INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

E2E-S2S-VC: End-To-End Sequence-To-Sequence Voice Conversion

Takuma Okamoto[1], *Tomoki Toda*[2,1], *Hisashi Kawai*[1]

[1]National Institute of Information and Communications Technology, Japan
[2]Information Technology Center, Nagoya University, Japan

okamoto@nict.go.jp, tomoki@icts.nagoya-u.ac.jp, hisashi.kawai@nict.go.jp

Abstract

This paper proposes end-to-end (E2E) non-autoregressive sequence-to-sequence (S2S) voice conversion (VC) models that extend two E2E text-to-speech models, VITS and JETS. In the proposed E2E-S2S-VC models, VITS-VC and JETS-VC, the input text sequences of VITS and JETS are replaced by the source speaker's acoustic feature sequences, and E2E models (including HiFi-GAN waveform synthesizers) are trained using monotonic alignment search (MAS) without external aligners. To successfully train MAS for VC, the proposed models use a reduction factor only for the encoder. The voice of a source speaker is converted directly to that of a target speaker using a single neural network in the proposed models in an S2S manner; the duration and prosody between the source and target speech can be directly converted. The results of experiments using 1,000 parallel utterances of Japanese male and female speakers demonstrate that the proposed JETS-VC outperformed cascade non-autoregressive S2S VC models.

Index Terms: end-to-end voice conversion, monotonic alignment search, sequence-to-sequence voice conversion

1. Introduction

Voice conversion (VC), which converts the voice of a source speaker to that of a target speaker while preserving the linguistic content of the speech, is an important technology for speech communication [1]. Similarly to text-to-speech (TTS) [2, 3], high-quality VC can be achieved by recently developed neural networks [4]. In typical neural-network-based TTS and VC, input text sequences or sequences of acoustic features of source speakers are first converted to acoustic features of target speakers by acoustic models. Speech waveforms of target speakers are then synthesized from the converted acoustic features by neural vocoders [4]. In TTS [2, 3] and framewise VC [5], end-to-end (E2E) models have recently been proposed. These can directly convert input text sequences or speech waveforms of source speakers to speech waveforms of target speakers using a single neural network without intermediate acoustic features. They outperform conventional cascade models with intermediate acoustic features. However, because the conventional E2E framewise VC performs frame-by-frame conversion, leaving the temporal structure of the source speech unchanged, it is difficult to convert the duration and prosody between the source and target speech; this results in limited conversion quality.

In contrast to framewise VC models, sequence-to-sequence (S2S) VC methods can directly control the duration and prosody between source and target speech. In addition to their use for normal VC [6, 7], S2S methods have been investigated for emotional VC [8], singing VC [9], normal-to-dysarthric VC [10], and electrolaryngeal speech enhancement [11]. Although S2S VC methods typically require parallel data for training, in contrast to non-parallel framewise VC, non-parallel S2S VC has also been investigated [12]. In S2S VC, the temporal alignment between the source and target sequences is trained by the attention mechanism [13] without external aligners. Compared with recurrent neural network (RNN)-based S2S VC models [6, 7], Transformer-based S2S VC models [14], such as Voice Transformer Network (VTN) [15–17], can achieve faster training and higher conversion quality. However, because VTN uses the attention mechanism and must have an autoregressive (AR) structure, its inference speed is slow, and converted voices are sometimes unstable due to the attention prediction error.

To achieve fast and stable S2S VC, a non-AR S2S VC model has been proposed [18]. This model is based on a non-AR neural TTS model, Conformer-based FastSpeech 2 (CFS2) [19], which uses a Conformer [20]-based encoder and decoder instead of Transformer. In non-AR S2S VC, the alignment between the source and target sequences is first obtained by using an AR teacher VTN, similarly to FastSpeech [21] in TTS. The non-AR S2S VC model is then trained using durations predicted by the teacher VTN based on CFS2 with variance (energy and fundamental frequency f_{o}) conversion. Finally, converted speech waveforms are synthesized from converted mel-spectrograms by Parallel WaveGAN (PWG) [22]. The non-AR S2S VC model (CFS2+PWG) can improve both the inference speed and conversion accuracy, compared with VTN [18]. Additionally, CFS2+PWG-based and RNN-based streaming models have been used for real-time applications [23, 24]. However, CFS2+PWG still has the following problems. **P1)** Three types of neural network models (AR teacher VTN, CFS2-based VC model, and PWG) are trained separately, and the prediction errors from the teacher VTN and CFS2 are propagated to the final PWG, resulting in low conversion quality. **P2)** The alignment between the source and target sequences is still unstable despite the use of teacher forcing in the teacher VTN. **P3)** There is scope for improving the inference speed and synthesis quality of the neural vocoder because HiFi-GAN [25] outperforms PWG in this respect. **P4)** In addition to mel-spectrograms, the energy and f_{o} of a source speaker are extracted in the inference.

To simultaneously solve the above problems in CFS2+PWG, this paper proposes E2E non-AR S2S VC (E2E-S2S-VC) models by extending two E2E neural TTS models, VITS [2] and JETS [3]. VITS and JETS are E2E TTS models that perform joint training of Glow-TTS [26] and HiFi-GAN or FastSpeech 2 [27] and HiFi-GAN with monotonic alignment search (MAS) [26] without external aligners. The proposed E2E-S2S-VC models, VITS-VC and JETS-VC, in which the input text sequences of VITS and JETS for TTS are replaced with the source speaker's acoustic feature sequences, were implemented using ESPnet2-TTS [28].

10.21437/Interspeech.2023-2518

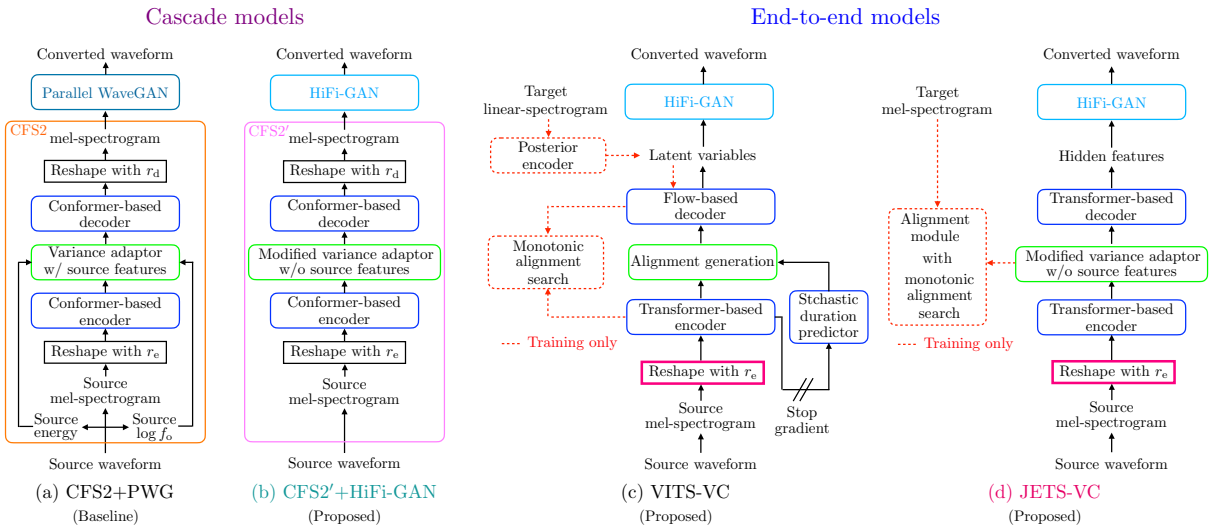

Figure 1: *Network architectures of cascade and end-to-end non-autoregressive sequence-to-sequence voice conversion models. (a) Baseline Conformer-based FastSpeech 2 with Parallel WaveGAN. (b) Proposed Conformer-based FastSpeech 2 with modified variance adaptor and HiFi-GAN. (c) Proposed VITS-based end-to-end model. (d) Proposed JETS-based end-to-end model with modified variance adaptor.* r_e *and* r_d *are the reduction factors for the encoder and decoder, respectively.*

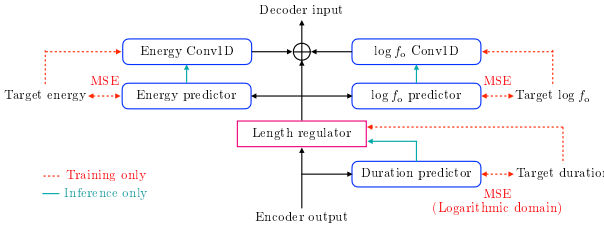

Figure 2: *Network architecture of modified variance adaptor without source energy and* f_o *input while predicting target energy and* f_o.

The results of preliminary experiments indicate that it is very difficult to achieve E2E-S2S-VC by simply applying the original architectures of VITS- and JETS-based TTS to the VC task because the input source feature sequence for VC is much longer than the input text sequence for TTS, and is too long to train MAS between source and target features. To successfully train MAS for VC, the proposed models use a reduction factor (RF)[1] [15–18] only for the encoder. This can stabilize the training of MAS by making the source feature sequence shorter than the target feature sequence. The voice of a source speaker is then converted directly to that of a target speaker using a single neural network in the proposed E2E-S2S-VC models in an S2S manner, such that the duration and prosody between the source and target speech can be directly converted. Experiments were conducted using a Japanese speaker dataset, which consists of 1,000 parallel utterances of one male speaker and one female speaker [18, 23]. The results demonstrate that the proposed JETS-VC outperformed the conventional CFS2+PWG [18] with respect to conversion quality and inference speed. To ensure the reproducibility of this study, some of the speech samples and the PyTorch source code used in the experiments are available on the demo page[2]. Furthermore, the dataset used in the experiments will be published for the purpose of accelerating speech synthesis research [30].[3]

[1]The RF reduces the length of sequences. If the RF is N, a tensor $[B, T, C]$ is reshaped to $[B, T/N, NC]$, where B, T, and C are the batch size, length of sequence, and number of channels, respectively.

2. Conventional VC and TTS models

2.1. Non-AR S2S VC model: CFS2+PWG

The network architecture of the conventional non-AR S2S VC model, CFS2+PWG, is shown in Figure 1(a). The input source mel-spectrograms, energy sequences, and $\log f_o$ sequences are analyzed from source speech waveforms and the input mel-spectrograms (with RF for the encoder r_e) are converted to the hidden features by the Conformer-based encoder. The input energy sequences and $\log f_o$ sequences are then resampled by the length regulator according to the input durations and converted to those of target speech waveforms in the variance adaptor. During the training, durations predicted by a teacher VTN and ground-truth target energy sequences and $\log f_o$ sequences are used. During the inference, they are predicted in the variance adaptor and the predicted energy sequences and $\log f_o$ sequences are added to the resampled hidden features. The hidden features are then converted to the target mel-spectrograms (with RF for the decoder r_d) and the target speech waveforms are synthesized by a separately trained PWG [18].

2.2. E2E TTS models: VITS and JETS

VITS, an extension of Glow-TTS [26], was proposed as an E2E TTS model. During the training of Glow-TTS, the target mel-spectrograms are converted to Gaussian white noise by a Flow-based decoder, and alignment between the hidden features (con-

[2]https://ast-astrec.nict.go.jp/demo_samples/e2e-s2s-vc/index.html In addition to Japanese speech samples, English speech samples trained using CMU-ARCTIC [29] are available on the demo page.

[3]The dataset includes 19,056 utterances of one female speaker and 19,058 utterances of one male speaker (about 18,800 utterances are parallel) for Japanese used in [31]. It also includes 14,000 utterances of one female speaker and one male speaker (about 13,000 utterances are parallel) for English. The sampling frequency of the dataset is 48 kHz.

verted from the input text) and converted white noise is gradually obtained by MAS [26] without external aligners. During the inference, the upsampled hidden features are converted to the target mel-spectrograms by Flow-based inverse transformation. In VITS, the target linear spectrograms are converted to latent variables by a variational auto-encoder (VAE) [32], and the latent variables (instead of mel-spectrograms) are converted both to Gaussian white noise (by a Flow-based decoder) and to the target speech waveforms (by a HiFi-GAN generator). All the network components are jointly trained, and the intermediate latent variables are optimized to minimize the training loss. VITS can achieve higher-quality TTS than the cascade model with Glow-TTS and HiFi-GAN [2].

Compared with VITS, which efficiently uses three types of deep generative models (Flow [33], VAE [32], and GAN [34]), JETS is a simpler E2E TTS model but it achieves higher synthesis quality than VITS [3]. JETS performs joint training of a FastSpeech 2-based acoustic model and a HiFi-GAN-based speech waveform synthesizer with neither intermediate mel-spectrograms nor external aligners; in contrast, CFS2-based TTS models [19, 28] require external aligners. JETS uses an alignment training framework proposed in [35] with MAS, and alignment between the hidden features (converted from the input text sequences) and the target mel-spectrogram sequences is gradually obtained during training, similarly to VITS.

3. Proposed methods

3.1. Alternative cascade model with modified variance adaptor and HiFi-GAN: CFS2′+HiFi-GAN

Before proposing E2E-S2S-VC models, an alternative cascade model is proposed, named CFS2′+HiFi-GAN (Figure 1(b)). HiFi-GAN, which can perform real-time and high-fidelity speech waveform synthesis with a CPU, is used instead of PWG. Additionally, a modified variance adaptor, without source energy and $\log f_{\mathrm{o}}$ input, is proposed (Figure 2). In contrast to the variance adaptor used in CFS2+PWG [18], the hidden features resampled by the length regulator directly predict the target energy and $\log f_{\mathrm{o}}$ sequences without the need for source energy and $\log f_{\mathrm{o}}$ sequences. Using the modified variance adaptor, the analysis of the source energy and $\log f_{\mathrm{o}}$ can be avoided while predicting the target energy and $\log f_{\mathrm{o}}$ sequences only from the source mel-spectrogram sequences, to achieve high-fidelity conversion. In CFS2′+HiFi-GAN, simple repetition-based resampling is applied in the length regulator, as used in CFS2+PWG. As reported in [25, 28] for TTS, the synthesis quality of HiFi-GAN is degraded when predicted mel-spectrograms are used. To improve the final synthesis quality for TTS, both joint fine-tuning of pre-trained CFS2 and HiFi-GAN models (ft) and joint training of CFS2 and HiFi-GAN models from scratch (jt) were investigated [3, 28]. Therefore, both joint fine-tuning and joint training are used for CFS2′+HiFi-GAN. By using fine-tuning, joint training, and HiFi-GAN, the problems of CFS2+PWG, with the exception of **P2**), can be solved.

3.2. E2E-S2S-VC models: VITS-VC and JETS-VC

As explained in Section 2.2, VITS and JETS successfully perform E2E TTS with MAS. To construct E2E-S2S-VC models, VITS- and JETS-based E2E structures are applied to VC models, in models named VITS-VC and JETS-VC (Figure 1(c) and (d)), respectively. In VITS-VC and JETS-VC, the input text sequences used in VITS and JETS for TTS are directly replaced with the source speaker's acoustic feature sequences. As in CFS2+PWG and CFS2′+HiFi-GAN, simple mel-spectrograms are used as the source speaker's acoustic feature sequences in both VITS-VC and JETS-VC. As explained in Section 1, it is difficult to achieve E2E-S2S-VC by simply applying the original architectures of VITS- and JETS-based TTS to the VC task because the input source feature sequence for VC is much longer than the input text sequence for TTS, and is too long to train MAS between source and target features. To successfully train MAS for VC, the proposed models include an RF only for the encoder. By using an E2E framework and stable alignment based on MAS, instead of a cascade framework and unstable alignment predicted by an AR teacher VTN, all the problems of CFS2+PWG, from **P1)** to **P4)**, can be solved.

In JETS-VC, the modified variance adaptor of CFS2′+HiFi-GAN is also used, so that the target energy and $\log f_{\mathrm{o}}$ are predicted only from the source mel-spectrogram sequences, to improve the conversion quality. Similarly to JETS for TTS [3], Gaussian resampling [36] is used in the modified variance adaptor of JETS-VC. In the proposed CFS2′+HiFi-GAN, VITS-VC, and JETS-VC, the same network structures, discriminators for HiFi-GAN, and loss functions are used as in CFS2, VITS, and JETS, respectively. Although CFS2+PWG and CFS2′+HiFi-GAN use RFs for both the encoder and decoder, VITS-VC and JETS-VC use an RF only for the encoder.

4. Experiments

4.1. Experimental conditions

Experiments were conducted to evaluate the proposed E2E-S2S-VC models and compare them with the conventional cascade models. All the S2S VC models were implemented in PyTorch based on ESPnet2-TTS [28] and trained using NVIDIA Tesla A100 GPUs. The experiments were conducted with the Japanese speaker dataset [30] introduced in Section 1. To align the experimental conditions with those of [18], only 1,000 parallel utterances were used in the experiments. The training, validation, and test sets contained 950, 25, and 25 utterances, respectively [18]. The sampling frequency was 24 kHz.

As acoustic features, 80-dimensional mel-spectrograms were analyzed. The FFT, window, and hop sizes were 1024, 1024, and 256, respectively. $\log f_{\mathrm{o}}$ sequences were analyzed by the Dio and Stonemask algorithm [37], implemented in ESPnet2-TTS. To obtain the alignment between the source and target sequences for cascade models, teacher AR VTN models were first trained, following [18]. The RFs for the encoder and decoder were set to 3 in the teacher VTN and all the cascade models, following [18]. CFS2+PWG and CFS2′+HiFi-GAN with joint fine-tuning (ft) and joint training (jt) were then trained as cascade models. To evaluate the effectiveness of the modified variance adaptor, CFS2′+PWG was additionally evaluated. As E2E-S2S-VC models, VITS-VC and JETS-VC were trained, with the RF for the encoder r_{e} set to 2 or 3. All the model configurations were the same as those used in ESPnet2-TTS.

As the objective evaluation criteria, mel-cepstral distortion (MCD), $\log f_{\mathrm{o}}$ root mean square error (RMSE), and character error rate (CER) of automatic speech recognition (ASR) were measured, following [3, 18, 28]. The MCD and $\log f_{\mathrm{o}}$ RMSE were calculated by the ESPnet2-TTS toolkit [28], following [3]. The CER was calculated by a Conformer-based ASR, trained using the CSJ corpus [38] by ESPnet [19]. The real-time factors (RTFs) of all the S2S VC models for inference were measured on an Intel Xeon 6152 CPU (1 core).

Table 1: *Results of objective evaluations. The values in the columns for mel-cepstral distortion (MCD) and* $\log f_{\mathrm{o}}$ *root mean square error (RMSE) are the means and standard deviations. CER, RTF, ft, jt, and* r_{e} *are the character error rate of automatic speech recognition, the real-time factor on an Intel Xeon 6152 CPU (1 core), joint fine-tuning, joint training, and the reduction factor for the encoder, respectively.*

	Male $\longrightarrow$ Female			Female $\longrightarrow$ Male			
Method	MCD [dB]	$\log f_{\mathrm{o}}$ RMSE	CER [%]	MCD [dB]	$\log f_{\mathrm{o}}$ RMSE	CER [%]	RTF
Original	N/A	N/A	1.0	N/A	N/A	1.2	
(Baseline) CFS2+PWG	5.83 ± 0.52	0.25 ± 0.07	3.4	4.74 ± 0.26	0.20 ± 0.04	4.4	3.44
CFS2′+PWG	5.50 ± 0.45	0.24 ± 0.08	3.0	4.76 ± 0.23	0.18 ± 0.06	6.8	3.41
CFS2′+HiFi-GAN (ft)	5.31 ± 0.58	$\mathbf{0.22 \pm 0.07}$	4.4	$\mathbf{4.49 \pm 0.31}$	0.19 ± 0.08	5.8	**0.72**
CFS2′+HiFi-GAN (jt)	5.95 ± 0.60	0.25 ± 0.06	12.7	4.80 ± 0.32	0.22 ± 0.08	12.5	**0.72**
VITS-VC ($r_{\mathrm{e}} = 2$)	5.31 ± 0.43	0.23 ± 0.08	5.2	4.50 ± 0.30	$\mathbf{0.18 \pm 0.05}$	3.2	0.77
VITS-VC ($r_{\mathrm{e}} = 3$)	5.36 ± 0.43	$\mathbf{0.22 \pm 0.07}$	5.4	4.58 ± 0.28	0.19 ± 0.06	5.8	0.76
JETS-VC ($r_{\mathrm{e}} = 2$)	$\mathbf{5.28 \pm 0.42}$	0.23 ± 0.07	**2.2**	4.78 ± 0.36	0.21 ± 0.09	**2.2**	0.79
JETS-VC ($r_{\mathrm{e}} = 3$)	5.38 ± 0.41	0.25 ± 0.09	2.8	4.59 ± 0.25	0.21 ± 0.09	3.0	0.78

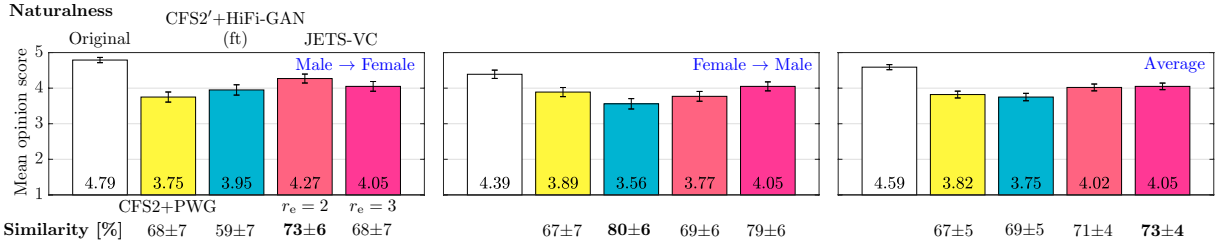

Figure 3: *Results of MOS tests to evaluate naturalness and paired comparison tests to evaluate speaker similarity with 20 listening subjects. The confidence level is 95%.*

To evaluate the converted speech subjectively, mean opinion score (MOS) tests were conducted to evaluate naturalness, and paired comparison tests were conducted to evaluate speaker similarity. According to the results of the objective evaluations, CFS2+PWG, CFS2′+HiFi-GAN (ft), JETS-VC ($r_{\mathrm{e}} = 2$), and JETS-VC ($r_{\mathrm{e}} = 3$) were compared. The VITS-VC models were not evaluated because the CERs of these models were higher than those of JETS-VC. To evaluate naturalness, each subject evaluated 100 samples and rated the naturalness of each sample on a five-point scale. To evaluate speaker similarity, each subject evaluated 80 pairs comprising the target and converted samples to judge whether the two samples were produced by the same speaker with confidence (sure or not sure). Twenty Japanese adult native speakers without hearing loss participated using headphones.

4.2. Results of experiments

The results of the objective evaluations are shown in Table 1. The RTF results show that models with a HiFi-GAN generator achieved real-time inference, whereas CFS2+PWG and CFS2′+PWG did not. Because the decoder in CFS2′+HiFi-GAN had an RF of 3, this model achieved slightly faster synthesis than VITS-VC and JETS-VC. Comparing the results of the $\log f_{\mathrm{o}}$ RMSE for CFS2+PWG and CFS2′+PWG reveals that the target $\log f_{\mathrm{o}}$ was successfully predicted only from source mel-spectrograms by the modified variance adaptor. Additionally, VITS-VC achieved a lower $\log f_{\mathrm{o}}$ RMSE than CFS2+PWG, even though VITS-VC has no variance adaptor. Similarly to the results of JETS and VITS for TTS [3], JETS-VC achieved a lower CER than VITS-VC, and JETS-VC ($r_{\mathrm{e}} = 2$) achieved the lowest CER. The results of the MOS tests and paired comparison tests are shown in Figure 3. For male to female conversion, JETS-VC ($r_{\mathrm{e}} = 2$) achieved the highest MOS value and similarity. For female to male conversion, JETS-VC ($r_{\mathrm{e}} = 3$) achieved the highest MOS value and similarity comparable to that of CFS2′+HiFi-GAN (ft). The averaged results show that JETS-VC ($r_{\mathrm{e}} = 3$) achieved the highest similarity and significantly higher naturalness than the cascade models.

In summary, E2E-S2S-VC could solve the problems (identified in Section 1) in the conventional cascade S2S VC, and the proposed JETS-VC achieved higher conversion quality than the cascade models. Future work includes improving inference speed by introducing faster generator models [31, 39], investigating efficient training frameworks that require only a small amount of parallel data [11, 17] for practical applications, introducing neural-network-based data-driven trainable feature extraction (following [5]) to improve the quality of conversion, and integrating fundamental frequency and speech rate control [40] to improve its controllability.

5. Conclusion

This paper proposed E2E non-AR S2S VC models by extending E2E TTS models. In the proposed VITS-VC and JETS-VC, the input text sequences for TTS are replaced with the source speaker's acoustic feature sequences with an RF only for the encoder, and E2E models including HiFi-GAN waveform synthesizers are trained with MAS without external aligners. The results of the experiments demonstrate that the proposed JETS-VC outperformed the cascade non-AR S2S VC models.

6. References

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, Apr. 2017.

[2] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, July 2021, pp. 5530–5540.

[3] D. Lim, S. Jung, and E. Kim, "JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech," in *Proc. Interspeech*, Sept. 2022, pp. 21–25.

[4] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 132–157, 2021.

[5] B. Nguyen and F. Cardinaux, "NVC-Net: End-to-end adversarial voice conversion," in *Proc. ICASSP*, May 2022, pp. 7012–7016.

[6] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 631–644, Mar. 2019.

[7] H. Kameoka, K. Tanaka, D. Kwaśny, T. Kaneko, and N. Hojo, "ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1849–1863, 2020.

[8] Z. Yang, X. Jing, A. Triantafyllopoulos, M. Song, I. Aslan, and B. W. Schuller, "An overview & analysis of sequence-to-sequence emotional voice conversion," in *Proc. Interspeech*, Sept. 2022, pp. 4915–4919.

[9] J. Shi, S. Guo, N. Huo, Y. Zhang, and Q. Jin, "Sequence-to-sequence singing voice synthesis with perceptual entropy loss," in *Proc. ICASSP*, June 2021, pp. 76–80.

[10] W.-C. Huang, B. M. Halpern, L. P. Violeta, O. Scharenborg, and T. Toda, "Towards identity preserving normal to dysarthric voice conversion," in *Proc. ICASSP*, May 2022, pp. 6672–6676.

[11] D. Ma, L. Violeta, K. Kobayashi, and T. Toda, "Two-stage training method for Japanese electrolaryngeal speech enhancement based on sequence-to-sequence voice conversion," in *Proc. SLT*, Jan. 2023, pp. 949–954.

[12] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 540–552, 2020.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, May 2015.

[14] R. Liu, X. Chen, and X. Wen, "Voice conversion with transformer network," in *Proc. ICASSP*, May 2020, pp. 7759–7763.

[15] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proc. Interspeech*, Oct. 2020, pp. 4676–4680.

[16] H. Kameoka, W.-C. Huang, K. Tanaka, T. Kaneko, N. Hojo, and T. Toda, "Many-to-many voice transformer network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 656–670, 2021.

[17] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Pretraining techniques for sequence-to-sequence voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 745–755, 2021.

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

[19] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, "Recent developments on ESPnet toolkit boosted by Conformer," in *Proc. ICASSP*, June 2021, pp. 5874–5878.

[20] 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*, Oct. 2020, pp. 5036–5040.

[21] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, Dec. 2019, pp. 3165–3174.

[22] 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 *Proc. ICASSP*, May 2020, pp. 6199–6203.

[23] T. Hayashi, K. Kobayashi, and T. Toda, "Investigation of streaming non-autoregressive sequence-to-sequence voice conversion," in *Proc. ICASSP*, May 2022, pp. 6802–6806.

[24] K. Tanaka, H. Kameoka, T. Kaneko, and S. Seki, "Distilling sequence-to-sequence voice conversion models for streaming conversion applications," in *Proc. SLT*, Jan. 2023, pp. 1022–1028.

[25] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, Dec. 2020, pp. 17 022–17 033.

[26] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *Proc. NeurIPS*, Dec. 2020, pp. 8067–8077.

[27] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, May 2021.

[28] T. Hayashi, R. Yamamoto, T. Yoshimura, P. Wu, J. Shi, T. Saeki, Y. Ju, Y. Yasuda, S. Takamichi, and S. Watanabe, "ESPnet2-TTS: Extending the edge of TTS research," *arXiv:2110.07840*, 2021.

[29] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Proc. SSW5*, June 2004, pp. 223–224.

[30] T. Okamoto, Y. Shiga, and H. Kawai, "Hi-Fi-CAPTAIN: High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT," https://ast-astrec.nict.go.jp/en/release/hi-fi-captain/, 2023.

[31] T. Okamoto, T. Toda, and H. Kawai, "Multi-stream HiFi-GAN with data-driven waveform decomposition," in *Proc. ASRU*, Dec. 2021, pp. 610–617.

[32] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, Apr. 2014.

[33] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. ICML*, July 2015, pp. 1530–1538.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Dec. 2014, pp. 2672–2680.

[35] R. Badlani, A. Łańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, "One TTS alignment to rule them all," in *Proc. ICASSP*, May 2022, pp. 6092–6096.

[36] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *Proc. ICLR*, May 2021.

[37] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE trans. Inf. Syst.*, vol. E99-D, no. 7, pp. 1877–1884, July 2016.

[38] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. SSPR*, Apr. 2003, pp. 7–12.

[39] M. Kawamura, Y. Shirahata, R. Yamamoto, and K. Tachibana, "Lightweight and high-fidelity end-to-end text-to-speech with multi-band generation and inverse short-time Fourier transform," in *Proc. ICASSP*, June 2023.

[40] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, and H. Kawai, "Harmonic-Net: Fundamental frequency and speech rate controllable fast neural vocoder," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1902–1915, 2023.