Investigation of the effectiveness of converted speech auditory feedback in low-latency real-time voice conversion

Kiseki Niwa* and Kazuhiro Kobayashi* and Tomoki Toda*
* Nagoya University, Japan
E-mail: niwa.kiseki@g.sp.m.is.nagoya-u.ac.jp, kobayashi.kazuhiro@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

***Abstract*—Low-latency real-time voice conversion (VC) technique transforming voice characteristics while preserving linguistic content allows users to freely speak with desired target voices. This technique can instantaneously present auditory feedback of the converted speech to the users, enabling them to adapt their vocalization to the VC system. On the other hand, it is also possible that the converted auditory feedback causes any adverse effects, such as stuttering, similar to those caused by delayed auditory feedback (DAF). In this study, we investigate the effectiveness of the converted auditory feedback using eight participants in a real-time VC system with 77ms end-to-end latency. We compare real-time (RT) feedback conditions where participants hear converted speech during utterance with recording (Rec) feedback conditions where participants hear their converted speech after utterance completion. Experimental results demonstrate that the converted auditory feedback yields statistically significant improvements in both speech naturalness (UT-MOS) and speaker similarity. Objective analysis reveals that users subconsciously increase their fundamental frequency (F0) variability to better match the VC model, while Character Error Rate (CER) analysis confirms no adverse DAF-like effects.**

I. Introduction

With recent technological advancements, individuals can engage in recreational activities such as anonymous communication and video streaming in virtual reality spaces using virtual avatars. This enables them to transcend physical constraints such as gender and appearance, allowing them to embody their ideal selves. However, voice remains strongly constrained by physical factors, and there is still no widely established method for easily altering it. As a result, communicating with others using one's ideal voice remains a challenge.

To address this issue, research on real-time streaming voice conversion (VC) technology has been conducted [1], [2], [3], [4]. Real-time VC converts the original voice into a voice with desired characteristics while preserving linguistic content, with low latency. However, there is a trade-off between latency and conversion quality, including aspects such as sound quality and the accuracy of speaker characteristic conversion. Achieving low latency while maintaining high conversion quality, as seen in most of high-quality VC methods that perform utterance-by-utterance conversion [5], [6], [7], [8], [9], [10], [11], remains a challenge. Additionally, it is also challenging to reduce computational cost to achieve low latency processing. For instance, real-time VC based on noncausal conversion models, such as Retrieval-based Voice Conversion[1] (RVC), require at least 300–500 ms of processing time, making it hard to use VC in speech communication. If higher-quality real-time VC can be realized, it is expected to have various applications, such as singing VC [12] and voice enhancement for laryngectomees [13].

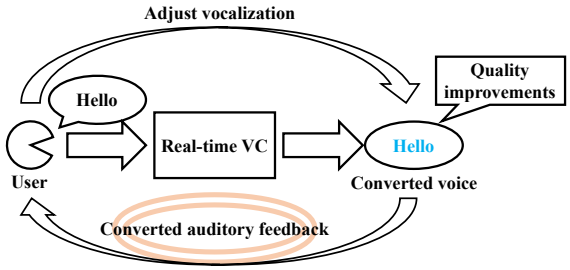


Fig. 1: Overview of quality improvements by adjusting vocalization based on converted auditory feedback.

There are two main approaches to improving conversion quality in VC. The first is to develop a VC model that achieves higher-quality conversion during training. This can be accomplished by using more complex model architectures, tuning hyperparameters, and increasing the amount of training data. The second is to enhance the suitability of input speech for conversion by preventing data drift during inference. In real-time VC, although variations in the recording cannot be completely eliminated, users are expected to improve the quality by appropriately adjusting their vocalization.

In this research, we investigate the effectiveness of users' vocalization control during inference focusing on the effects of the auditory feedback of the converted speech by using a real-time VC system capable of low-latency feedback within 100 ms. Figure 1 illustrates this framework: in addition to hearing their own voice, users receive auditory feedback of the converted speech, allowing them to adjust their vocalization to better match the target speaker's characteristics.

Real-time VC systems introduce additional latency compared to natural auditory feedback, which may affect speech production. For example, CycleVAE with an MWDLP vocoder [3] introduces at least 30 ms of latency due to algorith-

[1]https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI

979-8-3315-7206-8/25/$31.00

mic delays, potentially triggering delayed auditory feedback (DAF) [14]. DAF typically disrupts speech fluency when latency falls between 30 ms and 300 ms, with the strongest effect observed around 200 ms [14], [15].

While DAF can negatively impact speech fluency, previous studies have shown potential benefits of real-time VC. DigitalSpeechMakeup [16] demonstrated that conversing with converted speech can reduce implicit biases, including those related to self-perception. Nevertheless, many aspects related to user voice control and conversion quality improvement remain unclear. In this study, we investigate how converted auditory feedback influences conversion quality through speaking practice experiments using real-time VC. We analyze how users adjust loudness, pitch, and speaking style to improve the converted speech, and find a suitable latency and volume setting of the converted auditory feedback.

II. Real-time VC based on CycleVAE and MWDLP vocoder

CycleVAE with an MWDLP vocoder [17], [3] is a real-time VC system that converts any input speaker's speech to a target speaker's voice with low latency. Like variational autoencoder (VAE)-VC [18], [19], CycleVAE [17] learns the conversion model using non-parallel utterances from multiple speakers. In VAE-VC, the input speech's mel-spectrogram is first encoded into a latent representation. The input mel-spectrogram is then reconstructed by feeding the combined vector of this latent representation and the input speaker's code into the decoder. The encoder and decoder are optimized based on reconstruction loss for both the spectrogram and Kullback–Leibler divergence loss for the latent representation. In contrast, CycleVAE generates a converted mel-spectrogram using the latent representation and a different speaker's code. After re-inputting it into the encoder, a reconstructed mel-spectrogram is obtained using the input speaker's code, capable of using another reconstruction loss that accounts for conversion errors. This approach enables higher-quality non-parallel VC compared to VAE-VC. Additionally, CycleVAE achieves high-quality conversion even in cross-gender transformations and reduces F0 estimation errors by jointly optimizing the excitation signal and mel-spectrogram generation.

The MWDLP vocoder is a low-latency neural vocoder that generates speech waveforms from mel-spectrograms. Based on WaveRNN [20], it reduces computational cost via band-splitting while stabilizing waveform generation through data-driven linear prediction. Moreover, optimizing CycleVAE with MWDLP loss enables end-to-end optimization of the conversion system.

III. Speaking practice experiment

To investigate the effect of using converted auditory feedback in real-time VC, we conduct speaking practice experiments. Additionally, by allowing the adjustment of the latency and volume of the converted speech provided as auditory feedback, we examine the desirable settings for auditory feedback.

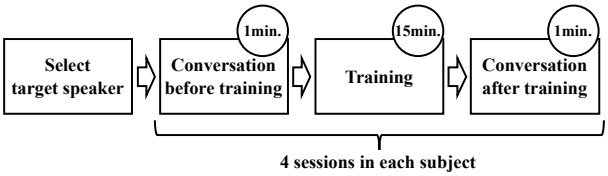


Fig. 2: Speaking practice flow.

TABLE I: Speaking practice order for each experimental group*

Group	Session	Feedback	Target Speaker
RT-Rec	1st practice (15 min)	RT	Easy-to-practice
	2nd practice (15 min)	RT	Difficult-to-practice
	3rd practice (15 min)	Rec	Easy-to-practice
	4th practice (15 min)	Rec	Difficult-to-practice
Rec-RT	1st practice (15 min)	Rec	Easy-to-practice
	2nd practice (15 min)	Rec	Difficult-to-practice
	3rd practice (15 min)	RT	Easy-to-practice
	4th practice (15 min)	RT	Difficult-to-practice

* RT: Real-time feedback condition (converted speech heard during utterance)
Rec: Recording feedback condition (converted speech heard after utterance)

We set a task for participants to improve the quality of the real-time VC system by adjusting vocalization such as the loudness, pitch, and speaking style of their voices during speaking practice. Figure 2 shows the speaking practice flow for each participant. Since the conversion quality varies depending on the target speaker, we properly select target speakers for each participant in the following manner. We prepare a total of eight Japanese speakers (five males and three females) as target speakers. Participants listen to one natural speech utterance from each target speaker and then speak the same sentence using the real-time VC system. Each participant evaluates the "ease of practice" from the perspectives of the naturalness and speaker similarity of the converted speech on a 10-point scale. They evaluate all target speakers and select two: one with the highest score as the "easy-to-practice speaker" and the other with the lowest score as the "difficult-to-practice speaker."

During the speaking practice, participants attempt to adjust their vocalization based on the auditory feedback of the converted speech from the real-time VC system to make the converted speech similar to the target speaker's natural voice. The condition where the converted speech can be heard after the utterance is referred to as the recording (Rec) feedback condition. The condition where real-time feedback of the converted speech is provided during the utterance is referred to as the real-time (RT) feedback condition. Each participant practices speaking under both conditions for the two selected target speakers.

Participants are divided into two groups, and each group follows the practice order shown in Table I. Each participant practices speaking a total of four times, with each practice session lasting up to 15 minutes. During the practice sessions, participants can freely review the content of 10 natural speech

utterances by the target speakers and the corresponding text.

In the RT feedback condition, the delay time and volume can be adjusted. Regarding delay time, an additional delay of up to 10 seconds can be added to the inherent delay of the conversion system. For example, in the RT feedback condition, setting a large delay can simulate a situation similar to the Rec feedback condition, where the converted speech is reviewed after the utterance. Regarding volume, it is adjustable on an 11-point scale from 0 to 10, with the baseline volume set to 5, as determined by the participants before the experiment. For example, setting the volume to 0 would create a condition where the real-time auditory feedback of the converted speech is not used. We ask each participant to manually optimize the delay time and volume. Therefore, it is expected that if the participants prefer the long delay setting or the zero volume setting, the RT feedback is not helpful due to possible DAF adverse effects.

To assess the effect of speaking practice, participants engage in a one-minute conversation using an online meeting system before and after each speaking practice session. The conversation includes a fixed script mimicking light conversation and free conversation with a few questions to the participants. After the Rec feedback condition practice, conversations are conducted without auditory feedback of the converted speech. In contrast, after the RT feedback condition practice, conversations are conducted with real-time auditory feedback of the converted speech. In this case, the delay time and volume of the converted auditory feedback are set to the values adjusted during the speaking practice. Before the speaking practice, no auditory feedback is provided. In the above procedures, all speech and converted speech of the participants are recorded to analyze the impact of converted auditory feedback on the quality of the converted speech.

IV. Evaluation

A. Experimental conditions

Fourteen Japanese males in their twenties, with little or no experience using real-time VC systems, participated in the experiment. The experiment was conducted in a soundproof room. We used the Steinberg UR22C as the audio interface, and audio data was recorded at 48 kHz, 32-bit (floating point) mono. Participants were divided into two groups: the RT-Rec group and the Rec-RT group, with seven participants in each. Each group practiced speaking and recorded conversations with both the "easy-to-practice speaker" and the "difficult-to-practice speaker."

For the real-time VC system, we used CycleVAE with an MWDLP vocoder [3], as described in Sect. II. The conversion model was trained on a total of 62 speakers extracted from multiple datasets in different languages. The VC system's training and conversion were conducted at a sampling rate of 24 kHz. The total latency from microphone input to headphone output was approximately 77 ms. As detailed in Table II, this total latency comprises two main components: inherent algorithmic delays (36.75 ms) and system-level I/O delays (40.48 ms). Target speakers were selected from those with Japanese training data, and all experiments were conducted in Japanese. We compared the quality of the converted speech using five fixed sentences and one or two free sentences recorded before and after each speaking practice session.

TABLE II: Latency analysis of the real-time VC system. The total measured latency from microphone input to headphone output was approximately 77 ms.

Source / Component	Latency (ms)
Algorithmic Delay	
Opus frame length	3.0
STFT frame	13.75
CycleVAE processing	10.0
MWDLP vocoder processing	10.0
System & I/O Delay	
Audio device buffer	1.5
ASIO driver input	8.5
ASIO driver output	11.48
Other (e.g., scheduling, buffering)	≈19.0
Total Estimated Latency	**77.23**

B. Objective evaluation

As objective evaluation metrics, we used the estimated mean opinion score (MOS) based on UT-MOS [21][2], cosine similarity using ECAPA-TDNN speaker embeddings [22][3], and character error rate (CER) using Whisper [23][4]. In this experiment, we evaluated improvements from speaking practice under the RT and Rec conditions separately, without considering the order in which they were administered.

Table III (a) presents the results of the estimated MOS. In both RT and Rec feedback conditions, the MOS values of the converted speech improved due to speaking practice. Furthermore, when comparing the converted speech and the original voice, i.e., a user's input voice before conversion, a larger improvement was observed for the converted speech.

Table III (b) shows the cosine similarity of ECAPA embeddings. We computed the cosine similarity between the converted speech and the target voice, as well as between the original voice and the target voice. Similar to the MOS evaluation, the increase in cosine similarity for the converted speech indicates an improvement in speaker similarity resulting from speaking practice.

Table III (c) presents the CER of the original and converted speech. For CER, a significant improvement was observed for the converted speech, whereas the original voice showed no notable improvement.

These results suggest that practicing with feedback enhances the converted speech in terms of sound quality, speaker similarity, and intelligibility under both the Rec and RT conditions. On the other hand, no measurable changes were observed in the original voice due to practice.

[2]https://github.com/sarulab-speech/UTMOS22

[3]https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[4]https://huggingface.co/openai/whisper-small

TABLE III: Comparisons of objective measures before and after practice. (a) UT-MOS: 1-5 scale, higher is better. (b) Cosine similarity: 0-1 scale, higher is better. (c) CER: percentage, lower is better.

(a) UT-MOS

	Converted		Original	
	Before	**After**	**Before**	**After**
Rec	1.96	2.09	3.50	3.54
RT	1.94	2.13	3.48	3.58

(b) Cosine similarity of ECAPA-TDNN embeddings

	Converted		Original	
	Before	**After**	**Before**	**After**
Rec	0.43	0.47	0.20	0.20
RT	0.42	0.47	0.20	0.20

(c) CER

	Converted		Original	
	Before	**After**	**Before**	**After**
Rec	19.23	12.32	2.31	2.30
RT	13.89	9.89	2.74	2.14

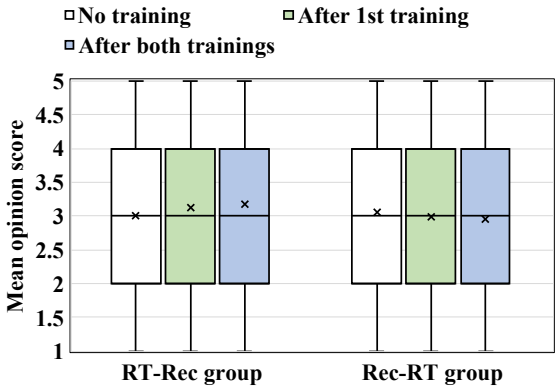
TABLE IV: Preference scores (%) for speaker similarity with 95% confidence intervals.

(a) Easy-to-practice speaker

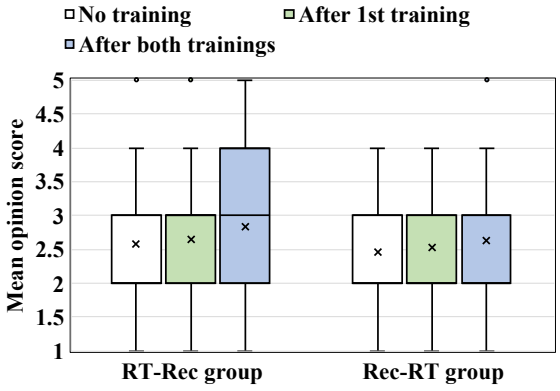
Group	**No training vs. After 1st training**	**After 1st vs. After both training**
RT-Rec	40.5 / 59.5 (±6.6)	43.8 / 56.2 (±6.7)
Rec-RT	21.4 / 78.6 (±5.6)	44.3 / 55.7 (±6.7)

(b) Difficult-to-practice speaker

Group	**No training vs. After 1st training**	**After 1st vs. After both training**
RT-Rec	26.2 / 73.8 (±6.0)	40.0 / 60.0 (±6.6)
Rec-RT	36.2 / 63.8 (±6.5)	37.6 / 62.4 (±6.6)



(a) Easy-to-practice speaker



(b) Difficult-to-practice speaker

Fig. 3: Results on MOS for naturalness.

C. Subjective evaluation

Subjective evaluation experiments were conducted to assess naturalness and speaker similarity. A total of 30 listeners participated in the evaluation.

In the first test, the naturalness of the converted speech was evaluated using a MOS scale. Converted samples were presented to listeners in random order, and they rated the naturalness on a five-point scale: 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for very poor. Each subject evaluated approximately 110 sentences.

The second test assessed identity conversion accuracy through a preference test. First, a natural voice sample of the target speaker was presented as a reference. Then, listeners were presented with a pair of converted speech samples in random order: either "No training" / "After 1st training" or "After 1st training" / "After both training" for the same utterance. They selected the sample that sounded more similar to the reference voice. Each subject evaluated 84 sample pairs and could replay each pair as many times as necessary.

Figure 3 presents the MOS results for naturalness, where the MOS values are shown in "×" marks. Consistent with the UT-MOS evaluation, speaking practice led to an improvement in naturalness under both RT and Rec feedback conditions, with higher scores for "After 1st training" compared to "No training." However, in the Rec-RT group, the average MOS for the "Easy-to-practice speaker" did not improve with continued practice. Additionally, a notable gap in average MOS values was observed between the easy-to-practice and difficult-to-practice speakers. This suggests that speaking practice generally enhances naturalness, though the extent of improvement is limited.

Table IV shows the preference scores for speaker similarity. Significant improvements were observed across all training conditions with increased practice. In particular, "After 1st training" resulted in notable improvements, regardless of RT or Rec conditions. These results demonstrate the effectiveness of the vocalization control in the real-time VC. The users could imitate prosodic features of the target speaker, such as speaking style, which are not directly converted by the real-time VC system, by interactively using the real-time VC system in the speaking practice.

TABLE V: Comparisons of F0 statistics. The values represent the mean F0, with the numbers in the brackets indicating the standard deviation.

(a) Converted

	Before	After
Rec	176.23 (30.10)	182.45 (33.15)
RT	176.54 (30.13)	182.72 (31.35)

(b) Original

	Before	After
Rec	105.91 (32.69)	111.88 (37.69)
RT	108.06 (34.82)	114.70 (37.27)

D. Analysis of the F0 statistics of converted and original voices

We investigated how prosodic features change through structured practice using converted auditory feedback. Table V presents the F0 statistical analysis for both the converted and original voices. For the converted speech, both the mean F0 and its standard deviation increased after practice. This trend is even more pronounced in the original voice; in particular, the substantial change in standard deviation suggests that vocalization facilitated greater F0 variability, which in turn contributed to improved conversion quality.

E. Analysis of the user control in RT feedback condition

We analyzed how users adjusted delay time and volume under the RT feedback condition.

Figure 4 presents the additional delay values manually optimized by participants when using RT feedback. All participants experimented with adjusting the delay during speaking practice. However, the most common final setting, chosen by ten participants, was to add no additional delay. This indicates that while every participant attempted to modify the delay at least once, they ultimately preferred a setting with no added delay. These results suggest that when real-time converted auditory feedback is available, participants tend to minimize delay, practicing under conditions different from those in the Rec feedback condition.

Figure 5 displays the modified volume values set by participants under RT feedback. Regarding volume, no participants set it to zero, thereby eliminating feedback. This suggests that real-time converted auditory feedback is helpful for users during speaking practice.

Based on the above results, several conclusions can be drawn regarding the effectiveness of practicing with converted auditory feedback. First, practicing while listening to converted feedback significantly improves speaker similarity and also slightly improves naturalness of the converted speech. Second, both Rec and RT feedback conditions contribute to enhancing conversion quality. Third, F0 analysis showed that both the mean F0 and its standard deviation increased after practice, with a stronger effect in the original voice, suggesting that vocalization adjustments played a key role in improving conversion quality. Finally, under the RT feedback condition, users actively adjusted parameters such as delay time and volume, indicating that real-time auditory feedback plays a crucial role in optimizing their speaking performance.

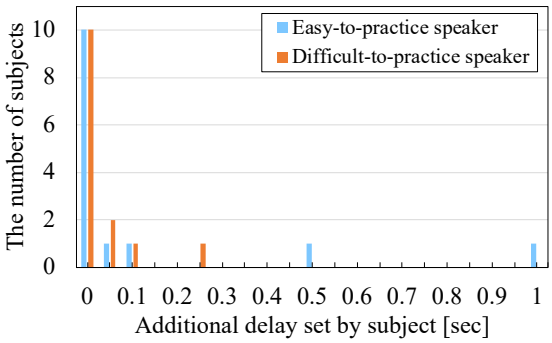


Fig. 4: Results on additional delay set by subjects.

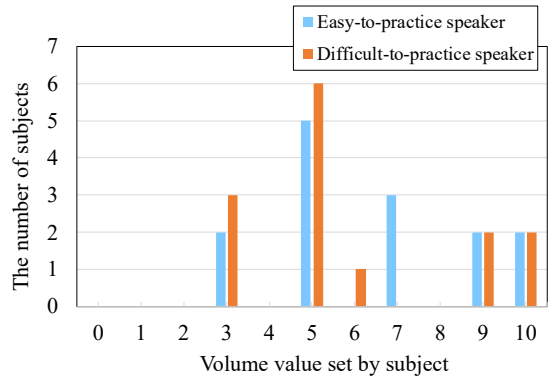


Fig. 5: Results on modified volume set by subjects.

V. Conclusions

In this paper, we investigated the effectiveness of auditory feedback of converted speech in real-time voice conversion. Experimental evaluation through speaking practice revealed that speakers can adjust their speaking styles to improve the quality of the converted speech. Our key finding is that practice with real-time auditory feedback yields statistically significant improvements across both objective and subjective metrics. Acoustic analysis further revealed that speakers adapted by producing more expressive speech, evidenced by an increase in both the mean and standard deviation of their fundamental frequency (F0).

Future work will focus on extending this investigation to diverse populations and other real-time VC systems to generalize our findings. Additionally, long-term practice effects and optimal feedback parameters warrant further investigation.

Acknowledgment

This work is partly supported by JST, CREST, JPMJCR19A3, JST, AIP accelerated research, JPMJCR25U5, and JSPS KAKENHI Grant Number JP25K00374.

REFERENCES

[1] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," *Proc. INTERSPEECH*, Sept. 2012.

[2] R. Arakawa, S. Takamichi, and H. Saruwatari, "Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device," *Proc. SSW*, pp. 93–98, 2019.

[3] 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," *Proc. SSW*, pp. 142–147, 2021.

[4] H. Kameoka, K. Tanaka, and T. Kaneko, "FastS2S-VC: Streaming non-autoregressive sequence-to-sequence voice conversion," *arXiv preprint arXiv:2104.06900*, 2021.

[5] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Trans. TASLP*, vol. 27, no. 3, pp. 631–644, 2019.

[6] W. Huang, H. Luo, H. Hwang, C. Lo, Y. Peng, Y. Tsao, and H. Wang, "Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion," *IEEE Trans. TETCI*, vol. 4, no. 4, pp. 468–479, 2020.

[7] J. Zhang, Z. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Trans. ASLP*, vol. 28, no. 1, pp. 540–552, 2020.

[8] 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," *Proc. INTERSPEECH*, pp. 836–840, 2021.

[9] D. Ronssin and M. Cernak, "AC-VC: Non-parallel low latency phonetic posteriorgrams based voice conversion," *Proc. ASRU*, pp. 710–716, 2021.

[10] T. Hayashi, W.-C. Huang, K. Kobayashi, and T. Toda, "Non-autoregressive sequence-to-sequence voice conversion," *Proc. ICASSP*, pp. 7068–7072, 2021.

[11] W.-C. Huang, S.-W. Yang, T. Hayashi, H.-Y. Lee, S. Watanabe, and T. Toda, "S3PRL-VC: Open-source voice conversion framework with self-supervised speech representations," *Proc. ICASSP*, pp. 6552–6556, 2022.

[12] S. Liu, Y. Cao, N. Hu, D. Su, and H. Meng, "FastSVC: Fast cross-domain singing voice conversion with feature-wise linear modulation," *Proc. ICME*, pp. 1–6, 2021.

[13] K. Kobayashi, T. Hayashi, and T. Toda, "Low-latency electrolaryngeal speech enhancement based on FastSpeech2-based voice conversion and self-supervised speech representation," *Proc. ICASSP*, pp. 1–5, 2023.

[14] A. J. Yates, "Delayed auditory feedback," *Psychological bulletin*, vol. 60, no. 3, pp. 213, 1963.

[15] A. Stuart, J. Kalinowski, M. P. Rastatter, and K. Lynch, "Effect of delayed auditory feedback on normal speakers at two speech rates," *The Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2237–2241, 2002.

[16] R. Arakawa, Z. Kashino, S. Takamichi, A. Verhulst, and M. Inami, "Digital Speech Makeup: Voice conversion based altered auditory feedback for transforming self-representation," *Proc. ICMI*, p. 159–167, 2021.

[17] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational autoencoder," *Proc. INTERSPEECH*, pp. 674–678, 2019.

[18] 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," *Proc. APSIPA*, pp. 1–6, Dec. 2016.

[19] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder," *IEEE/ACM Trans. ASLP*, vol. 27, no. 9, pp. 1432–1443, 2019.

[20] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *Proc. ICML*, pp. 2410–2419, 2018.

[21] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for voicemos challenge 2022," *Proc. INTERSPEECH*, pp. 4521–4525, 2022.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Proc. INTERSPEECH*, pp. 3830–3834, 2020.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*.