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 speech 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× 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.

https://is.gd/duv0DF

© 2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
2. **HI-FI-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. (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.

**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 $(B, 1, N \times T)$, 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 $1 \times 1$ 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].

---

**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.
tor (MRD) used in UnivNet [30] are employed. The loss functions of the generator and discriminators for MPD and MRD, \( \ell_{G,\text{MPD}} \), \( \ell_{G,\text{MRD}} \), \( \ell_{D,\text{MPD}} \), and \( \ell_{D,\text{MRD}} \) are the adversarial loss functions of the generator and discriminators for MPD and MRD, \( \ell_{F,\text{MPD}} \) and \( \ell_{F,\text{MRD}} \) are the feature matching loss functions for MPD and MRD, \( \ell_{\text{mel}} \) is the mel-spectrogram \( L1 \) loss between the ground-truth and synthesized speech waveforms, and \( w_{\text{MRD}} \) and \( w_{\text{mel}} \) 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].

### 3. VOCOS

In contrast to HiFi-GAN-based models, which have gradual up-sampling 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 \( \phi = \arctan2(\sin(\mathbf{p}), \cos(\mathbf{p})) \), respectively. The complex STFT spectrum is then obtained as \( \mathbf{M} \cdot e^{j\phi} \), 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:

\[
\mathcal{L}_G = \ell_{G,\text{MPD}} + w_{\text{MRD}} \ell_{G,\text{MRD}} + \ell_{F,\text{MPD}} + w_{\text{MRD}} \ell_{F,\text{MRD}} + w_{\text{mel}} \ell_{G,\text{mel}} \tag{1}
\]

\[
\mathcal{L}_D = \ell_{D,\text{MPD}} + w_{\text{MRD}} \ell_{D,\text{MRD}}, \tag{2}
\]

where \( \ell_{G,\text{MPD}}, \ell_{G,\text{MRD}}, \ell_{D,\text{MPD}}, \) and \( \ell_{D,\text{MRD}} \) are the adversarial loss functions of the generator and discriminators for MPD and MRD, \( \ell_{F,\text{MPD}} \) and \( \ell_{F,\text{MRD}} \) are the feature matching loss functions for MPD and MRD, \( \ell_{\text{mel}} \) is the mel-spectrogram \( L1 \) loss.

### 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 Vocos. 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
The final linear layer and reshaping component in WaveNeXt generator. $T$ is the number of frames and $l_{\text{FFT}}$ and $l_{\text{shift}}$ are the FFT length and shift length of acoustic feature analysis, respectively.

The 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 $l_{\text{FFT}}$ and $l_{\text{shift}}$, 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 $l_{\text{shift}} \times T$ and the reshaping component concatenates all the predicted speech waveform pieces and finally synthesizes a speech waveform with a length of $1 \times l_{\text{shift}} T$. 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 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:

\[
\mathcal{L}_{G, \text{E2E-TTS}} = \mathcal{L}_G + w_{\text{var}} \ell_{\text{var}} + w_{\text{align}} \ell_{\text{align}} \tag{3}
\]

\[
\mathcal{L}_{D, \text{E2E-TTS}} = \mathcal{L}_D, \tag{4}
\]

where $\ell_{\text{var}}$ and $\ell_{\text{align}}$ are the variance loss and alignment loss used in JETS [16], respectively, and $w_{\text{var}}$ and $w_{\text{align}}$ are the weighting coefficients for $\ell_{\text{var}}$ and $\ell_{\text{align}}$, 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 $f_s = 48$ 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 \( f_s \) 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.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Neural vocoder</th>
<th>RTF</th>
<th>MCD [dB]</th>
<th>log ( f_s )</th>
<th>RMSE [dB]</th>
<th>CER [%]</th>
<th>MCD [dB]</th>
<th>log ( f_s )</th>
<th>RMSE [dB]</th>
<th>CER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis-synthesis</td>
<td>HiFi-GAN V1 [11]</td>
<td>0.92</td>
<td>2.21 ± 0.09</td>
<td>0.16 ± 0.06</td>
<td>1.7</td>
<td>2.00 ± 0.10</td>
<td>0.12 ± 0.06</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HiFi-GAN V2 [11]</td>
<td>0.10</td>
<td>2.65 ± 0.10</td>
<td>0.18 ± 0.08</td>
<td>2.1</td>
<td>2.65 ± 0.09</td>
<td>0.13 ± 0.04</td>
<td>2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS-iSTFT-HiFi-GAN [35]</td>
<td>0.19</td>
<td>2.07 ± 0.10</td>
<td>0.15 ± 0.07</td>
<td>1.7</td>
<td>2.01 ± 0.09</td>
<td>0.13 ± 0.05</td>
<td>1.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS-FC-HiFi-GAN [36]</td>
<td>0.18</td>
<td>2.02 ± 0.09</td>
<td>0.16 ± 0.08</td>
<td>1.9</td>
<td>2.00 ± 0.08</td>
<td>0.14 ± 0.06</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vocos [37]</td>
<td>0.10</td>
<td>2.74 ± 0.11</td>
<td>0.16 ± 0.07</td>
<td>2.0</td>
<td>3.05 ± 1.16</td>
<td>0.14 ± 0.06</td>
<td>2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WaveNeXt</td>
<td>0.10</td>
<td>2.86 ± 0.12</td>
<td>0.17 ± 0.07</td>
<td>1.6</td>
<td>4.32 ± 0.33</td>
<td>0.13 ± 0.05</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JETS-based E2E TTS</td>
<td>HiFi-GAN V1 [16]</td>
<td>0.97</td>
<td>5.77 ± 0.61</td>
<td>0.23 ± 0.07</td>
<td>2.0</td>
<td>4.88 ± 0.06</td>
<td>0.19 ± 0.05</td>
<td>1.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HiFi-GAN V2</td>
<td>0.15</td>
<td>5.63 ± 0.47</td>
<td>0.22 ± 0.08</td>
<td>2.1</td>
<td>4.97 ± 0.68</td>
<td>0.20 ± 0.05</td>
<td>2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS-iSTFT-HiFi-GAN [35]</td>
<td>0.24</td>
<td>5.66 ± 0.57</td>
<td>0.22 ± 0.09</td>
<td>1.5</td>
<td>4.68 ± 0.58</td>
<td>0.20 ± 0.06</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS-FC-HiFi-GAN [36]</td>
<td>0.23</td>
<td>5.44 ± 0.47</td>
<td>0.22 ± 0.07</td>
<td>1.7</td>
<td>4.79 ± 0.68</td>
<td>0.21 ± 0.06</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vocos</td>
<td>0.15</td>
<td>5.49 ± 0.63</td>
<td>0.22 ± 0.08</td>
<td>2.1</td>
<td>4.77 ± 0.62</td>
<td>0.20 ± 0.06</td>
<td>2.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WaveNeXt</td>
<td>0.15</td>
<td>5.36 ± 0.49</td>
<td>0.21 ± 0.07</td>
<td>2.0</td>
<td>4.74 ± 0.50</td>
<td>0.18 ± 0.05</td>
<td>1.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ground truth: N/A

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 \( f_s = 24 \) kHz and the E2E TTS conditions with \( f_s = 24 \) 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 \( f_s = 24 \) kHz were 1,024 and 256 samples, and those for \( f_s = 48 \) 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 \( f_s = 24 \) 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 \([4, 4, 4]\) and \([8, 8, 4]\), respectively. For \( f_s = 48 \) kHz, the upsampling rates and kernel sizes 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 \([4, 4, 2]\) 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 Fast-Speech 2-based acoustic model is 0.05.

<table>
<thead>
<tr>
<th>Neural vocoder</th>
<th>Female (f_s = 48 kHz)</th>
<th>Male (f_s = 24 kHz)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RTF</td>
<td>MCD [dB]</td>
</tr>
<tr>
<td>HiFi-GAN V1</td>
<td>1.08</td>
<td>5.08 ± 0.34</td>
</tr>
<tr>
<td>HiFi-GAN V2</td>
<td>0.17</td>
<td>5.01 ± 0.37</td>
</tr>
<tr>
<td>MS-iSTFT-HiFi-GAN</td>
<td>0.31</td>
<td>5.37 ± 0.49</td>
</tr>
<tr>
<td>MS-FC-HiFi-GAN</td>
<td>0.30</td>
<td>4.97 ± 0.35</td>
</tr>
<tr>
<td>Vocos</td>
<td>0.16</td>
<td>5.87 ± 0.40</td>
</tr>
<tr>
<td>WaveNeXt</td>
<td>0.16</td>
<td>4.96 ± 0.33</td>
</tr>
<tr>
<td>Ground truth</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Evaluation criteria: The mel-cepstral distortion (MCD), log_{10} f_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_{10} f_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 × 31 conditions) and rated the naturalness of each sample on a five-point scale. Twenty adult Japanese native speakers without hearing loss participated using headpho

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_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_s = 48 kHz. MS-iSTFT-HiFi-GAN was not included in the MOS tests because it could not be trained successfully and a 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_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.

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%.
7. REFERENCES


