CONVNEXT-TTS AND CONVNEXT-VC: CONVNEXT-BASED FAST END-TO-END SEQUENCE-TO-SEQUENCE TEXT-TO-SPEECH AND VOICE CONVERSION

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

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

ABSTRACT

End-to-end (E2E) sequence-to-sequence (S2S) neural text-to-speech (TTS) models and E2E-S2S neural voice conversion (VC) models can achieve high-quality speech synthesis with a single neural network. To further improve the synthesis quality of E2E-S2S TTS and VC models and increase their inference speed, we propose a Transformer-free ConvNeXt-based encoder and decoder. Additionally, to further increase the inference speed, we propose ConvNeXt-TTS and ConvNeXt-VC, which include the WaveNeXt neural vocoder. This is also constructed from ConvNeXt blocks and can achieve much faster synthesis than HiFi-GAN. The results of experiments using the Hi-Fi-CAPTAIN corpus for the E2E-S2S-TTS and E2E-S2S-VC conditions demonstrate that the proposed ConvNeXt-based encoder and decoder can perform inference three times faster than a Transformer-based encoder and decoder while improving the synthesis quality. In particular, ConvNeXt-TTS and ConvNeXt-VC can achieve very fast E2E-S2S-TTS and E2E-S2S-VC with a real-time factor of 0.05 using a single-core CPU.

Index Terms— ConvNeXt, JETS, text-to-speech, voice conversion, WaveNeXt

1. INTRODUCTION

For sequence-to-sequence (S2S) conversion, such as machine translation and automatic speech recognition (ASR), Transformer-based models [1] can achieve high conversion quality. Autoregressive Transformer-based models have also been proposed [2, 3] for S2S text-to-speech (TTS) and S2S voice conversion (VC)[1]. In S2S-TTS and S2S-VC, non-autoregressive models using feedforward Transformer-based encoders and decoders with self-attention can achieve faster high-fidelity synthesis by using external aligners [5,7,8]. Additionally, end-to-end (E2E)-S2S-TTS models [9–12] have been designed that can synthesize speech waveforms directly from input text or phoneme sequences with a single neural network. In particular, JETS [10] can achieve fast high-fidelity E2E-S2S-TTS, outperforming conventional cascade models [13] and VITS, which is another E2E-S2S-TTS model with a Transformer-based encoder [9]. For E2E-S2S-VC, JETS-VC [6] outperforms the cascade model [5]. Transformer-based encoders and decoders with self-attention are the de facto standard for S2S-TTS and S2S-VC.

Transformer is also effective for various tasks in computer vision, and many Transformer-based models have been proposed, such as Swin Transformer [16]. To improve the recognition accuracy and inference speed without using a Transformer structure, while maintaining the same behavior as Swin Tranformer, ConvNeXt [14] has been proposed. ConvNeXt, which outperforms Swin Transformer, introduces layer normalization layers, depthwise convolution layers [17], pointwise convolution layers [18], and a Gaussian error linear unit (GELU) into ResNet [19] (Fig. 1). Vocos, which includes multiple ConvNeXt blocks and a short-time Fourier transform (STFT)-based upsampling layer, has been designed [15] for neural vocoding. To improve the synthesis quality while maintaining the inference speed, WaveNeXt has been proposed, in which the STFT layer is replaced with a trainable linear layer [20]. WaveNeXt can perform fast inference with a real-time factor (RTF) of 0.04 using a single-core CPU and can achieve higher synthesis quality than Vocos [15]. However, these tasks are framewise-based, not S2S conversion. ConvNeXt has never been used in S2S conversion tasks in which the lengths of the input and output sequences are different.

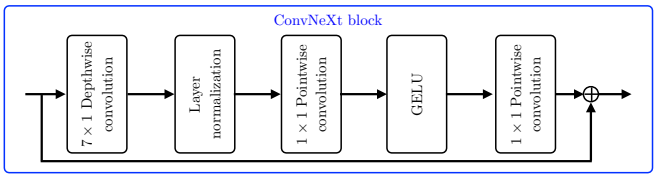

Fig. 1. Network architecture of ConvNeXt [14] with 1D convolutions used in Vocos [15]. GELU abbreviates Gaussian error linear unit.

This paper proposes Transformer-free ConvNeXt-based E2E-S2S TTS and VC models. This is the first time that the ConvNeXt architecture has been used in encoder–decoder acoustic models (AMs) for both TTS and VC frameworks as S2S conversion tasks. Additionally, to further increase the inference speed, ConvNeXt-TTS and ConvNeXt-VC are proposed, in which a ConvNeXt-based encoder and decoder are combined with a WaveNeXt neural vocoder. The results of experiments using the Hi-Fi-CAPTAIN corpus [21] for the E2E-S2S-TTS and E2E-S2S-VC conditions demonstrate that the proposed ConvNeXt-based encoder and decoder can perform inference three times faster than a Transformer-based encoder and decoder, while improving the synthesis quality; this is a consequence of the sophisticated structure of ConvNeXt. In particular, ConvNeXt-TTS and ConvNeXt-VC can achieve very fast E2E-S2S-TTS and E2E-S2S-VC with an RTF of 0.05 using a single-core CPU. Therefore, the proposed ConvNeXt-based encoder–decoder AM framework can be replaced with the conventional Transformer-based framework. To ensure the reproducibility of this study, some of the speech samples and the PyTorch source code used in the experiments are available[2].

[1]In contrast to framewise VC (e.g., [4]), S2S-VC can convert the duration and prosody between the source and target speech [5, 6].

[2]https://ast-astrec.nict.go.jp/demo_samples/convnext-tts_vc/index.html

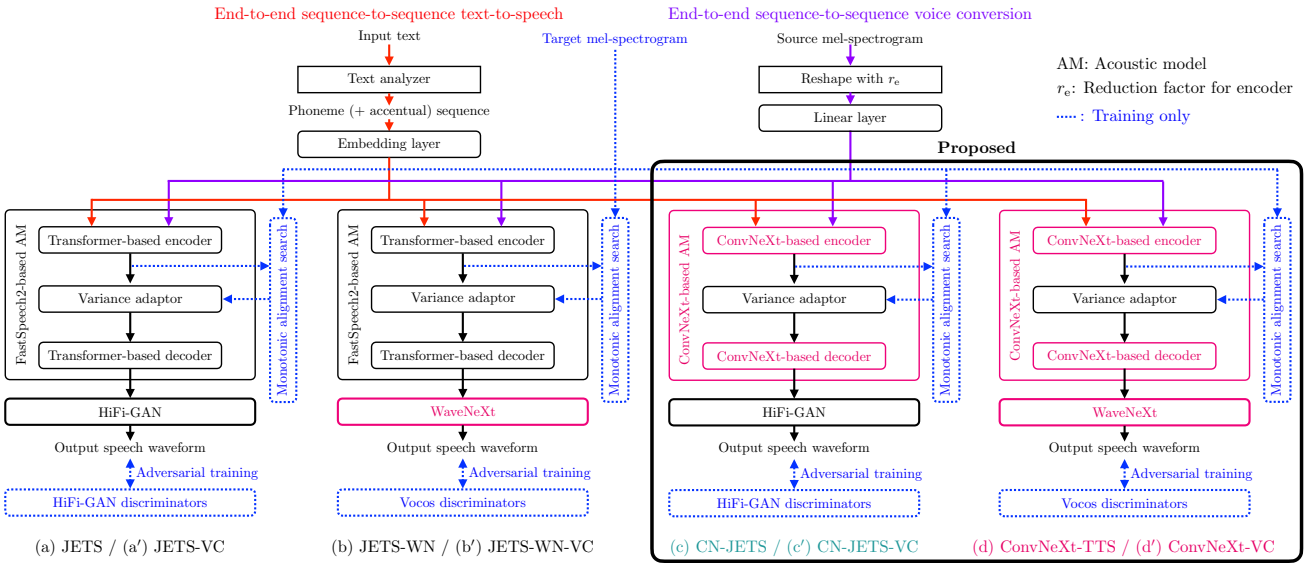

Fig. 2. Network architectures of end-to-end sequence-to-sequence text-to-speech and voice conversion models. (a) JETS [10] and (a′) JETS-VC [6]. (b) JETS-WN [20] and (b′) JETS-WN-VC. (c) CN-JETS and (c′) CN-JETS-VC. (d) WaveNeXt-TTS and (d′) WaveNeXt-VC. The variance adaptor predicts duration, energy, and fundamental frequency in the same manner as FastSpeech 2 [8].

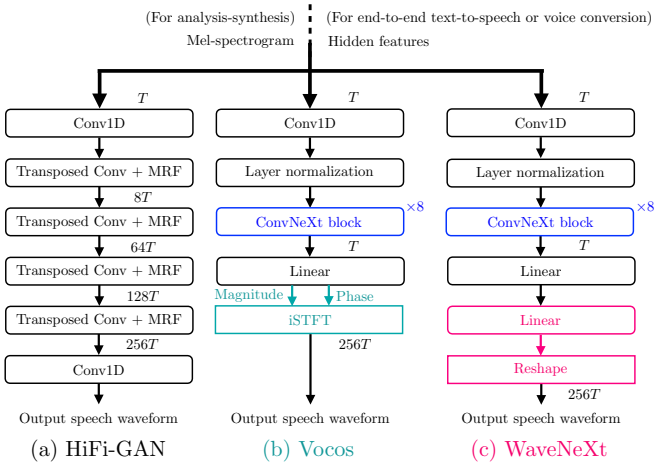

Fig. 3. Network architectures of neural vocoder generators with a sampling frequency of 24 kHz and a shift length of 256 samples. (a) HiFi-GAN [22]. (b) Vocos [15]. (c) WaveNeXt [20]. T is the number of frames. MRF abbreviates multi-receptive field fusion [22].

2. CONVENTIONAL METHODS

2.1. JETS and JETS-VC (Transformer + HiFi-GAN)

JETS and VITS [10] are both E2E-S2S-TTS models but JETS is simpler and achieves higher synthesis quality. JETS performs joint training of a FastSpeech-2-based AM [8], which includes a Transformer-based encoder and decoder with self-attention and a HiFi-GAN-based neural vocoder [22] [Fig. 3(a)] with neither intermediate mel-spectrograms nor external aligners. In contrast, FastSpeech-based TTS models [7, 8, 13] require external aligners. JETS includes an alignment training framework proposed in [23] with monotonic alignment search (MAS) [24]. The alignment between the hidden features (converted from the input text sequences) and the target mel-spectrogram sequences is obtained gradually during training. Similarly to FastSpeech 2, in addition to the duration, the energy and fundamental frequency (f_o) are also predicted in the variance adaptor for higher-quality synthesis. JETS uses the same discriminators as HiFi-GAN [22] [Fig. 2(a)].

JETS-VC for E2E-S2S-VC [6] [Fig. 2(a′)] is constructed as shown in Fig. 2. The input text sequence, text encoder, and embedding layer are replaced with the source mel-spectrogram sequence, reshaping block with reduction factor for encoder r_e [6], and linear layer. JETS-VC outperforms the conventional cascade model [5].

Although the inference speed of HiFi-GAN is high, its RTF is greater than 0.5 on a single CPU [25]. For example, when the duration of a waveform is 10 s, the inference time is greater than 5 s. To increase the inference speed of HiFi-GAN while maintaining the synthesis quality, some faster models have been designed [25–27], and HiFi-GAN in VITS and JETS can be replaced by these faster models [25].

2.2. WaveNeXt neural vocoder

As an alternative to transposed convolution-based upsampling models [22, 25–27], a generative adversarial network (GAN)-based fast neural vocoder, Vocos, has been proposed [15] [Fig. 3(b)]. In Vocos, high-resolution STFT spectra are predicted from input mel-spectrograms by multiple ConvNeXt blocks [14] without upsampling, and the predicted high-resolution STFT spectra are directly converted to speech waveforms by a final iSTFT layer. By introducing the sophisticated ConvNeXt structure, Vocos can achieve ten times faster inference on a CPU [15]. However, the synthesis quality of Vocos is inferior to that of HiFi-GAN-based models [20].

To improve the synthesis quality while maintaining the inference speed, WaveNeXt has been proposed, in which the STFT-based upsampling layer is replaced with a trainable linear layer and reshaping block [20] [Fig. 3(c)]. This is similar to FC-HiFi-GAN [25], in which the STFT-based upsampling layer in iSTFTNet [27] is

replaced with a trainable linear layer and reshaping block [25]. WaveNeXt uses the same discriminators as Vocos [15]. As a consequence of the trainable linear-layer-based upsampling, WaveNeXt can perform fast inference with an RTF of 0.04 using a single-core CPU and achieve a higher synthesis quality than Vocos and HiFi-GAN V2 with an initial channel of 128 [20].

2.3. JETS-WN and JETS-WN-VC (Transformer + WaveNeXt)

Vocos and WaveNeXt can be jointly trained with a FastSpeech 2-based AM, in the same manner as JETS. Fast E2E-S2S-TTS models, JETS-Vocos and JETS-WN [Fig. 2(b)], have been implemented [20] using this AM and the discriminators used in Vocos [15]. The loss functions of JETS-WN for the generator and discriminators are defined as, $\mathcal{L}_{\text{G,JETS-WN}} = \mathcal{L}_{\text{G}} + w_{\text{var}}\ell_{\text{var}} + w_{\text{align}}\ell_{\text{align}}$ and $\mathcal{L}_{\text{D,JETS-WN}} = \mathcal{L}_{\text{D}}$, where $\mathcal{L}_{\text{G}}$ and $\mathcal{L}_{\text{D}}$ are the generator and discriminator loss functions for Vocos [15], ℓ_{var} and ℓ_{align} are the variance loss and alignment loss used in JETS [10], and w_{var} and w_{align} are the weighting coefficients for ℓ_{var} and ℓ_{align}, respectively [20]. Similarly to WaveNeXt, JETS-WN can achieve higher synthesis quality than JETS-Vocos [20]. JETS-WN-VC [Fig. 2(b′)] can also be implemented by replacing HiFi-GAN in JETS-VC with WaveNeXt.

3. PROPOSED METHODS

3.1. CN-JETS and CN-JETS-VC (ConvNeXt + HiFi-GAN)

As described in Section 1, in computer vision, ConvNeXt [14] was designed to improve the accuracy and inference speed of image recognition without a Transformer structure, while maintaining the same behavior as Swin Transformer. ConvNeXt is constructed from layer normalization layers, depthwise convolution layers, pointwise convolution layers, and a GELU (Fig. 1). The depthwise convolution corresponds to the weighted sum in self-attention of Transformer. As a consequence of its sophisticated modifications, ConvNeXt can achieve faster inference and higher accuracy in image recognition than Swin Transformer. Additionally, ConvNeXt is used in two fast neural vocoders, Vocos [15] and WaveNeXt [20], as explained in Section 2.2.

Following the success of ConvNeXt in computer vision and neural vocoding, we propose two ConvNeXt-based E2E-S2S TTS and VC models, CN-JETS [Fig. 2(c)] and CN-JETS-VC [Fig. 2(c′)]. The proposed models use a ConvNeXt-based encoder and decoder in JETS and JETS-VC instead of a Transformer-based encoder and decoder. Specifically, they use the ConvNeXt blocks with one-dimensional (1D) convolutions that are used in Vocos and WaveNeXt (Fig. 1). Similarly to JETS and JETS-VC, MAS is used for alignment training and the discriminators used in HiFi-GAN are used [Fig. 2(c) and (c′)]. Moreover, in addition to duration, energy and f_{o} are also predicted in the variance adaptor. Because of the sophisticated structure of ConvNeXt, the proposed ConvNeXt-based encoder and decoder are expected to achieve faster inference and higher synthesis quality than those based on FastSpeech 2.

3.2. ConvNeXt-TTS and ConvNeXt-VC (ConvNeXt + WaveNeXt)

To further increase the inference speed, ConvNeXt-TTS [Fig. 2(d)] and ConvNeXt-VC [Fig. 2(d′)] are additionally proposed, in which the WaveNeXt neural vocoder is introduced into CN-JETS and CN-JETS-VC instead of HiFi-GAN. Similarly to JETS-WN and JETS-WN-VC, the discriminators loss functions and used in Vocos [15] are also used. Because ConvNeXt-TTS and ConvNeXt-VC are constructed from a ConvNeXt-based encoder and decoder and a WaveNeXt-based neural vocoder, they are expected to perform much faster E2E-S2S-TTS and E2E-S2S-VC and achieve a higher synthesis quality than JETS-WN and JETS-WN-VC.

4. EXPERIMENTS

To evaluate the proposed ConvNeXt-based models, experiments were conducted for both the E2E-S2S-TTS and E2E-S2S-VC conditions with a sampling frequency of 24 kHz. All the neural network models were implemented by modifying ESPnet2-TTS [13] and trained using an NVIDIA Tesla A100 GPU with 40 GB of memory.

4.1. Experimental conditions

Dataset: The experiments were conducted using the Japanese speech dataset (one female and one male speaker) of the Hi-Fi-CAPTAIN corpus [21]. For the E2E-S2S-TTS condition, 18,655 parallel utterances and 201 non-parallel utterances were used for the training set of female models, and 18,655 parallel utterances and 203 non-parallel utterances were used for the training set of male models. For the E2E-S2S-VC condition, 18,655 parallel utterance pairs were used for the training set of male-to-female conversion models and for that of female-to-male conversion models. For both the E2E-S2S-TTS and E2E-S2S-VC conditions, 100 utterances were used for the validation set and the same number were used for the test set, as specified in [21]. The input acoustic features for MAS and E2E-S2S-VC were 80-dimensional mel-spectrograms bandlimited to 7600 Hz. The STFT length and shift length were 1024 and 256 samples, respectively.

Model setting: In the experiments, all the models were trained and inferred by modifying the JETS-based E2E TTS model implemented in ESPnet2-TTS [13][3], following [20]. The Harvest algorithm [28] was used for f_{o} analysis, following [20]. For E2E-S2S-TTS in Japanese, the G2P function based on pyopenjtalk and enhanced with prosody symbols [29] was used, following [13, 20]. The model configuration of JETS with HiFi-GAN V1 was the default setting[4], except that the sampling frequency was changed from 22,050 Hz to 24 kHz. For the E2E-S2S-VC condition, r_{e} was set to 3. For the proposed ConvNeXt-based encoder and decoder, the ConvNeXt blocks with 1D convolutions implemented in an official implementation of Vocos[5] were used. These ConvNeXt blocks were also used for WaveNeXt. For the proposed ConvNeXt-based encoder and decoder in CN-JETS, CN-JETS-VC, ConvNeXt-TTS, and ConvNeXt-VC, the number of input channels, dimensionality of the intermediate layer, and number of ConvNeXt blocks were 256, 1024, and 4, respectively. Additionally, the stochastic depth [30] was used with a weight of 0.2 only for the ConvNeXt-based encoder and decoder. These parameters were the same as those of the FastSpeech 2 model in JETS. w_{var} and w_{align} in the loss function were both set to 1.0.

Evaluation criteria: Mel-cepstral distortion (MCD), $\log f_{\text{o}}$ root-mean-square error (RMSE), and character error rate (CER) of ASR were used as the objective evaluation criteria, following [6, 10, 13, 20]. The MCD and $\log f_{\text{o}}$ RMSE were calculated by the ESPnet2-TTS toolkit [10, 13]. The CER was calculated by the pretrained Transformer-based ASR model for Japanese used in [13][6]. The RTFs

[3] https://is.gd/vbqxeB
[4] https://is.gd/a5UHnP
[5] https://github.com/charactr-platform/vocos
[6] https://zenodo.org/record/4037458

Table 1. Results of objective evaluations. The values of mel-cepstral distortion (MCD) and log f_o root-mean-square error (RMSE) in the table are the means and standard deviations. CER is the character error rate of automatic speech recognition. RTF is the real-time factor on an AMD EPYC 7542 CPU (1 core) using PyTorch 2.0.1. The RTFs of the FastSpeech-2-based acoustic model, ConvNeXt-based acoustic model, HiFi-GAN-based neural vocoder, and WaveNeXt-based neural vocoder are 0.03, 0.01, 0.80, and 0.04, respectively.

			Female (Japanese)			Male (Japanese)		
Condition	Model (Acoustic model + Neural vocoder)	RTF	MCD [dB]	logf_o RMSE	CER [%]	MCD [dB]	logf_o RMSE	CER [%]
E2E-S2S-TTS	JETS (Transformer + HiFi-GAN) [10]	0.83	5.96 ± 0.63	0.21 ± 0.05	**0.4**	5.09 ± 0.56	$\mathbf{0.19 \pm 0.05}$	0.9
	JETS-WN (Transformer + WaveNeXt) [20]	0.07	5.75 ± 0.57	0.21 ± 0.07	**0.4**	5.01 ± 0.62	$\mathbf{0.19 \pm 0.05}$	0.5
	CN-JETS (ConvNeXt + HiFi-GAN)	0.81	5.76 ± 0.61	0.20 ± 0.07	0.6	4.98 ± 0.58	$\mathbf{0.19 \pm 0.05}$	0.6
	ConvNeXt-TTS (ConvNeXt + WaveNeXt)	**0.05**	$\mathbf{5.67 \pm 0.59}$	$\mathbf{0.20 \pm 0.06}$	**0.4**	$\mathbf{4.87 \pm 0.54}$	0.20 ± 0.06	**0.4**
			Male to Female (Japanese)			Female to Male (Japanese)		
E2E-S2S-VC	JETS-VC (Transformer + HiFi-GAN) [6]	0.83	5.55 ± 0.51	$\mathbf{0.20 \pm 0.06}$	1.2	4.90 ± 0.48	0.18 ± 0.06	3.4
	JETS-WN-VC (Transformer + WaveNeXt)	0.07	5.43 ± 0.50	0.21 ± 0.06	1.1	4.87 ± 0.47	$\mathbf{0.18 \pm 0.05}$	4.9
	CN-JETS-VC (ConvNeXt + HiFi-GAN)	0.81	5.52 ± 0.54	$\mathbf{0.20 \pm 0.06}$	1.0	4.75 ± 0.46	0.19 ± 0.05	1.3
	ConvNeXt-VC (ConvNeXt + WaveNeXt)	**0.05**	$\mathbf{5.40 \pm 0.52}$	0.21 ± 0.07	**0.8**	$\mathbf{4.69 \pm 0.48}$	$\mathbf{0.18 \pm 0.05}$	**0.4**
	Ground truth	N/A	N/A	N/A	0.0	N/A	N/A	0.0

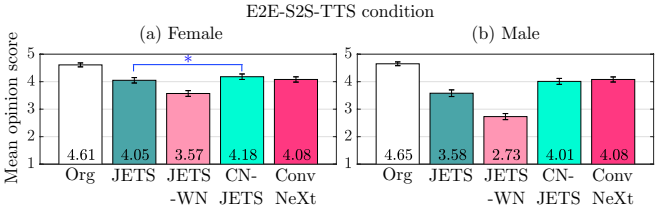

Fig. 4. Results of MOS tests for the E2E-S2S-TTS condition with 23 listening subjects. The confidence level is 95%. "Org" indicates the original samples.

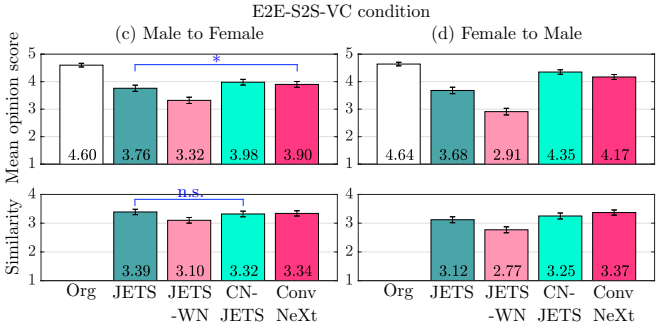

Fig. 5. Results of MOS tests and speaker similarity tests for the E2E-S2S-VC condition with 23 listening subjects. The confidence level is 95%. "Org" indicates the original samples and "n.s." abbreviates non-significant.

of all the models for inference were measured on an AMD EPYC 7542 CPU (1 core). To evaluate the synthesized speech subjectively, mean opinion score (MOS) tests were conducted for both the E2E-S2S-TTS and E2E-S2S-VC conditions. Each subject evaluated 280 samples: 14 utterances $\times$ 5 conditions $\times$ 2 speakers (female and male) $\times$ 2 conditions (E2E-S2S-TTS and E2E-S2S-VC). The naturalness of each sample was rated on a five-point scale: (1) bad, (2) poor, (3) fair, (4) good, and (5) excellent. To evaluate speaker similarity for the E2E-S2S-VC condition, each subject evaluated 112 pairs comprising the target and converted samples: $14 \times 4 \times 2$. The subjects were then asked to rate the speaker similarity of the two samples on a four-point scale: (4) same speaker, absolutely sure, (3) same speaker, not sure, (2) different speaker, not sure, (1) different speaker, absolutely sure [31]. Twenty-three adult Japanese native speakers without hearing loss participated using headphones.

4.2. Results of experiments

Table 1 and Figs. 4 and 5 show the results of the objective and subjective evaluations. The RTFs of the FastSpeech-2-based AM, ConvNeXt-based AM, HiFi-GAN- and WaveNeXt-based neural vocoders were 0.03, 0.01, 0.80, and 0.04, respectively. First, the proposed ConvNeXt-based encoder and decoder performed inference three times faster than the Transformer-based encoder and decoder while improving the synthesis quality. In particular, the proposed ConvNeXt-TTS and ConvNeXt-VC performed very fast E2E-S2S-TTS and E2E-S2S-VC with an RTF of 0.05 using a single-core CPU. In addition, ConvNeXt-TTS and ConvNeXt-VC achieved the highest ASR accuracy and lowest MCD, and significantly improved the synthesis quality and speaker similarity compared with JETS-WN and JETS-WN-VC. The proposed CN-JETS and CN-JETS-VC also significantly improved the synthesis quality and speaker similarity compared with JETS and JETS-VC, with the exception of the speaker similarity score for male-to-female conversion.

In summary, the proposed ConvNeXt-based encoder and decoder could achieve faster and higher-quality synthesis than the Transformer-based encoder and decoder, as expected. Future work includes further improvement of the synthesis quality of ConvNeXt-TTS and ConvNeXt-VC by introducing sophisticated discriminators [32] and duration modeling [33].

5. CONCLUSION

This paper proposed Transformer-free ConvNeXt-based models, CN-JETS and CN-JETS-VC, in which a ConvNeXt-based encoder and decoder were introduced into JETS and JETS-VC instead of a Transformer-based encoder and decoder. To further increase the inference speed, ConvNeXt-TTS and ConvNeXt-VC were additionally proposed, in which WaveNeXt was used instead of HiFi-GAN. The results of the experiments show that the proposed ConvNeXt-based encoder and decoder could perform inference three times faster than a Transformer-based encoder and decoder, while improving the synthesis quality. In particular, ConvNeXt-TTS and ConvNeXt-VC could achieve very fast E2E-S2S-TTS and E2E-S2S-VC with an RTF of 0.05 using a single-core CPU.

6. REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, Dec. 2017, pp. 5998–6008.

[2] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Neural speech synthesis with Transformer network," in *Proc. AAAI*, Jan. 2019, pp. 6706–6713.

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

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

[5] 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.

[6] T. Okamoto, T. Toda, and H. Kawai, "E2E-S2S-VC: End-to-end sequence-to-sequence voice conversion," in *Proc. Interspeech*, Aug. 2023, pp. 2043–2047.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] J. Kong, J. Park, B. Kim, J. Kim, D. Kong, and S. Kim, "VITS2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design," in *Proc. Interspeech*, Aug. 2023, pp. 4374–4378.

[12] B. T. Vecino, A. Gabrys, D. Matwicki, A. Pomirski, T. Iddon, M. Cotescu, and J. Lorenzo-Trueba, "Lightweight end-to-end text-to-speech synthesis for low resource on-device applications," in *Proc. SSW*, Aug. 2023, pp. 225–229.

[13] 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.

[14] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. CVPR*, June 2022, pp. 11976–11986.

[15] H. Siuzdak, "Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis," *arXiv:2306.00814*, 2023.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision Transformer using shifted windows," in *Proc. ICCV*, Oct. 2021, pp. 9992–10002.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR*, July 2017, pp. 1251–1258.

[18] B.-S. Hua, M.-K. Tran, and S.-K. Yeung, "Pointwise convolutional neural networks," in *Proc. CVPR*, June 2018, pp. 984–993.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, June 2016, pp. 770–778.

[20] T. Okamoto, H. Yamashita, Y. Ohtani, T. Toda, and H. Kawai, "WaveNeXt: ConvNeXt-based fast neural vocoder without iSTFT layer," in *Proc. ASRU*, Dec. 2023.

[21] 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.

[22] 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. 17022–17033.

[23] R. Badlani, A. Lańcucki, 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.

[24] 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.

[25] H. Yamashita, T. Okamoto, R. Takashima, Y. Ohtani, T. Takiguchi, T. Toda, and H. Kawai, "Fast neural waveform generation model with fully connected upsampling," *IEICE Tech. Rep.*, vol. 123, no. 88, SP2023-15, pp. 73–78, June 2023, (in Japanese).

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

[27] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform," in *Proc. ICASSP*, May 2022, pp. 6207–6211.

[28] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Proc. Interspeech*, Aug. 2017, pp. 2321–2325.

[29] K. Kurihara, N. Seiyama, and T. Kumano, "Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural TTS," *IEICE trans. Inf. Syst.*, vol. E104-D, no. 2, pp. 302–311, Feb. 2021.

[30] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. ECCV*, Sept. 2016, pp. 646–661.

[31] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *Proc. Interspeech*, Sept. 2016, pp. 1632–1636.

[32] D. S. Dang, T. L. Nguyen, B. T. Ta, T. T. Nguyen, T. N. A. Nguyen, D. L. Le, N. M. Le, and V. H. Do, "LightVoc: An upsampling-free GAN vocoder based on Conformer and inverse short-time Fourier transform," in *Proc. Interspeech*, Aug. 2023, pp. 3043–3047.

[33] Y. Ma, J. He, M. Wu, G. Hu, and H. Fei, "Eden-TTS: A simple and efficient parallel text-to-speech architecture with collaborative duration-alignment learning," in *Proc. Interspeech*, Aug. 2023, pp. 4449–4453.