WAVENEXT: CONVNEXT-BASED FAST NEURAL VOCODER WITHOUT ISTFT LAYER

Takuma Okamoto[1]*, Haruki Yamashita*[2,1]*, Yamato Ohtani*[1]*, Tomoki Toda*[3,1]*, and Hisashi Kawai*[1]

[1]National Institute of Information and Communications Technology, Japan
[2]Kobe University, Japan [3]Nagoya University, Japan

ABSTRACT

A recently proposed neural vocoder, Vocos, can perform inference ten times faster than HiFi-GAN because of its use of ConvNeXt layers that can predict high-resolution short-time Fourier transform (STFT) spectra and an inverse STFT layer. To improve synthesis quality while preserving inference speed, this paper proposes an alternative ConvNeXt-based fast neural vocoder, WaveNeXt, in which the inverse STFT layer in Vocos is replaced with a trainable linear layer that can directly predict speech waveform samples without STFT spectra. Additionally, by integrating the JETS-based end-to-end text-to-speech (E2E TTS) framework, E2E TTS models can also be constructed with Vocos and WaveNeXt. Furthermore, full-band models with a sampling frequency of 48 kHz were investigated. The results of experiments for both the analysis–synthesis and E2E TTS conditions demonstrate that the proposed WaveNeXt can achieve higher quality synthesis than Vocos while preserving its inference speed.

***Index Terms*—** ConvNext, end-to-end text-to-speech, linear layer-based upsampling, neural vocoder, Vocos

1. INTRODUCTION

Since the advent of WaveNet [1], many types of real-time neural-network-based generative models for speech waveforms (neural vocoders) have been proposed [2–8]. Although these models can synthesize high-fidelity speech waveforms, a GPU is required to achieve real-time inference. In contrast to these models, MelGAN [9], Multi-band MelGAN [10], and HiFi-GAN [11], which are based on a generative adversarial network (GAN) [12], can achieve real-time inference with a single CPU because of their use of gradual-upsampling-based generators.

In particular, HiFi-GAN can achieve high-quality synthesis and is a de facto standard for neural vocoders. Therefore, HiFi-GAN is widely used both for end-to-end text-to-speech (E2E TTS), to synthesize speech waveforms directly from input text or phoneme sequences with a single neural network [13–17], and for various speech and audio applications. These applications include end-to-end voice conversion [18, 19], singing voice synthesis [20], speech enhancement [21, 22], bandwidth extension [21], neural audio codec [23], automatic spoken language acquisition [24], fundamental frequency controllable neural vocoders [25,26], speech rate conversion [26,27], and sound field reconstruction [28]. Additionally, extended models have also been investigated [29–32]. Despite the high inference speed of HiFi-GAN, its real-time factor (RTF) is about 0.7 on a single CPU [11]. If the duration of a speech waveform is 10 s, the inference time is about 7 s, which is not suitable for real-time applications. Therefore, it is important to further accelerate the inference speed of high-fidelity neural vocoders on a single CPU for practical applications.

To accelerate the inference speed of HiFi-GAN while maintaining its synthesis quality, Multi-stream (MS)-HiFi-GAN [33] and iSTFTNet [34] have been proposed, in which the final $4\times$ upsampling layers of HiFi-GAN are replaced with lightweight fast upsampling layers. Additionally, by efficiently combining these models, MS-iSTFT-HiFi-GAN [35] has been proposed as a VITS-based E2E TTS model: this can perform inference four times faster than VITS (with HiFi-GAN-based waveform synthesizer) [13] while maintaining its synthesis quality. Both iSTFTNet and MS-iSTFT-HiFi-GAN perform upsampling based on the inverse short-time Fourier transform (iSTFT), using fixed weights based on the Fourier basis. More recently, Fully connected (FC)-HiFi-GAN and MS-FC-HiFi-GAN [36] have been proposed, in which this iSTFT-based upsampling is replaced with simple linear-layer-based upsampling using trainable weights without the overlap-add operation. By introducing the trainable upsampling and avoiding the overlap-add operation, FC-HiFi-GAN and MS-FC-HiFi-GAN can achieve slightly faster inference and higher synthesis quality for the E2E TTS condition, compared with iSTFTNet and MS-iSTFT-HiFi-GAN [36].

As an alternative to gradual-upsampling-based generators [11, 33–36], a GAN-based fast neural vocoder, Vocos, has recently been proposed [37]. In Vocos, high-resolution STFT spectra are predicted from input mel-spectrograms by ConvNeXt blocks [38] without upsampling, and the predicted high-resolution STFT spectra are converted directly to speech waveforms by a final iSTFT layer. By introducing the sophisticated ConvNeXt structure, Vocos can perform inference ten times faster on a CPU and achieve higher objective evaluation scores than HiFi-GAN in experiments using the VCTK corpus [39]. Although Vocos can achieve faster inference than HiFi-GAN-based models [11, 35, 36], the synthesis quality of Vocos is lower than that of HiFi-GAN-based models, as demonstrated by the results of experiments reported in Section 5.

To improve the synthesis quality of a ConvNeXt-based fast neural vocoder while preserving its inference speed, this paper proposes an alternative ConvNeXt-based fast neural vocoder, WaveNeXt. WaveNeXt replaces the iSTFT layer in Vocos with a trainable linear layer that can directly predict speech waveform samples without STFT spectra, similarly to FC-HiFi-GAN and MS-FC-HiFi-GAN [36]. Additionally, by integrating the JETS-based E2E TTS framework [16], E2E TTS models can also be constructed with Vocos and WaveNeXt, similarly to HiFi-GAN. Furthermore, full-band models with a sampling frequency f_{s} of 48 kHz were investigated. The results of experiments using the Hi-Fi-CAPTAIN corpus [40] for both the analysis–synthesis and E2E TTS conditions demonstrate that the proposed WaveNeXt can achieve higher quality synthesis than Vocos while preserving its inference speed. Some of the speech samples and the PyTorch [41] source code based on ESPNet2-TTS [42] used in the experiments are available[1].

[1]https://is.gd/duv0DF

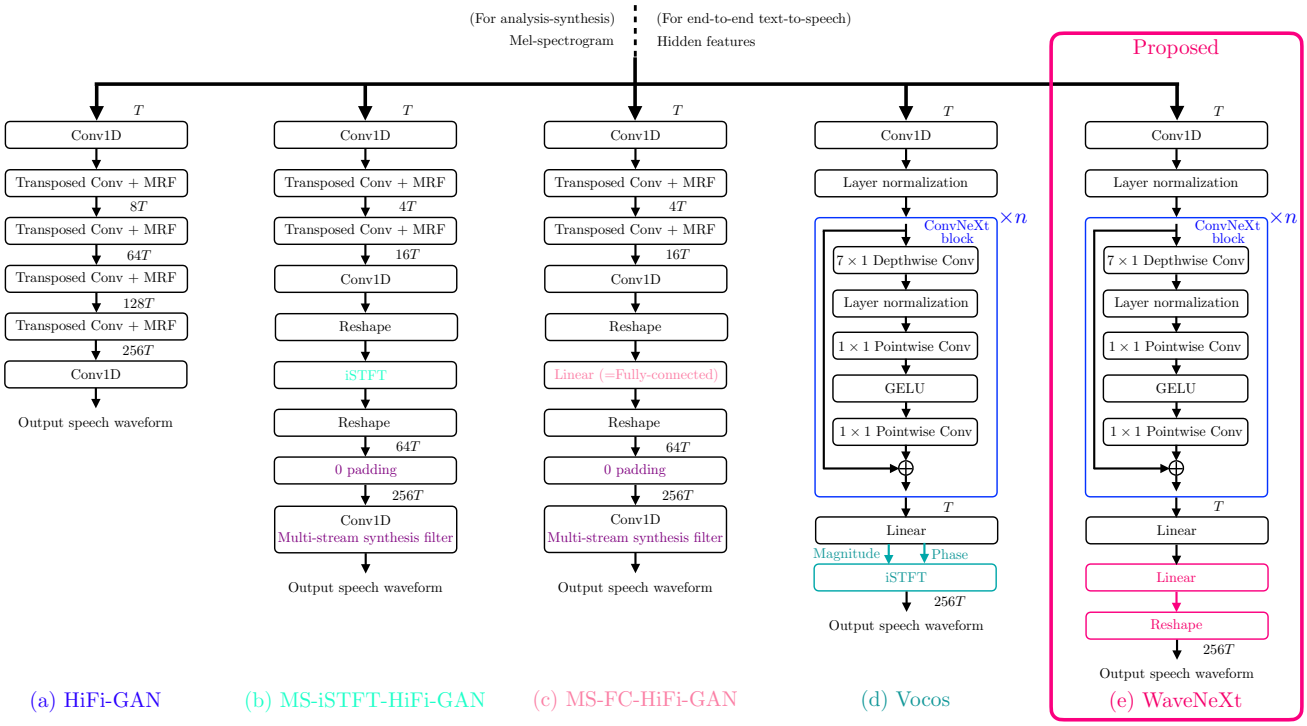


Fig. 1. Network architectures of generative adversarial network (GAN)-based neural vocoders with a sampling frequency of 24 kHz and a shift length of acoustic feature analysis of 256 samples. (a) HiFi-GAN generator [11]. (b) MS-iSTFT-HiFi-GAN generator [35]. (c) MS-FC-HiFi-GAN generator [36]. (d) Vocos generator [37]. (e) Proposed WaveNext generator. T is the number of frames of mel-spectrograms for the analysis–synthesis condition or the number of hidden features for the end-to-end text-to-speech condition. MRF abbreviates multi-receptive field fusion [11]. GELU abbreviates Gaussian error linear unit [43]. n is the number of ConvNeXt blocks.

2. HIFI-GAN-BASED FAST NEURAL VOCODERS

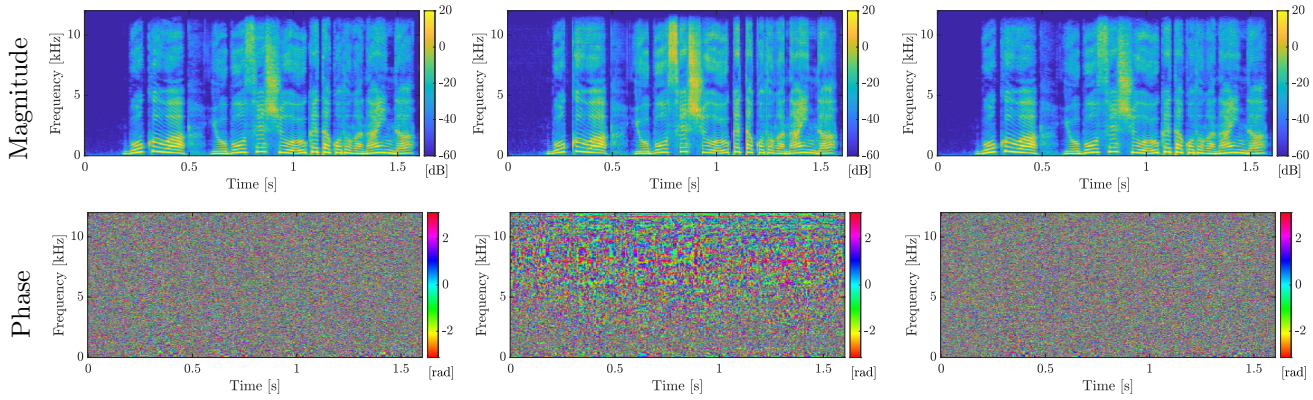
HiFi-GAN [11]: HiFi-GAN is a GAN-based high-fidelity and fast neural vocoder consisting of a generator and two superior discriminators. The generator synthesizes speech waveforms from acoustic features with a shift length of 256 samples, such as mel-spectrograms, by gradually upsampling the input features ($8\times \rightarrow 8\times \rightarrow 2\times \rightarrow 2\times$) using transposed convolutional layers with multi-receptive field fusion blocks, as shown in Fig. 1(a).

MS-HiFi-GAN [33]: Similarly to Multi-band MelGAN [10], HiFi-GAN can easily be accelerated by replacing the final $4\times$ upsampling with a subband synthesis filter [44–47] based on multirate signal processing [48], where the four subband output waveforms are upsampled by zero-padding and then a full-band speech waveform is synthesized. MS-HiFi-GAN replaces the subband synthesis filter, which can be regarded as a convolutional layer with fixed weights without bias, with a trainable convolutional layer without bias. It can be successfully trained by decomposing the target waveform into the four output waveforms in a data-driven manner.

iSTFTNet [34]: Similarly to the subband synthesis filter in Multi-band MelGAN [10], iSTFT can be regarded as an upsampling operation. To accelerate HiFi-GAN, iSTFTNet replaces the last two layers for the final $4\times$ upsampling of HiFi-GAN with iSTFT-based fast upsampling. In iSTFTNet, the amplitude and phase components of the STFT spectra are predicted by a one-dimensional convolutional layer before the iSTFT layer.

MS-iSTFT-HiFi-GAN [35]: By combining a trainable convolutional layer-based upsampling for MS-HiFi-GAN with iSTFT-based upsampling for iSTFTNet, MS-iSTFT-HiFi-GAN has been proposed to further accelerate HiFi-GAN-based neural vocoders. MS-iSTFT-HiFi-GAN is used in the speech waveform synthesizer component for VITS-based E2E TTS. The architecture of the MS-iSTFT-HiFi-GAN generator is depicted in Fig. 1(b). Although MS-iSTFT-HiFi-GAN is twice as fast as MS-HiFi-GAN and iSTFTNet, it can still maintain the same synthesis quality.

FC-HiFi-GAN and MS-FC-HiFi-GAN [36]: With the success of the trainable upsampling in MS-HiFi-GAN [33], FC-HiFi-GAN and MS-FC-HiFi-GAN have been proposed. In these models, the iSTFT layer-based upsampling in iSTFTNet and MS-iSTFT-HiFi-GAN is replaced with a simple linear (fully connected) layer-based trainable fast upsampling without the overlap-add operation. In the linear layer-based upsampling, $N\times$ upsampling is performed simply by a trainable linear layer with output channels of N and by reshaping the output tensor shape from (B, N, T) to (B, 1, NT), where B and T are the batch size and number of frames, respectively. The linear layer-based upsampling is equivalent to sub-pixel convolution (pixel-shuffler) based upsampling [49] with a 1×1 convolutional layer. A trainable linear layer without bias is introduced [33]. By introducing the trainable upsampling and avoiding the overlap-add operation, FC-HiFi-GAN and MS-FC-HiFi-GAN can achieve slightly faster inference and higher quality synthesis for the E2E TTS condition, compared with iSTFTNet and MS-iSTFT-HiFi-GAN [36].



(a) Ground truth (b) Vocos (estimated) (c) Vocos (reanalyzed)

Fig. 2. (a) Magnitude and phase components of the STFT spectrum of a ground-truth female speech waveform. (b) Components estimated by Vocos trained using the female corpus. (c) Components reanalyzed from the speech waveform synthesized by using (b).

3. VOCOS

In contrast to HiFi-GAN-based models, which have gradual upsampling based generators [11, 33–36], an alternative GAN-based fast neural vocoder, Vocos, has recently been proposed [37]. In the Vocos generator, high-resolution magnitude and phase components of the STFT spectra are predicted from input mel-spectrograms by ConvNeXt blocks [38] without upsampling, and the predicted high-resolution STFT spectra are converted directly to speech waveforms by a final iSTFT layer, as shown in Fig. 1(d). ConvNeXt blocks were first proposed for image classification and have outperformed Swin-Transformer [50]. They are constructed from layer normalization layers [51], depthwise convolution layers [52], pointwise convolution layers [53], and Gaussian error linear unit (GELU) activations [43], as shown in Fig. 1(d). By introducing the sophisticated ConvNeXt structure, Vocos can directly predict high-resolution magnitude and phase components of the STFT spectra, while performing inference ten times faster on a CPU and achieving higher objective evaluation scores than HiFi-GAN in experiments using the VCTK corpus [39], even though predicting high-resolution STFT spectra is a challenging problem [37]. In Vocos, to predict high-resolution magnitude and phase components of the STFT spectra by considering the phase wrapping in the desired range $(-\pi, \pi]$, the hidden features output from the linear layer are split into $\mathbf{m}$ and $\mathbf{p}$, and the magnitude and phase components are predicted as $\mathbf{M} = \exp(\mathbf{m})$ and $\boldsymbol{\varphi} = \mathrm{atan2}(\sin(\mathbf{p}), \cos(\mathbf{p}))$, respectively. The complex STFT spectrum is then obtained as $\mathbf{M} \cdot e^{j\boldsymbol{\varphi}}$, where $j = \sqrt{-1}$. An inverse modified discrete cosine transform layer has also been investigated, in addition to the iSTFT layer in Vocos [37]. To train the Vocos generator, the multi-period discriminator (MPD) used in HiFi-GAN [11] and the multi-resolution discriminator (MRD) used in UnivNet [30] are employed. The loss functions of the generator and discriminators are then defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{G}} =& \ell_{\mathrm{G,MPD}} + w_{\mathrm{MRD}}\ell_{\mathrm{G,MRD}} \\ &+ \ell_{\mathrm{FM,MPD}} + w_{\mathrm{MRD}}\ell_{\mathrm{FM,MRD}} + w_{\mathrm{mel}}\ell_{\mathrm{G,mel}} \quad (1)\\ \mathcal{L}_{\mathrm{D}} =& \ell_{\mathrm{D,MPD}} + w_{\mathrm{MRD}}\ell_{\mathrm{D,MRD}}, \quad (2)\end{aligned}$$

where $\ell_{\mathrm{G,MPD}}$, $\ell_{\mathrm{G,MRD}}$, $\ell_{\mathrm{D,MPD}}$, and $\ell_{\mathrm{D,MRD}}$ are the adversarial loss functions of the generator and discriminators for MPD and MRD, $\ell_{\mathrm{FM,MPD}}$ and $\ell_{\mathrm{FM,MRD}}$ are the feature matching loss functions for MPD and MRD, $\ell_{\mathrm{G,mel}}$ is the mel-spectrogram $L1$ loss between the ground-truth and synthesized speech waveforms, and w_{MRD} and w_{mel} are the weighting coefficients of the loss functions for MRD and mel-spectrogram $L1$ loss, respectively. The hinge loss formulation [54] is used in Vocos [37].

4. PROPOSED METHOD: WAVENEXT

4.1. Is iSTFT layer-based upsampling really necessary for GAN-based training in the time domain?

As explained in Section 1, although Vocos can achieve faster inference than HiFi-GAN-based models [11,35,36], the synthesis quality of Vocos is lower than that of HiFi-GAN-based models, as demonstrated by the results of the experiments reported in Section 5. To explain the behavior of Vocos, Fig. 2 shows the magnitude and phase components of the STFT spectrum of a ground-truth female speech waveform, together with those estimated by Vocos trained using the female corpus and those reanalyzed from the speech waveform synthesized by using the estimated STFT spectrum. The estimated magnitude component (Fig. 2(b)) is slightly degraded, compared with that of the ground truth (Fig. 2(a)), and the estimated phase component (Fig. 2(b)) differs from that of the ground truth (Fig. 2(a)). These results indicate that Vocos cannot perfectly predict the magnitude and phase components of the STFT spectra. However, the reanalyzed magnitude and phase components (Fig. 2(c)) are indistinguishable from those of the ground truth (Fig. 2(a)). This is because the estimated magnitude and phase components, which differ from those of the ground truth, can still synthesize high-fidelity speech waveforms because of the redundancy of the overlap-add operation and GAN-based training in the time domain. Conversely, Vocos is trained to estimate STFT spectra for synthesizing high-quality speech waveforms using the overlap-add operation, and GAN-based training in the time domain has no restriction in the STFT domain. Therefore, direct estimation of speech waveform samples in the time domain is more suitable for GAN-based training in the time domain than the indirect estimation of STFT spectra used in Vocos.

4.2. WaveNeXt

To predict speech waveform samples directly, without the use of iSTFT layer-based upsampling, WaveNeXt, an alternative ConvNeXt-based fast neural vocoder, is proposed. In WaveNeXt, the iSTFT

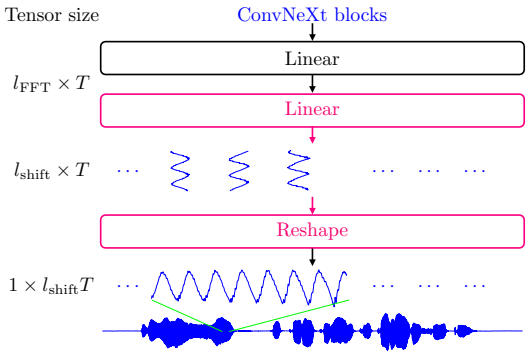


Fig. 3. Final linear layer and reshaping component in WaveNeXt generator. T is the number of frames and l_{FFT} and l_{shift} are the FFT length and shift length of acoustic feature analysis, respectively.

layer in Vocos is replaced with a trainable linear layer without bias, similarly to FC-HiFi-GAN and MS-FC-HiFi-GAN [36]. The proposed WaveNeXt generator is depicted in Fig. 1(e). To adjust the iSTFT layer in Vocos, the input and output channels of the final linear layer are set to l_{FFT} and l_{shift}, which are the FFT length and shift length of acoustic feature analysis, respectively. As shown in Fig. 3, the final linear layer in the proposed WaveNeXt generator directly predicts speech waveform samples with a tensor size of $l_{\mathrm{shift}} \times T$ and the reshaping component concatenates all the predicted speech waveform pieces and finally synthesizes a speech waveform with a length of $1 \times l_{\mathrm{shift}} T$. The proposed WaveNeXt uses the same discriminators and loss functions as those used in Vocos. Compared with FC-HiFi-GAN and MS-FC-HiFi-GAN, which use linear layers for relatively small $4\times$ upsampling, the large and direct $256\times$ upsampling using the final linear layer in the proposed WaveNeXt is a challenging problem. However, because of the sophisticated ConvNeXt layers and final trainable linear layer for direct speech waveform sample prediction, the proposed WaveNeXt is expected to achieve higher quality synthesis than Vocos with iSTFT layer-based upsampling while maintaining the same inference speed.

4.3. JETS-based end-to-end text-to-speech models with Vocos and WaveNeXt

Although HiFi-GAN-based models have been used for E2E TTS [13–17], and high-fidelity and fast TTS can be performed [35, 36], Vocos has only been investigated for the analysis–synthesis condition with mel-spectrogram input [37]. To perform E2E TTS much faster than HiFi-GAN-based models, E2E TTS models using Vocos and WaveNeXt are additionally proposed in this subsection.

Compared with VITS [13], JETS is a simpler E2E TTS model that achieves higher synthesis quality than VITS [16]. Therefore, the JETS-based E2E TTS framework is integrated into Vocos and WaveNeXt neural vocoders. JETS is implemented by joint training of a FastSpeech 2 [55]-based acoustic model and a HiFi-GAN-based neural vocoder using the same discriminators as those used for HiFi-GAN, with neither intermediate mel-spectrograms nor external aligners; however, FastSpeech 2 [55] requires an external aligner, such as Montreal Forced Aligner [56]. In JETS, an alignment training framework proposed in [57] with monotonic alignment search (MAS) [58] is used, and the alignment between the hidden features converted from the input text sequences and the target mel-spectrogram sequences is gradually obtained in the training, in the same manner as VITS. The JETS-based generator for E2E TTS

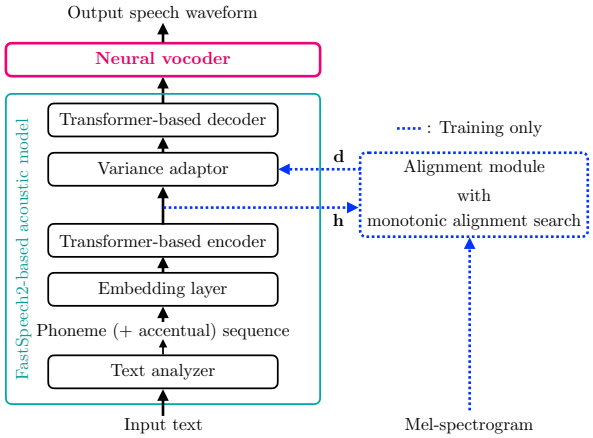


Fig. 4. JETS-based generator for end-to-end text-to-speech models. **h** and **d** are the hidden feature and phoneme duration sequences.

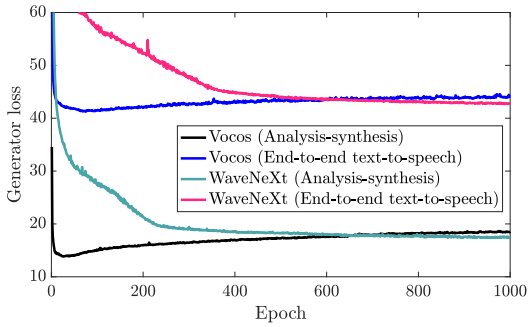


Fig. 5. Generator loss values of Vocos and the proposed WaveNeXt using the female corpus in the training for both the analysis–synthesis and end-to-end text-to-speech conditions.

models is shown in Fig. 4. Although the loss functions for HiFi-GAN-based models [36] are the same as those of JETS [16], the loss functions of ConvNeXt-based models with Vocos and WaveNeXt for the generator and discriminators are defined as follows:

$$\mathcal{L}_{\mathrm{G,E2ETTS}} = \mathcal{L}_{\mathrm{G}} + w_{\mathrm{var}}\ell_{\mathrm{var}} + w_{\mathrm{align}}\ell_{\mathrm{align}} \quad (3)$$

$$\mathcal{L}_{\mathrm{D,E2ETTS}} = \mathcal{L}_{\mathrm{D}}, \quad (4)$$

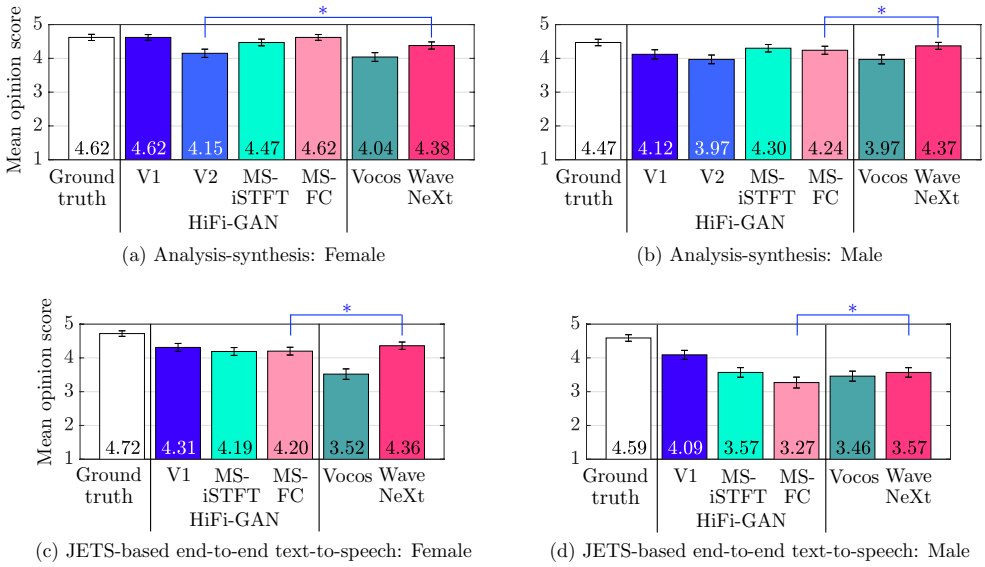
where ℓ_{var} and ℓ_{align} are the variance loss and alignment loss used in JETS [16], respectively, and w_{var} and w_{align} are the weighting coefficients for ℓ_{var} and ℓ_{align}, respectively.

4.4. Full-band end-to-end text-to-speech models with Vocos and WaveNeXt

To further improve the synthesis quality, full-band E2E TTS models with $f_{\mathrm{s}} = 48$ kHz, covering the human auditory frequency range, were additionally investigated. In HiFi-GAN-based models, additional $2\times$ upsampling layers were introduced, as in [36, 42]. In contrast, ConvNeXt-based models can easily perform full-band synthesis by simply changing the FFT and shift lengths of the STFT calculation. In contrast to HiFi-GAN-based models, the inference speed of ConvNeXt-based models is almost the same as that with a sampling frequency of 48 kHz. Therefore, very fast full-band E2E TTS can be achieved by Vocos and WaveNeXt neural vocoders.

Table 1. Results of objective evaluations for the analysis–synthesis and JETS-based E2E TTS conditions using the Hi-Fi-CAPTAIN corpus. The values in the mel-cepstral distortion (MCD) and $\log f_{\mathrm{o}}$ root-mean-square error (RMSE) columns represent the means and standard deviations. f_{s}, CER, and RTF are the sampling frequency, character error rate of automatic speech recognition, and real-time factor on an AMD EPYC 7542 CPU (1 core) using PyTorch 1.13.1. The RTF of FastSpeech 2-based acoustic model in JETS-based E2E TTS is 0.05.

			Female ($fs = 24$ kHz)			Male ($fs = 24$ kHz)		
Condition	Neural vocoder	RTF	MCD [dB]	$\log f_{\mathrm{o}}$ RMSE	CER [%]	MCD [dB]	$\log f_{\mathrm{o}}$ RMSE	CER [%]
Analysis-synthesis	HiFi-GAN V1 [11]	0.92	2.21 ± 0.09	0.16 ± 0.06	1.7	2.00 ± 0.10	$\mathbf{0.12 \pm 0.06}$	2.0
	HiFi-GAN V2 [11]	**0.10**	2.65 ± 0.10	0.18 ± 0.08	2.1	2.65 ± 0.09	0.13 ± 0.04	2.2
	MS-iSTFT-HiFi-GAN [35]	0.19	2.07 ± 0.10	$\mathbf{0.15 \pm 0.07}$	1.7	2.01 ± 0.09	0.13 ± 0.05	**1.7**
	MS-FC-HiFi-GAN [36]	0.18	$\mathbf{2.02 \pm 0.09}$	0.16 ± 0.08	1.9	$\mathbf{1.90 \pm 0.08}$	0.14 ± 0.06	2.0
	Vocos [37]	**0.10**	2.74 ± 0.11	0.16 ± 0.07	2.0	3.05 ± 1.16	0.14 ± 0.06	2.4
	WaveNeXt	**0.10**	2.86 ± 0.12	0.17 ± 0.07	**1.6**	4.32 ± 0.33	0.13 ± 0.05	2.0
JETS-based E2E TTS	HiFi-GAN V1 [16]	0.97	5.77 ± 0.61	0.23 ± 0.07	2.0	4.88 ± 0.06	0.19 ± 0.05	**1.7**
	HiFi-GAN V2	**0.15**	5.63 ± 0.47	0.22 ± 0.08	2.1	4.97 ± 0.68	0.20 ± 0.05	2.1
	MS-iSTFT-HiFi-GAN [36]	0.24	5.66 ± 0.57	0.22 ± 0.09	**1.5**	$\mathbf{4.68 \pm 0.58}$	0.20 ± 0.06	2.0
	MS-FC-HiFi-GAN [36]	0.23	5.44 ± 0.47	0.22 ± 0.07	1.7	4.79 ± 0.68	0.21 ± 0.06	2.0
	Vocos	**0.15**	5.49 ± 0.63	0.22 ± 0.08	2.1	4.77 ± 0.62	0.20 ± 0.06	2.5
	WaveNeXt	**0.15**	$\mathbf{5.36 \pm 0.49}$	$\mathbf{0.21 \pm 0.07}$	2.0	4.74 ± 0.50	$\mathbf{0.18 \pm 0.05}$	**1.7**
	Ground truth	N/A	N/A	N/A	1.7	N/A	N/A	1.9



(a) Analysis-synthesis: Female

(b) Analysis-synthesis: Male

(c) JETS-based end-to-end text-to-speech: Female

(d) JETS-based end-to-end text-to-speech: Male

Fig. 6. Results of MOS tests for the analysis–synthesis and JETS-based end-to-end text-to-speech conditions with 20 listening subjects. The confidence level is 95%. The blue asterisks and connecting line indicate the significant differences between two models in the T-test.

5. EXPERIMENTS

To compare the proposed WaveNeXt with Vocos [37] and HiFi-GAN-based models [11, 35, 36], experiments were conducted for both the analysis–synthesis condition with $f_{\mathrm{s}} = 24$ kHz and the E2E TTS conditions with $f_{\mathrm{s}} = 24$ kHz and 48 kHz. All the neural network models were implemented using PyTorch [41] and trained on an NVIDIA Tesla A100 GPU with 40 GB of memory.

5.1. Experimental conditions

Dataset: The experiments were conducted using Japanese single-speaker speech corpora of female and male professional speakers from the Hi-Fi-CAPTAIN corpus [40]. Each corpus contained 18,662 utterances (about 20 hours), all of which were parallel between the female and male corpora. The training, validation, and test sets contained 18,372, 250, and 40 utterances, respectively. The input acoustic features were 80-dimensional mel-spectrograms bandlimited to 7600 Hz. The FFT and shift lengths for $f_{\mathrm{s}} = 24$ kHz were 1,024 and 256 samples, and those for $f_{\mathrm{s}} = 48$ kHz were 2,048 and 512 samples, respectively.

Model setting: In the experiments, all the models were trained and inferred by modifying a JETS-based E2E TTS model implemented using ESPnet2-TTS [42] (`https://is.gd/vbqxeB`). For the analysis–synthesis condition, the FastSpeech 2-based acoustic model and MAS in JETS were omitted, and mel-spectrogram sequences were directly input to the neural vocoder models. Each model was trained for up to 1 million iterations. For $f_{\mathrm{s}} = 24$ kHz (Fig. 1), the upsampling rates and kernel sizes of the transposed convolutional layers for HiFi-GAN were [8, 8, 2, 2] and [16, 16, 4, 4], and those for MS-iSTFT-HiFi-GAN and MS-FC-HiFi-GAN were [4, 4] and [8, 8], respectively. For $f_{\mathrm{s}} = 48$ kHz, the upsampling rates and kernel sizes for HiFi-GAN were [8, 8, 2, 2, 2] and [16, 16, 4, 4, 4], and those for MS-iSTFT-HiFi-GAN and MS-FC-HiFi-GAN were [4, 4, 2] and [8, 8, 4], respectively. Both HiFi-GAN V1 with

Table 2. Results of objective evaluations for the full-band JETS-based end-to-end text-to-speech condition. The RTF of the FastSpeech 2-based acoustic model is 0.05.

		Female ($fs = 48$ kHz)		
Neural vocoder	RTF	MCD [dB]	$\log f_\mathrm{o}$ RMSE	CER [%]
HiFi-GAN V1	1.08	5.08 ± 0.34	$\mathbf{0.23 \pm 0.09}$	**1.7**
HiFi-GAN V2	0.17	5.01 ± 0.37	0.24 ± 0.08	2.4
MS-iSTFT-HiFi-GAN	0.31	5.37 ± 0.49	0.25 ± 0.07	2.0
MS-FC-HiFi-GAN	0.30	4.97 ± 0.35	0.24 ± 0.08	2.1
Vocos	**0.16**	5.87 ± 0.40	0.33 ± 0.15	**1.7**
WaveNeXt	**0.16**	$\mathbf{4.96 \pm 0.33}$	0.25 ± 0.10	**1.7**
Ground truth	N/A	N/A	N/A	1.6

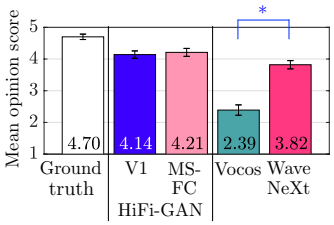


Fig. 7. Results of MOS tests for full-band end-to-end text-to-speech condition with 20 listening subjects. The confidence level is 95%.

an initial channel of 512 (high-fidelity model) and HiFi-GAN V2 with an initial channel of 128 (fast model) were evaluated [11]. The Harvest algorithm [59] was used for fundamental frequency f_o analysis instead of the Dio and Stonemask algorithms [60]. For E2E TTS in Japanese, the G2P function, based on pyopenjtalk enhanced with prosody symbols [61], was used [42]. The model configurations of JETS with HiFi-GAN V1 used the default settings (`https://is.gd/a5UHnP`) with only the sampling frequency changed from 22,050 Hz to 24 kHz. The model configurations of MS-iSTFT-HiFi-GAN and MS-FC-HiFi-GAN were modified from the default settings. The Vocos neural vocoder and JETS-based E2E TTS with Vocos were implemented by integrating an official implementation of Vocos[2] into ESPnet2-TTS. All the model configurations were the same as that of the official implementation, where $n = 8$ in Fig. 1(d), and w_MRD and w_mel in (1) and (2) and w_var and w_align in (3) were 0.1, 45.0, 1.0, and 2.0, respectively. The proposed WaveNeXt neural vocoder and JETS-based E2E TTS with WaveNeXt were implemented by simply replacing the iSTFT layer in Vocos with a linear layer, as shown in Fig. 1(e).

Evaluation criteria: The mel-cepstral distortion (MCD), $\log f_\mathrm{o}$ root-mean-square error (RMSE), and the character error rate (CER) of automatic speech recognition (ASR) were used as the objective evaluation criteria, as in [16, 42]. The MCD and $\log f_\mathrm{o}$ RMSE were calculated by the ESPnet2-TTS toolkit [16, 42]. The CER was calculated by a Conformer-based ASR system, trained using the CSJ corpus [62] by ESPnet2 [63]. The RTFs of all the inference models were measured on an AMD EPYC 7542 CPU (1 core). To evaluate the synthesized speech subjectively, mean opinion score (MOS) tests [64] were conducted. Each subject evaluated 310 samples (ten utterances $\times$ 31 conditions) and rated the naturalness of each sample on a five-point scale. Twenty adult Japanese native speakers without hearing loss participated using headphones.

5.2. Results of experiments

To compare the characteristics of the Vocos and WaveNeXt generators, the generator loss values of these models, including E2E TTS models trained using the female corpus, are plotted in Fig. 5. Although the loss values of the Vocos-based models converged quickly, those of the WaveNeXt-based models converged more gradually and finally became slightly lower than those of the Vocos-based models. These results indicate that the proposed WaveNeXt generator is superior to the Vocos generator, provided that sufficient training is performed.

Table 1 and Fig. 6 show the results of the objective and subjective evaluations for $f_\mathrm{s} = 24$ kHz, respectively. HiFi-GAN V2 was not included in the MOS tests for the E2E TTS condition because it was unable to outperform WaveNeXt for the analysis–synthesis condition. The proposed WaveNeXt achieved high inference speed, comparable to that of HiFi-GAN V2 and Vocos, and higher than that of the other models. Although WaveNeXt underperformed Vocos with respect to MCD for the analysis–synthesis condition, WaveNeXt outperformed Vocos with respect to the other objective evaluation criteria. Importantly, WaveNeXt outperformed Vocos with respect to synthesis quality for all the conditions (Fig. 6). In particular, WaveNeXt outperformed the other models for the analysis–synthesis condition with the male corpus (Fig. 6(b)) and the E2E TTS condition with the female corpus (Fig. 6(c)), while achieving the fastest inference. Table 2 and Fig. 7 show the results of the objective and subjective evaluations, respectively, for the full-band E2E TTS condition trained using the female corpus with $f_\mathrm{s} = 48$ kHz. MS-iSTFT-HiFi-GAN was not included in the MOS tests because it could not be trained successfully and an aliasing component appeared around 3 kHz in all the synthesized waveforms. Vocos and WaveNeXt achieved faster full-band E2E TTS than the other models, and WaveNeXt outperformed the other models with respect to MCD and CER. Although WaveNeXt underperformed HiFi-GAN and MS-FC-HiFi-GAN with respect to synthesis quality, it significantly outperformed Vocos (Fig. 7).

These results indicate that the trainable linear layer, which can directly predict speech waveform samples, introduced in the proposed WaveNeXt is more suitable for GAN-based training in the time domain than the iSTFT layer used in Vocos. Consequently, replacing the iSTFT layer in Vocos with the trainable linear layer enables WaveNeXt to achieve higher synthesis quality than Vocos while preserving its inference speed. Although WaveNeXt outperformed Vocos, it underperformed HiFi-GAN-based models for some conditions, particularly the full-band E2E TTS condition. Therefore, future work includes improving the synthesis quality of WaveNeXt by introducing extended neural network models [65, 66].

6. CONCLUSION

This paper proposed an alternative ConvNeXt-based fast neural vocoder, WaveNeXt, in which the iSTFT layer in Vocos is replaced with a trainable linear layer that can directly predict speech waveform samples without STFT spectra. Additionally, JETS-based E2E TTS models can also be constructed with Vocos and WaveNeXt. Furthermore, full-band models with $f_\mathrm{s} = 48$ kHz were investigated. The results of experiments for both the analysis–synthesis and E2E TTS conditions demonstrate that the proposed WaveNeXt can achieve higher quality synthesis than Vocos while preserving its inference speed. These results indicate that the sophisticated ConvNeXt component is important for fast neural vocoding and the iSTFT operation is not necessarily required.

[2] `https://github.com/charactr-platform/vocos`

7. REFERENCES

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proc. SSW9*, Sept. 2016, p. 125.

[2] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. ICML*, July 2018, pp. 2415–2424.

[3] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, May 2019, pp. 5826–7830.

[4] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga, and H. Kawai, "Full-band LPCNet: A real-time neural vocoder for 48 khz audio with a CPU," *IEEE Access*, vol. 9, pp. 94923–94933, 2021.

[5] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," in *Proc. ICLR*, May 2021.

[6] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *Proc. ICLR*, May 2021.

[7] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Noise level limited sub-modeling for diffusion probabilistic vocoders," in *Proc. ICASSP*, June 2021, pp. 6014–6018.

[8] Y. Koizumi, K. Yatabe, H. Zen, and M. Bacchiani, "WaveFit: An iterative and non-autoregressive neural vocoder based on fixed-point iteration," in *Proc. SLT*, Jan. 2023.

[9] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. NeurIPS*, Dec. 2019, pp. 14910–14921.

[10] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech," in *Proc. SLT*, Jan. 2021, pp. 492–498.

[11] 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.

[12] 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.

[13] 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.

[14] H. Chung, S.-H. Lee, and S.-W. Lee, "Reinforce-Aligner: Reinforcement alignment search for robust end-to-end text-to-speech," in *Proc. Interspeech*, Aug. 2021, pp. 3635–3639.

[15] E. Casanova, J. Weber, C. D Shulby, A. C. Junior, E. Gölge, and M.r A Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proc. ICML*, July 2022, pp. 2709–2720.

[16] 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.

[17] B. Nguyen, F. Cardinaux, and S. Uhlich, "AutoTTS: End-to-end text-to-speech synthesis through differentiable duration modeling," in *Proc. ICASSP*, June 2023.

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

[19] 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.

[20] Z. Zhang, Y. Zheng, X. Li, and L. Lu, "WeSinger 2: Fully parallel singing voice synthesis via multi-singer conditional adversarial training," in *Proc. ICASSP*, June 2023.

[21] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "HiFi++: A unified framework for bandwidth extension and speech enhancement," in *Proc. ICASSP*, June 2023.

[22] K. Kobayashi, T. Hayashi, and T. Toda, "Low-latency electrolaryngeal speech enhancement based on Fastspeech2-based voice conversion and self-supervised speech representation," in *Proc. ICASSP*, June 2023.

[23] Y.-C. Wu, I. D. Gebru, D. Marković, and A. Richard, "AudioDec: An open-source streaming high-fidelity neural audio codec," in *Proc. ICASSP*, June 2023.

[24] R. Komatsu, Y. Kimura, T. Okamoto, and T. Shinozaki, "Continuous action space-based spoken language acquisition agent using residual sentence embedding and transformer decoder," in *Proc. ICASSP*, June 2023.

[25] R. Yoneyama, Y.-C. Wu, and T. Toda, "Source-Filter HiFi-GAN: Fast and pitch controllable high-fidelity neural vocoder," in *Proc. ICASSP*, June 2023.

[26] 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.

[27] E. Cohen, F. Kreuk, and J. Keshet, "Speech time-scale modification with GANs," *IEEE Signal Process. Lett.*, vol. 29, pp. 1067–1071, 2022.

[28] E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstoft, "Generative models for sound field reconstruction," *J. Acoust. Soc. Am.*, vol. 153, no. 2, pp. 1179–1190, Feb. 2023.

[29] J.-H. Kim, S.-H. Lee, J.-H. Lee, and S.-W. Lee, "Fre-GAN: Adversarial frequency-consistent audio synthesis," in *Proc. Interspeech*, Aug. 2021, pp. 2197–2201.

[30] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," in *Proc. Interspeech*, Aug. 2021, pp. 2207–2211.

[31] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive GAN for conditional waveform synthesis," in *Proc. ICLR*, Apr. 2022.

[32] T. Kaneko, H. Kameoka, K. Tanaka, and S. Seki, "Wave-U-Net discriminator: Fast and lightweight discriminator for generative adversarial network-based speech synthesis," in *Proc. ICASSP*, June 2023.

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

[34] 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.

[35] 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.

[36] 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).

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

[38] 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.

[39] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," Nov. 2019.

[40] 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.

[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, Dec. 2019, pp. 8024–8035.

[42] 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.

[43] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," *arXiv:1606.08415*, 2016.

[44] T. Nguyen, "Near-perfect-reconstruction pseudo-QMF banks," *IEEE Trans. Signal Process.*, vol. 42, no. 1, pp. 65–76, Jan. 1994.

[45] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "Subband WaveNet with overlapped single-sideband filterbanks," in *Proc. ASRU*, Dec. 2017, pp. 698–704.

[46] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "An investigation of subband WaveNet vocoder covering entire audible frequency range with limited acoustic features," in *Proc. ICASSP*, Apr. 2018, pp. 5654–5658.

[47] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Improving FFTNet vocoder with noise shaping and subband approaches," in *Proc. SLT*, Dec. 2018, pp. 304–311.

[48] P. P. Vaidyanathan, "Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial," *Proc. IEEE*, vol. 78, no. 1, pp. 56–93, Jan. 1990.

[49] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. CVPR*, June 2016, pp. 1874–1883.

[50] 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.

[51] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.

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

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

[54] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 495–507, 2021.

[55] 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.

[56] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech*, Aug. 2017, pp. 498–502.

[57] 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.

[58] 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.

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

[60] 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.

[61] 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.

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

[63] 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.

[64] ITU-T Recommendation P. 800, *Methods for subjective determination of transmission quality*, 1996.

[65] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. CVPR*, June 2023, pp. 16113–16142.

[66] 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.