

ConvNeXt-based fast end-to-end sequence-to-sequence text-to-speech and voice conversion

Takuma Okamoto¹, Yamato Ohtani¹, Tomoki Toda^{2,1}, and Hisashi Kawai¹

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

Demo samples, source code and preprint

Demo samples: Hi-Fi-CAPTAIN corpus for Japanese used in experiments

Source code based on ESPnet2-TTS

- Recipe for Hi-Fi-CAPTAIN corpus used in experiments

https://ast-astrec.nict.go.jp/demo_samples/convnext-tts_vc/

Hi-Fi-CAPTAIN corpus:

High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT

1 female and 1 male (English): 14K utts (parallel: 13K)

1 female and 1 male (Japanese): 19K utts (parallel: 18.5K)

ESPnet2-TTS recipe for JETS-based E2E-TTS

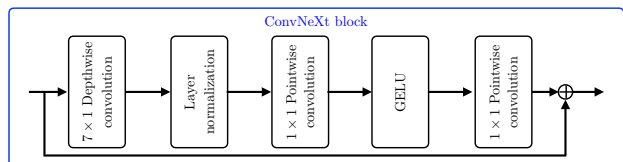
<https://ast-astrec.nict.go.jp/en/release/hi-fi-captain/>

1. Introduction

- Fast and high-fidelity neural text-to-speech (TTS) and voice conversion (VC) models
- End-to-end (E2E) sequence-to-sequence (S2S) TTS model: MS-FC-JETS (Yamashita+ IEEE Access 2024)
 - Realizing high-fidelity and fast synthesis with real-time factor (RTF) of 0.14 using a CPU
- E2E-S2S-VC model: JETS-VC (Okamoto+ Interspeech 2023)
 - Realizing higher quality conversion than cascade models
- Motivation of proposed method
 - Transformer blocks are introduced to encoder and decoder of acoustic models as de facto standard for S2S-TTS and S2S-VC
 - ConvNeXt-based model is proposed and outperforms Swin-Transformer in image recognition
 - ConvNeXt-based very fast neural vocoders: Vocos (Siuzdak ICLR 2024) and WaveNeXt (Okamoto+ ASRU 2023)
- Proposed methods
 - ConvNeXt blocks are introduced to encoder and decoder in acoustic models instead of Transformer blocks
- Results: Proposed models can improve inference speed while keeping synthesis quality

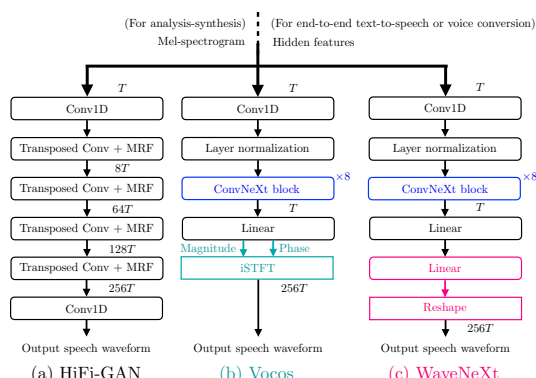
2. Conventional models

- ConvNeXt (Liu+ CVPR 2022)
 - ResNet-based fast and high-fidelity model by introducing essence of Transformer
 - Depthwise convolution corresponds to weighted sum in self-attention of Transformer



- WaveNeXt (Okamoto+ ASRU 2023)

- Replacing iSTFT-based upsampling layer in Vocos (Siuzdak ICLR 2024) with trainable linear layer similar to MS-FC-JETS
- Improving synthesis quality while keeping inference speed

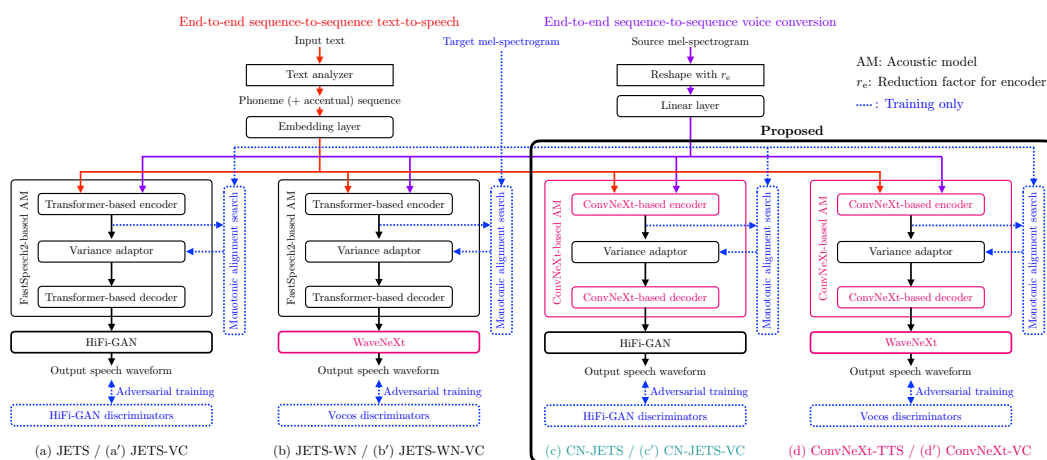


Preprint of WaveNeXt

- (a) JETS (Lim+ Interspeech 2022) and JETS-VC (Okamoto+ IS 2023)
- E2E-S2S-TTS and VC models realized by joint training of FastSpeech 2 and HiFi-GAN with monotonic alignment search

3. Proposed method:

- (c) CS-JETS and CS-JETS-VC
 - ConvNeXt blocks are introduced only to encoder and decoder in acoustic models of JETS and JETS-VS instead of Transformer blocks
- (d) ConvNeXt-TTS and ConvNeXt-VC
 - ConvNeXt blocks are introduced not only to encoder and decoder in acoustic models instead of Transformer blocks but also to speech waveform generative models as WaveNeXt instead of HiFi-GAN
 - All the modules are constructed from ConvNeXt blocks



4. Experiments

- Experimental conditions
 - Dataset: One female and one male in Hi-Fi-CAPTAIN corpus
 - Sampling frequency f_s : 24 kHz
 - Objective evaluation criteria: MCD, $\log f_o$ RMSE, CER and RTF
 - Subjective evaluation criteria ($N=23$): Naturalness and similarity
- Results of experiments

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

