AN INVESTIGATION OF STREAMING NON-AUTOREGRESSIVE SEQUENCE-TO-SEQUENCE VOICE CONVERSION

Tomoki Hayashi[1,2], *Kazuhiro Kobayashi*[1,2], *Tomoki Toda*[2]

[1]TARVO Inc., Japan, [2]Nagoya University, Japan
hayashi.tomoki@g.sp.m.is.nagoya-u.ac.jp

ABSTRACT

Recent advances in sequence-to-sequence (S2S) models have improved the quality of voice conversion (VC), but it requires the entire sequence to perform inference, which prevents using it in real-time applications. To address this issue, this paper extends the non-autoregressive (NAR) S2S-VC model to enable us to perform streaming VC. We introduce streamable architectures such as causal convolution and self-attention with causal masking for the FastSpeech2-based NAR-S2S-VC model. The streamable architecture also tries to convert durations, which are kept as is in conventional real-time VC methods. To further improve the performance of the streaming VC model, we utilize an instant knowledge distillation with a dual-mode architecture, which performs non-causal and causal inference by sharing the network parameters. Through the experimental evaluation with Japanese parallel corpus, we investigate the impact on performance caused by the streamable architecture. The experimental results reveal that the use of future context frames increases latency, but it improves the conversion quality and that the difference in the speaking rate affects the performance of streaming inference.

***Index Terms*—** Voice conversion, streaming, non-autoregressive, sequence-to-sequence

1. INTRODUCTION

Voice conversion (VC) [1] is a technology that converts the source speaker's voice into the target speaker's voice while preserving the linguistic content. With the recent growth of deep learning techniques, attention-based autoregressive sequence-to-sequence (AR-S2S) VC methods have attracted attention [2, 3, 4, 5, 6] and replaced the conventional frame-by-frame VC [1, 7, 8]. The attention-based AR-S2S models enable us to learn the alignment between the source and target sequences in a data-driven manner, allowing the conversion of the duration and prosody components. Although attention-based AR-S2S models have achieved remarkable performance, they cause generation errors and slower computation due to AR architecture, making it challenging to explicitly control prosody components.

To address these issues, our previous study proposed a novel non-autoregressive (NAR) S2S VC method [9]. Our previous model was based on the FastSpeech2 architecture [10], extended with the Conformer architecture [11], making it possible to capture both local and global context information from the input sequence. Furthermore, we introduced variance converters to explicitly convert the source speaker's prosody components such as pitch and energy into the target speaker. The model has achieved more stable, faster, and better conversion than AR-S2S models such as Tacotron 2 [2] and Transformer [5]. However, it requires the entire input sequence as the input due to the property of S2S model, making it difficult to use it for the real-time application.

Toward real-time applications, this paper extends our previous NAR-S2S model to allow performing streaming processing. The contributions of this paper are summarized as follows:

- We extend the model architecture of our previous work [9] to enable us to perform streaming processing for real-time applications. To do this, we replace non-causal convolution layers with causal convolution and introduce the causal masking for multi-headed self-attention layers. This streamable architecture also tries to convert durations, which are kept as is in conventional real-time VC methods [12, 13, 14, 15].
- In order to compensate for the performance degradation caused by the streamable architecture, we introduce the use of future context frames and an instance knowledge distillation using the dual-mode architecture, which shares parameters with causal and non-causal models [16]. Furthermore, we utilize the HiFi-GAN discriminator [17] to improve the quality of the neural vocoder.
- We conduct an experimental evaluation using a Japanese speaker dataset consisting of two speakers of 1,000 utterances. The evaluation results demonstrate the proposed streaming VC model performance and reveal the impact on performance caused by the streamable architecture. Furthermore, we demonstrate the controllability of our model, enabling us to control prosody components according to various use cases.

2. NAR-S2S-VC

In this section, we briefly explain our previous study [9], which proposed NAR-S2S-VC based on FastSpeech2 with Conformer architecture [11]. The overview of the architecture is shown in Fig. 1. The model consists of a source encoder, a duration predictor, a length regulator, variance converters, a target decoder, and a neural vocoder. The source encoder consists of Conformer blocks [11], and it converts the Mel-spectrogram of the source speaker into a hidden representation sequence. The input Mel-spectrogram of the source speaker is normalized to be mean zero and variance one over the source speaker's training data. The duration predictor predicts each frame's duration from the encoder hidden representation. To solve the mismatch between the length of the source and the target, the length regulator replicates each frame of the source components with duration information [18]. The variance converters convert the source speaker's pitch and energy into the target speaker's. Then, the target decoder, which consists of another Conformer block, predicts the Mel-spectrogram of the target speaker with replicated encoder hidden representations, converted pitch, and converted energy, and then the following Postnet [19] refines it. We use the ground-truth of duration, pitch, and energy as the decoder inputs during the

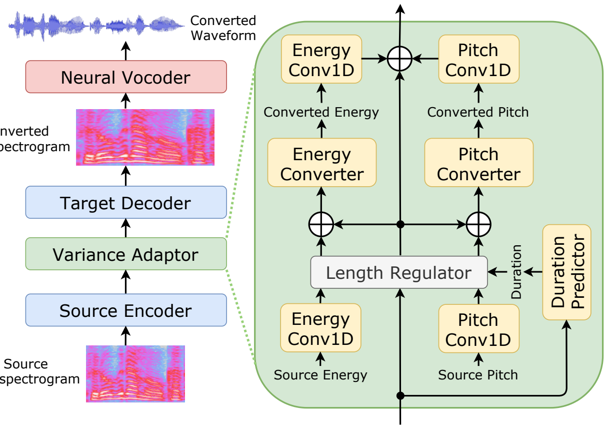


Fig. 1: *An overview of the NAR-S2S-VC model [9].*

training. The encoder and decoder are optimized to minimize the mean absolute error between the predicted Mel-spectrogram and the target Mel-spectrogram. The target Mel-spectrogram is normalized over the target speaker's training data. Finally, we employ Parallel WaveGAN [20] as a neural vocoder and generate the waveform with the converted Mel-spectrogram.

From the components of the above modules, convolution and self-attention layers assume the use of future information to calculate the outputs. Therefore, we need to modify these architectures to perform streaming inference.

3. EXTENSION TO STREAMING NAR-S2S-VC

3.1. Streamable architecture

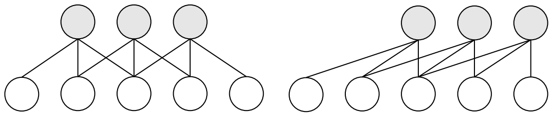
To realize streaming inference, we add the following two modifications to the model architecture: 1) replacing all convolution layers with causal convolution and 2) using causal masking for self-attention layers. The causal convolution is convolution without convolving future information to make sure the model cannot violate the causality, as shown in Fig. 2. The causal masking for self-attention is applying a mask not to attend future information, as shown in Fig 3.

These modifications make it possible to streaming inference, but they do not allow the model to use the future context, causing performance degradation. To address this issue, we use a non-causal convolution as the input layer of the NAR-S2S model to consider a few future context frames. For instance, if we use one future context frame, the kernel size of the input layer is three (from $t-1$ to $t+1$ around the current frame t). This extension slightly increases the latency but improves the conversion quality.

For the neural vocoder, we introduce the HiFi-GAN discriminator [17] and replace the Parallel WaveGAN (PWG) discriminator during the training. Also, we introduce the feature matching loss function and replace the multi-resolutional short time Fourier transform (STFT) loss function [20] with the Mel-spectrogram loss function [17]. These changes lead to slower convergence but better sound quality.

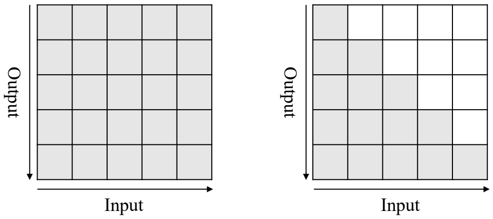
3.2. Instant knowledge distillation with a dual-mode

To further improve the performance of the streamable model, we introduce the instant knowledge distillation with a dual-mode architecture [16]. The model with the dual-mode architecture can perform both causal and non-causal calculations by sharing the parameters.



(a) *Non-causal* (b) *Causal*

Fig. 2: *Non-causal convolution vs. causal convolution.*



(a) *Non-causal* (b) *Causal*

Fig. 3: *Non-causal masking vs. causal masking.*

The calculation of the convolution operation is illustrated in Fig. 4. In the non-causal mode, the convolution operation is performed with all kernels, but the kernel that convolves future frames is disabled in the causal mode. Consequently, the same kernel parameters are used, but the kernel size is different between the causal and the non-causal modes. Also, the same parameters of self-attention modules are used, but the mask for attention calculation is different between the causal and non-causal modes, as shown in Fig. 3.

We perform the instant knowledge distillation using the dual-mode architecture, which minimizes the loss between the non-causal and causal predictions. The final loss function $\mathcal{L}$ is calculated as follows:

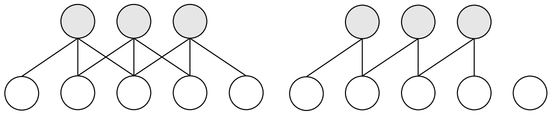
$$\mathcal{L} = \mathcal{L}_{\text{noncausal}} + \mathcal{L}_{\text{causal}} + \mathcal{L}_{\text{dist}}, \tag{1}$$

where $\mathcal{L}_{\text{noncausal}}$ represents the loss calculated between ground-truth and non-causal predictions, $\mathcal{L}_{\text{causal}}$ represents that calculated between ground-truth and causal predictions, and $\mathcal{L}_{\text{dist}}$ represents that calculated between non-causal predictions and causal predictions. Note that each loss is the sum of four losses: mean square errors (MSE) of pitch sequences, energy sequences, duration sequences in log-domain, and L1 loss of Mel-spectrograms.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

To check the performance, we conducted an experimental evaluation using the Japanese parallel speech dataset as the same as our previous work [9]. The dataset consisted of one female speaker and one male speaker of 1,000 utterances. The dataset was recorded in a low reverberation and quiet room, and the sampling rate was set to 24



(a) *Non-causal mode* (b) *Causal mode*

Fig. 4: *Non-causal inference vs. causal inference in the dual-mode architecture. The kernel size is changed from three to two in the causal mode.*

Table 1: *Objective evaluation results, where "STD" represents a standard deviation and "k" represents the number of future context frames.*

Male → Female	MCD ± STD	F_0 RMSE ± STD	CER
Ground-truth	N/A	N/A	7.7
Non-causal	5.88 ± 0.34	0.133 ± 0.023	11.6
Causal ($k = 0$)	6.46 ± 0.28	0.140 ± 0.021	20.3
+ Knowledge dist.	6.18 ± 0.35	0.142 ± 0.020	18.4
Causal ($k = 1$)	6.19 ± 0.31	0.151 ± 0.025	14.1
+ Knowledge dist.	6.03 ± 0.33	0.144 ± 0.025	15.8
Causal ($k = 2$)	6.10 ± 0.30	0.146 ± 0.021	13.3
+ Knowledge dist.	5.95 ± 0.23	0.142 ± 0.023	14.1

Female → Male	MCD ± STD	F_0 RMSE ± STD	CER
Ground-truth	N/A	N/A	6.0
Non-causal	4.99 ± 0.22	0.208 ± 0.036	12.1
Causal ($k = 0$)	5.65 ± 0.22	0.195 ± 0.030	43.3
+ Knowledge dist.	5.59 ± 0.27	0.199 ± 0.032	41.7
Causal ($k = 1$)	5.40 ± 0.25	0.204 ± 0.034	28.1
+ Knowledge dist.	5.25 ± 0.26	0.198 ± 0.031	22.5
Causal ($k = 2$)	5.31 ± 0.31	0.199 ± 0.040	16.0
+ Knowledge dist.	5.13 ± 0.23	0.199 ± 0.032	19.4

kHz. We used 950 utterances for training, 25 for validation, and 25 for evaluation.

We compared three settings: non-causal, causal, and dual-mode models. We extracted an 80-dimensional Mel-spectrogram with 2,048 FFT points, 300 shift points, a Hanning window of 1,200 points, and a Mel-basis range of 80-7600 Hz. The number of encoder blocks and decoder blocks was set to four, and that of heads of multi-headed attention was set to two. The attention dimension was set to 384, and the kernel size of the convolutional module in the Conformer block was 15. The number of channels in each of the duration predictor and variance converters was set to 256. The number of layers of the duration predictor and the energy converter was two, and their kernel size was three. The number of layers of the pitch converter was five, and the kernel size of the pitch converter was five. The reduction factor was set to one for both the source and the target sides in order to reduce the latency for all models. We trained 100k iterations using Noam optimizer [21], and the warmup steps were set to 4,000. For the non-causal model, we used a PWG generator with 400k iterations using PWG discriminator and multi-resolution STFT loss [20]. For the causal and dual-mode models, we used a causal PWG generator with 1.5M iterations using HiFi-GAN discriminator, feature matching loss, and Mel-spectrogram loss [17]. Except for the causality, all models have the same architecture.

4.2. Objective evaluation

At first, we conducted an objective evaluation. As the objective evaluation metrics, we used Mel-cepstral distortion (MCD) [dB], root mean square error of the natural logarithm of F_0 (F_0 RMSE), and character error rate (CER) [%]. To get the alignment between the prediction and the reference, we used dynamic time warping (DTW) [22]. To calculate the MCD, we calculated 0-34 order Mel-cepstrum with SPTK toolkit[1]. The CER was calculated by the Transformer-based ASR model trained on the corpus of spontaneous Japanese (CSJ) [23], which was provided by ESPnet [24][2]. MCD and F_0 RMSE reflect speaker, prosody, and phonetic content similarities, and CER represents the intelligibility and correlates to naturalness [25]. Note that MCD and F0 RMSE are less sensitive to speech discontinuity than CER because MCD and F0 RMSE are calculated frame-by-frame while the CER is calculated by ASR model that takes the sequential information into account.

Table 2: *Statistics of utterance length over the training set. The silence part is trimmed to be the same length.*

Speaker	Avg [sec]	Min [sec]	Max [sec]
Male	3.90	2.61	4.99
Female	4.39	3.02	5.87

The objective evaluation results are shown in Table 1. First, we focus on the results of male-to-female conversion. Compared to the non-causal model, the performance of causal models was degraded if we did not use any future context frames ($k = 0$); however, using future context frames ($k = 1, 2$) improved the performance while it increased the latency (12.5 msec per future context frame). Introducing the instant knowledge distillation with a dual-mode architecture constantly improved MCD, but if we used future context frames, the improvement became limited, especially in terms of intelligibility. This is because the model with the distillation could produce speech without discontinuity, but sometimes it included ambiguous pronunciation. The distillation model with the causal mode tries to mimic the outputs of the non-causal mode. In other words, the model with the causal mode tries to predict missing future context, and the prediction is not perfect. When we do not use the future context frame (k=0), the advantage outweighs the disadvantage; however, with future context frames (k=2), the disadvantage outweighs its advantage since the future context frames give the model enough information, resulting in ambiguous pronunciation.

Next, let us focus on female-to-male conversion results. The tendency was similar to the male-to-female conversion, but CER was drastically degraded when not using future context frames. The degradation was caused by the difference in the speed of utterances. Table 2 showed the statistics of utterance lengths of the training data for each speaker. From Table 2, we can see that the speaking rate of the male speaker is faster than the female speaker, and therefore, the causal model was forced to predict the linguistic contents of the target speaker without the source information. This resulted in discontinuous speech generation, degrading intelligibility. Although using future context frames could solve a part of this issue, it is necessary to develop a method that can handle pairs of speakers of arbitrary speaking rates.

4.3. Subjective evaluation

Next, we conducted subjective evaluation tests on naturalness and speaker similarity to check the perceptual quality. For naturalness, each subject listened to each sample and rated the naturalness on a 5-point scale: 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad. For speaker similarity, each subject listened to pairs of the target sample and the converted sample to judge whether the presented samples were produced by the same speaker with confidence (e.g., sure or not sure). In total, 49 native Japanese speakers participated in the subjective evaluation. The evaluation samples

[1] `http://sp-tk.sourceforge.net/`

[2] `https://zenodo.org/record/4037458`

Table 3: *Subjective evaluation results, where "CI" represents 95% confidence interval.*

Male → Female	Naturalness ± CI	Similarity ± CI
Ground-truth	3.82 ± 0.12	N/A
Non-causal	3.51 ± 0.13	85% ± 5%
Causal (k=0)	3.09 ± 0.12	84% ± 5%
Causal (k=2)	3.23 ± 0.11	74% ± 5%
+ Knowledge dist.	3.09 ± 0.12	82% ± 5%

Female → Male	Naturalness ± CI	Similarity ± CI
Ground-truth	4.60 ± 0.08	N/A
Non-causal	3.64 ± 0.12	70% ± 6%
Causal (k=0)	1.47 ± 0.08	20% ± 5%
Causal (k=2)	2.79 ± 0.12	40% ± 6%
+ Knowledge dist.	2.42 ± 0.12	40% ± 6%

are available online at `https://kan-bayashi.github.io/icassp2022-streaming-vc`.

The subjective evaluation results are shown in Table 3. From the results, in the case of the male-to-female conversion, the causal model with future context frames can achieve reasonable performance; however, there is a gap in the case of female-to-male conversion. This is because even if we use future context frames, it is difficult to solve the issue of the difference in the speaking rate as discussed in Section 4.2, resulting in discontinuous speech. In addition, we could not confirm the improvement when combining the instant knowledge distillation with future context frames in the subjective evaluation. From our perception, samples with the distillation have better sound quality, but some include ambiguous word pronunciation, as discussed in Section 4.2, which may affect the naturalness.

Thanks to the duration predictor, we can make the converted speech slower by multiplying a constant value to the predicted duration. Therefore, we can prepare the slower speech converted by the non-causal model as the target, which may solve the issue of the speaking rate difference. We will work on it in future work.

4.4. Controllability analysis

The constant delay is more preferred than the variable one when speaking with monitoring the converted voice in real-time applications. In addition, sometimes, we may want to keep the pitch contour while converting the speaker identity (e.g., singing a song). Since our S2S model explicitly uses duration, pitch, and energy sequences, it can use the source sequences instead of the predicted target sequences, which can deal with the above demand. The example of conversion with the causal model ($k = 2$) is illustrated in Fig 5. "Conversion with fixed duration" in Fig. 5 represents the converted speech using the source speaker's duration as is without duration conversion. We can confirm that the pitch contour is changed while the source speaker's temporal structure is kept. This enables streaming VC applications to produce the converted speech with the fixed latency, making it easy to speak through the VC system with feedback. "Conversion with fixed pitch, energy, and duration" in Fig. 5 represents the converted speech using the source speaker's pitch, energy, and duration as are. We can see that the converted speech has a pitch contour and temporal structure similar to the source speaker's one, which is helpful for the cases such as singing a song. Consequently, our model can select the components to be converted according to the specific use case.

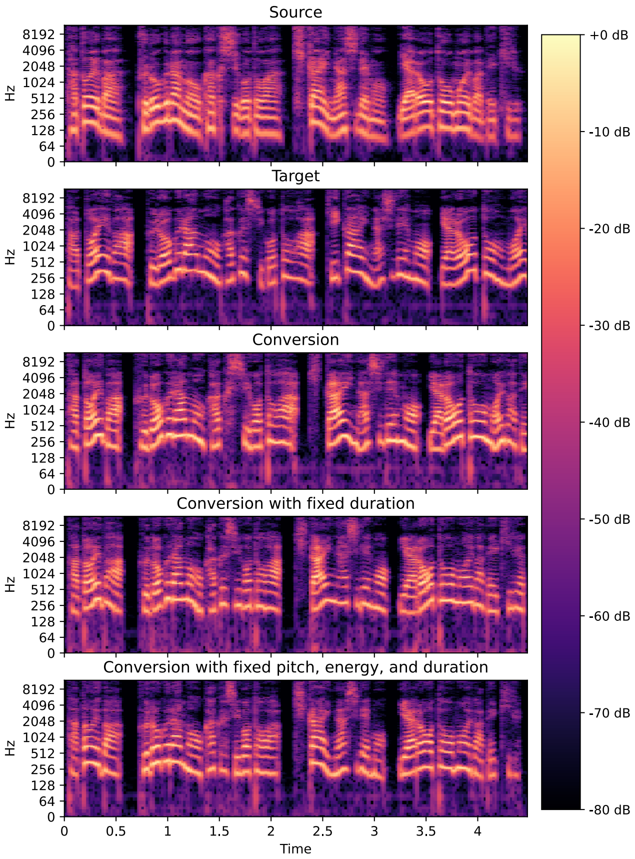


Fig. 5: *Spectrograms of converted examples with the causal ($k = 2$) model on a logarithmic scale. "Source" and "Target" represent the ground-truth speech of source and target speakers, respectively. "Conversion" represents the converted speech, and "Conversion" with fixed components represents the converted speech using the sequence calculated source speech instead of the converted sequence.*

5. CONCLUSION

Toward real-time VC applications, this paper extended the NAR-S2S-VC model to allow performing conversion in streaming fashion and investigated the impact on performance caused by the streamable architecture. The experimental results revealed that the use of just a few future context frames improved the conversion quality while slightly increasing the latency and the speaking rate difference greatly affected the performance in the causal S2S model. Moreover, we demonstrated the controllability of our model, enabling us to control prosody components according to the use cases.

In future work, we will work on the causal training with speech converted by the non-causal model to solve the issue of the difference in speaking rate as well as the development of real-time VC applications.

6. ACKNOWLEDGEMENT

This work was partly supported by JSPS KAKENHI Grant-in-Aid for JSPS Research Fellow Number 19K20295, and JST, CREST Grant Number JPMJCR19A3.

7. REFERENCES

[1] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[2] Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, and Nobukatsu Hojo, "AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019, pp. 6805–6809.

[3] Hirokazu Kameoka, Kou Tanaka, Takuhiro Kaneko, and Nobukatsu Hojo, "ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion," *arXiv preprint arXiv:1811.01609*, 2018.

[4] Jing-Xuan Zhang, Zhen-Hua Ling, Yuan Jiang, Li-Juan Liu, Chen Liang, and Li-Rong Dai, "Improving sequence-to-sequence voice conversion by adding text-supervision," in *Proc. International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019, pp. 6785–6789.

[5] 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," *arXiv preprint arXiv:1912.06813*, 2019.

[6] Hirokazu Kameoka, Wen-Chin Huang, Kou Tanaka, Takuhiro Kaneko, Nobukatsu Hojo, and Tomoki Toda, "Many-to-many voice transformer network," *arXiv preprint arXiv:2005.08445*, 2020.

[7] Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Proc. Interspeech*, 2014, pp. 2514–2518.

[8] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4869–4873.

[9] Tomoki Hayashi, Wen-Chin Huang, Kazuhiro Kobayashi, and Tomoki Toda, "Non-autoregressive sequence-to-sequence voice conversion," in *Proc. International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2021, pp. 7068–7072.

[10] Yi Ren, Chenxu Hu, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "FastSpeech 2: Fast and high-quality end-to-end text-to-speech," *arXiv preprint arXiv:2006.04558*, 2020.

[11] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented Transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[12] Tomoki Toda, Takashi Muramatsu, and Hideki Banno, "Implementation of computationally efficient real-time voice conversion," in *Proc. Interspeech*, 2012, pp. 94–97.

[13] Riku Arakawa, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Implementation of dnn-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device," in *Proc. Speech Synthesis Workshop (SSW)*, 2019, pp. 93–98.

[14] Takaaki Saeki, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Real-time, full-band, online dnn-based voice conversion system using a single cpu.," in *Proc. Interspeech*, 2020, pp. 1021–1022.

[15] Patrick Lumban Tobing and Tomoki Toda, "Low-latency real-time non-parallel voice conversion based on cyclic variational autoencoder and multiband WaveRNN with data-driven linear prediction," in *Proc. Speech Synthesis Workshop (SSW)*, 2021, pp. 142–147.

[16] Jiahui Yu, Wei Han, Anmol Gulati, Chung-Cheng Chiu, Bo Li, Tara N Sainath, Yonghui Wu, and Ruoming Pang, "Dual-mode asr: Unify and improve streaming asr with full-context modeling," *arXiv preprint arXiv:2010.06030*, 2020.

[17] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.

[18] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. Advances in neural information processing systems (NeurIPS*, 2019, pp. 3165–3174.

[19] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *Proc. International conference on acoustics, speech and signal processing (ICASSP)*, 2018, pp. 4779–4783.

[20] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. Advances in neural information processing systems (NeurIPS)*, 2017, pp. 5998–6008.

[22] Stan Salvador and Philip Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[23] Kikuo Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[24] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, et al., "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[25] Rohan Kumar Das, Tomi Kinnunen, Wen-Chin Huang, Zhenhua Ling, Junichi Yamagishi, Yi Zhao, Xiaohai Tian, and Tomoki Toda, "Predictions of subjective ratings and spoofing assessments of voice conversion challenge 2020 submissions," *arXiv preprint arXiv:2009.03554*, 2020.