion Challenge 2020: On the Effectiveness of Sequence-to-sequence Models and Autoregressive Neural Vocoders," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 165–169.
- [15] D. Ronsson and M. Cernak, "Ac-vc: non-parallel low latency phonetic posteriorgrams based voice conversion," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 710–716.
- [16] L. Zheng, J. Tao, Z. Wen, and R. Zhong, "Casia voice conversion system for the voice conversion challenge 2020," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 136–139.
- [17] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.
- [18] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.
- [19] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech Model Pre-Training for End-to-End Spoken Language Understanding," in *Proc. Interspeech 2019*, 2019, pp. 814–818.
- [20] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.
- [21] L. Wan, Q. Wang, A. Papir, and I. L. 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.
- [22] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawlatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.
- [23] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2017-950>
- [24] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.
- [25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [27] 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," in *Proc. Interspeech 2019*, 2019, pp. 1541–1545.
- [28] K. El Hajal, M. Cernak, and P. Mainar, "MOSRA: Joint Mean Opinion Score and Room Acoustics Speech Quality Assessment," in *Proc. Interspeech 2022*, 2022, pp. 3313–3317.
- [29] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinunen, Z.-H. Ling, and T. Toda, "Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion —," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 80–98.