IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 November 2023, accepted 13 February 2024, date of publication 16 February 2024, date of current version 4 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3366707

RESEARCH ARTICLE

Fast Neural Speech Waveform Generative Models With Fully-Connected Layer-Based Upsampling

HARUKI YAMASHITA[1,2], TAKUMA OKAMOTO[2], (Member, IEEE), RYOICHI TAKASHIMA[1], (Member, IEEE), YAMATO OHTANI[2], (Member, IEEE), TETSUYA TAKIGUCHI[1], (Member, IEEE), TOMOKI TODA[2,3], (Senior Member, IEEE), AND HISASHI KAWAI[2], (Member, IEEE)

[1]Graduate School of System Informatics, Kobe University, Kobe 657-8501, Japan
[2]National Institute of Information and Communications Technology, Kyoto 619-0289, Japan
[3]Information Technology Center, Nagoya University, Nagoya 464-8601, Japan

Corresponding author: Haruki Yamashita (hyamashita@stu.kobe-u.ac.jp)

ABSTRACT Although end-to-end (E2E) text-to-speech (TTS) models with HiFi-GAN-based neural vocoder (e.g. VITS and JETS) can achieve human-like speech quality with fast inference speed, these models still have room to further improve the inference speed with a CPU for practical implementations because HiFi-GAN-based neural vocoder unit is a bottleneck. Additionally, HiFi-GAN is widely used not only for TTS but also for many speech and audio applications. To accelerate HiFi-GAN while maintaining the synthesis quality, Multi-stream (MS)-HiFi-GAN, iSTFTNet and MS-iSTFT-HiFi-GAN have been proposed. Although inverse short-term Fourier transform (iSTFT)-based fast upsampling is introduced in iSTFTNet and MS-iSTFT-HiFi-GAN, we first find that the predicted intermediate features input to the iSTFT layer are completely different from the original STFT spectra due to the redundancy of the overlap-add operation in iSTFT. To further improve the synthesis quality and inference speed, we propose FC-HiFi-GAN and MS-FC-HiFi-GAN by introducing trainable fully-connected (FC) layer-based fast upsampling without overlap-add operation instead of the iSTFT layer. The experimental results for unseen speaker synthesis and E2E TTS conditions show that the proposed methods can slightly accelerate the inference speed and significantly improve the synthesis quality in JETS-based E2E TTS than iSTFTNet and MS-iSTFT-HiFi-GAN. Therefore, the iSTFT layer can be replaced by the proposed trainable FC layer-based upsampling without overlap-add operation in HiFi-GAN-based neural vocoders.

INDEX TERMS End-to-end text-to-speech, fully-connected layer-based upsampling, iSTFTNet, multi-stream HiFi-GAN, neural vocoder.

I. INTRODUCTION

In recent years, text-to-speech (TTS) technology, which generates speech waveforms from input text, can synthesize high-quality speech as good as human speech by using deep learning techniques such as Tacotron 2 [1] combined with WaveNet-based neural vocoder [2]. However, this system requires large computing resources such as a GPU to generate speech waveforms, so it was necessary to reduce the model size and improve the inference speed. To achieve the high-speed inference while maintaining the synthesis quality, several end-to-end (E2E) TTS models have been proposed that can synthesize speech waveforms directly from input text or phoneme sequences with a single neural network [3], [4], [5], [6], [7], [8], [9]. Especially, VITS [7] and JETS [8] can achieve human-like quality and real-time inference. However, these models have room for improving the inference speed with a single CPU because HiFi-GAN [10]-based neural vocoder unit used in these models is a bottleneck in the inference speed although Glow-TTS [11]-based

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Huang.

acoustic model for VITS and Fastspeech 2 [4]-based acoustic model for JETS can realize quite fast inference with a single CPU.

An effective approach to improving the inference speed of TTS models is to accelerate neural vocoders. Vocoder is a speech waveform generator that can convert arbitrary parameters, such as acoustic features for analysis-synthesis or intermediate features in E2E TTS models, into speech waveforms. Initially, conventional TTS models have employed signal-processing-based source-filter vocoders (e.g. STRAIGHT [12] and WORLD [13]). However, the synthesis quality was not good enough. Subsequently, by using WaveNet vocoder [2], which introduces an autoregressive (AR) neural network, TTS models can synthesize speech waveforms that closely resemble human speech. However, to achieve human-like speech, neural vocoders require large computational resources such as a GPU. To realize real-time inference, many AR models [14], [15], [16], [17], [18], [19], [20] and non-AR models [21], [22], [23], [24], [25], [26] have been proposed. Although these models can synthesize high-fidelity speech waveforms, a GPU is required for real-time inference. Compared with these models, MelGAN [27], Multi-band MelGAN [28] and HiFi-GAN [10] are based on generative adversarial network (GAN) [29] and can realize real-time inference with a single CPU by introducing upsampling-based generators. Especially, HiFi-GAN can realize human-like quality synthesis for both single and multi-speaker models[1] and is becoming a de facto standard of neural vocoders and is widely used not only for TTS [6], [7], [8], [9], [31], [32] but also for many speech and audio applications, such as voice conversion [33], [34], singing voice synthesis [35], speech enhancement [36], [37], bandwidth extension [36], neural audio codec [38], automatic spoken language acquisition [39], fundamental frequency (f_0) [40] controllable neural vocoders [41], [42], [43], speech rate conversion [43], [44] and sound field reconstruction [45]. Additionally, extended models have also been investigated [46], [47], [48], [49]. Although the inference speed of HiFi-GAN is fast, the real-time factor (RTF) is more than 0.5 on a single CPU. If the duration of a waveform is 10 s, the inference time is more than 5 s. Therefore, it is important to further accelerate the inference speed of HiFi-GAN with a single CPU for practical applications.

To accelerate the inference speed of HiFi-GAN while maintaining the synthesis quality, Multi-stream (MS)-HiFi-GAN [50] and iSTFTNet [51] have been proposed by replacing the final 4× upsampling layers of HiFi-GAN with lightweight fast upsampling layers.[2] Additionally, by efficiently combining these models, MS-iSTFT-HiFi-GAN [53] has also been proposed in VITS-based E2E TTS model and can realize 4 times faster inference than vanilla HiFi-GAN while maintaining the synthesis quality. Focusing on iSTFTNet, this architecture can reasonably achieve the acceleration of HiFi-GAN by using the inverse short-term Fourier transform (iSTFT)-based fast upsampling. However, we first show that the intermediate features input to the iSTFT layer are completely different from the original STFT spectra due to the redundancy of the overlap-add operation in iSTFT. This means that the iSTFT-based upsampling does not work as expected and there is room for improvement.

To further improve the synthesis quality and inference speed of iSTFTNet and MS-iSTFT-HiFi-GAN, we propose simple but efficient models, FC-HiFi-GAN and MS-FC-HiFi-GAN by replacing iSTFT layer-based upsampling using fixed weights based on the Fourier basis and overlap-add operation with trainable fully-connected (FC) layer-based lightweight upsampling without overlap-add operation. In experiments for analysis-synthesis-based unseen speaker synthesis and VITS- and JETS-based E2E TTS conditions, we show that the proposed methods can also realize fast and high-fidelity synthesis as well as iSTFTNet and MS-iSTFT-HiFi-GAN, slightly improve the inference speed than iSTFTNet and MS-iSTFT-HiFi-GAN, and significantly improve the synthesis quality for JETS-based E2E TTS by trainable but lightweight upsampling without overlap-add operation.

The rest of this paper is organized as follows. Conventional HiFi-GAN-based fast neural vocoders and E2E TTS models, VITS and JETS, are briefly introduced in Sec. II. The issues for iSTFT-based upsampling are explained in Section III. FC-HiFi-GAN and MS-FC-HiFi-GAN are then proposed in Sec. IV. Section V describes experiments to compare the proposed FC-HiFi-GAN and MS-FC-HiFi-GAN with the conventional models for analysis-synthesis-based unseen speaker synthesis and VITS- and JETS-based E2E TTS conditions. Finally, conclusions are presented in Section VI.

II. CONVENTIONAL MODELS

A. HiFi-GAN-BASED FAST NEURAL VOCODERS

1) HiFi-GAN [10]

HiFi-GAN is a GAN-based neural vocoder consisting of a generator and two superior discriminators. The generator synthesizes speech waveforms from acoustic features, such as mel-spectrograms, by progressively upsampling the input features ($8\times \rightarrow 8\times \rightarrow 2\times \rightarrow 2\times$) using transposed convolutional layers with residual blocks as shown in Fig. 1(a). With the efficient upsampling-based generator and sophisticated discriminators, HiFi-GAN can realize high-fidelity and fast speech synthesis.

2) MS-HiFi-GAN [50]

As Multi-band MelGAN [28], HiFi-GAN can be easily accelerated by replacing the last two layers for final 4× upsampling to multi-rate signal processing [54]-based sub-band synthesis filter [55] as used in [16] where the four sub-band

[1] Recently, another GAN-based model, WaveFit [30], has been proposed. Although it can realize higher synthesis quality than HiFi-GAN and comparable inference speed to HiFi-GAN, the training cost is higher than HiFi-GAN and no implementation is provided.

[2] Although MISRNet [52] has also been proposed, it can only accelerate the inference speed of HiFi-GAN on a GPU and cannot accelerate the inference speed on a CPU. Therefore, it is not considered in this paper.

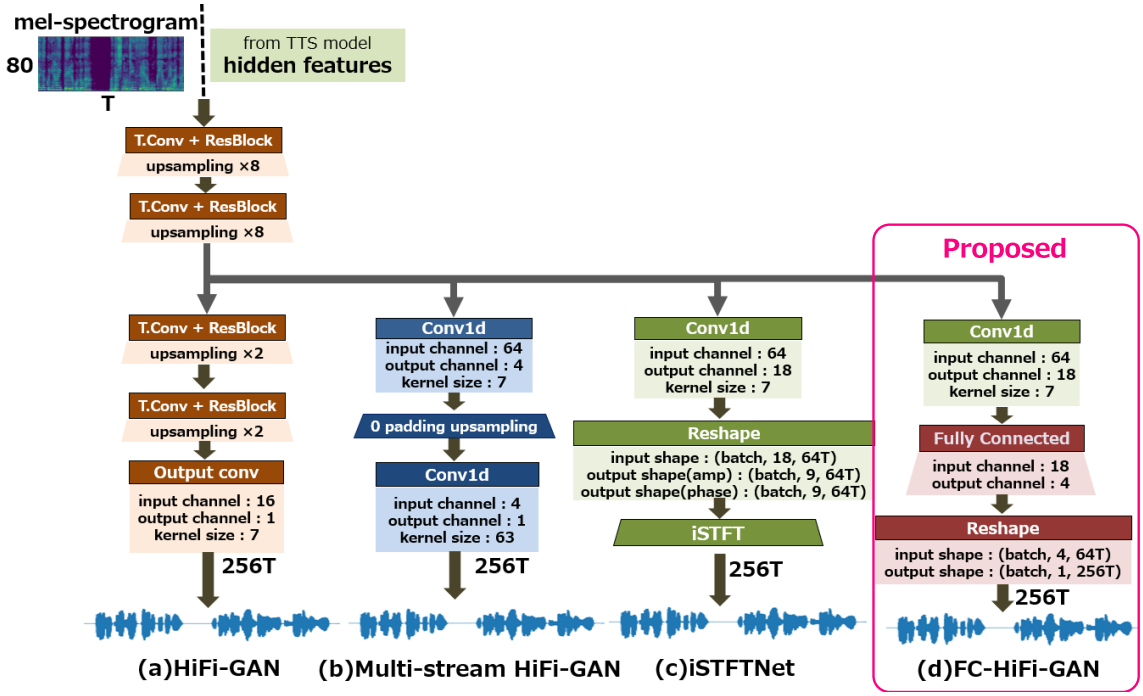


FIGURE 1. Architectures of (a) HiFi-GAN, (b) Multi-stream HiFi-GAN, (c) iSTFTNet, and (d) proposed FC-HiFi-GAN generators. T, T.Conv, ResBlock and Conv1d are the number of frames of mel-spectrograms for analysis-synthesis condition or hidden features for E2E TTS condition, transposed convolutional layer, residual block and 1-dimensional convolutional layer.

output waveforms are upsampled by zero-padding and then a full-band speech waveform is synthesized by the synthesis filter. However, the multi-band structure with constant synthesis filter is too restrictive to train HiFi-GAN because the sophisticated HiFi-GAN discriminators can easily distinguish between real and synthetic speech. By replacing the sub-band synthesis filter, which can be regarded as a convolutional layer with fixed weights without bias, with a trainable convolutional layer without bias, MS-HiFi-GAN can be successfully trained by decomposing the four output waveforms in a data-driven manner. Then, MS-HiFi-GAN can successfully accelerate the inference speed of HiFi-GAN while maintaining the synthesis quality. The architecture of the MS-HiFi-GAN generator is shown in Fig. 1(b).

3) iSTFTNet [51]

Similar to the sub-band synthesis filter in Multi-band MelGAN [28], iSTFT can also be regarded as an upsampling operation. For both accelerating HiFi-GAN and making the best use of input mel-spectrogram structure, iSTFTNet replaces the last two layers for final 4× upsampling of HiFi-GAN with iSTFT-based fast upsampling as shown in Fig. 1(c). In iSTFTNet, the amplitude and phase components of the STFT spectra are predicted by a 1D convolutional layer before the iSTFT layer. Compared with MS-HiFi-GAN with trainable lightweight upsampling, iSTFTNet can also successfully accelerate the inference speed of HiFi-GAN while maintaining the synthesis quality although the iSTFT layer with fixed weights based on the Fourier basis is not trainable.

4) MS-iSTFT-HiFi-GAN [53]

By combining a trainable convolutional layer-based upsampling for MS-HiFi-GAN and iSTFT-based upsampling for iSTFTNet, MS-iSTFT-HiFi-GAN has been proposed to further accelerate HiFi-GAN-based neural vocoder. MS-iSTFT-HiFi-GAN is introduced in the speech waveform synthesizer component for VITS-based E2E TTS. The architecture of the MS-iSTFT-HiFi-GAN generator is depicted in Fig. 2(e). Although MS-iSTFT-HiFi-GAN is twice as fast as MS-HiFi-GAN and iSTFTNet, it can still maintain the synthesis quality.

B. E2E TTS MODELS

1) VITS [7]

VITS is proposed as an E2E TTS model extended from Glow-TTS [11]. In the training of Glow-TTS, the target mel-spectrograms are converted to Gaussian white noise by the Flow [56]-based decoder, and the alignment between the hidden features converted from the input text and converted white noise is gradually obtained by monotonic alignment search (MAS) [11] without external aligners. In the inference, the upsampled hidden features are converted to the target mel-spectrograms by Flow-based inverse transformation. In VITS, the target linear-spectrograms are converted to the latent variables based on variational autoencoder (VAE) [57], and the latent variables instead of mel-spectrograms are converted not only to Gaussian white noise by the Flow-based decoder but also to the target speech waveforms by HiFi-GAN-based neural vocoder. All

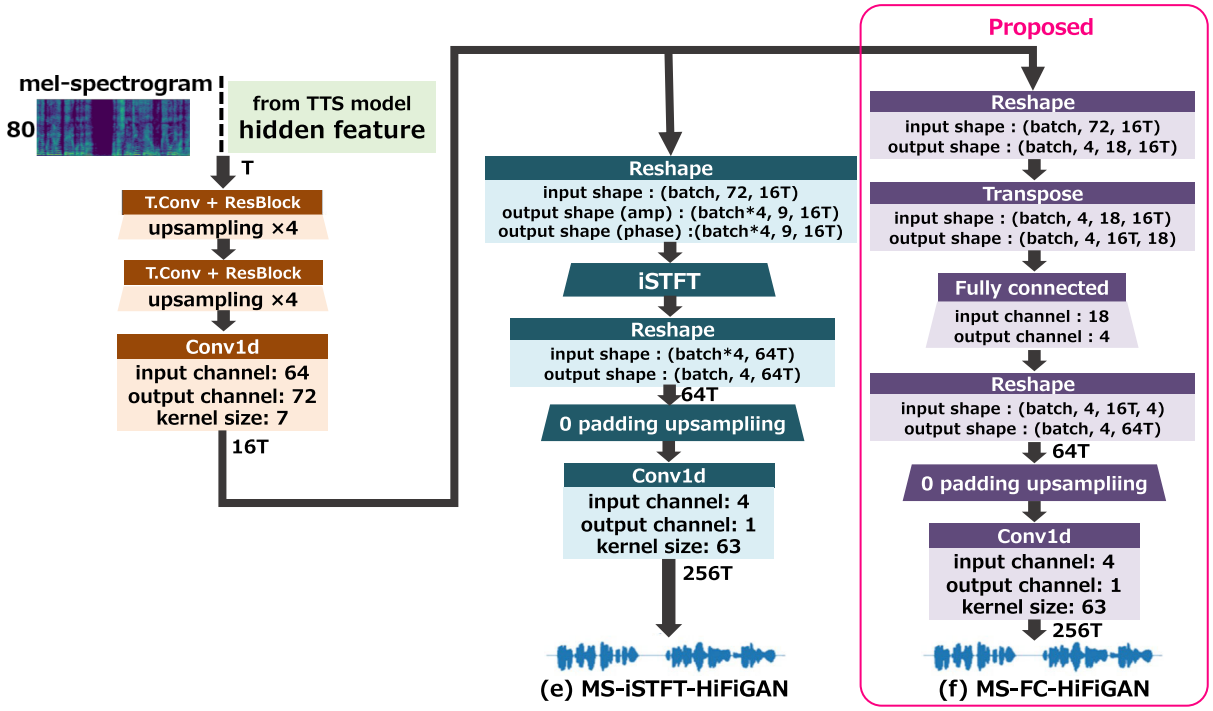


FIGURE 2. Architectures of (e) MS-iSTFT-HiFi-GAN and (f) proposed MS-FC-HiFi-GAN generators. T, T.Conv, ResBlock and Conv1d are the number of frames of mel-spectrograms for analysis-synthesis condition or hidden features for E2E TTS condition, transposed convolutional layer, residual block and 1-dimensional convolutional layer.

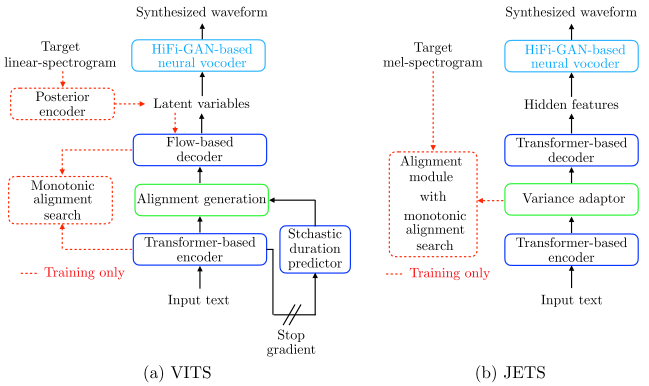


FIGURE 3. Architectures of VITS and JETS generators.

the network components are jointly trained with the same discriminators for HiFi-GAN, and the intermediate latent variables are optimized to minimize the training loss. Then, VITS can realize higher-quality TTS than the cascade model with Glow-TTS and HiFi-GAN [7]. The architecture of the VITS generator is shown in Fig. 3(a). In MS-iSTFT-VITS, MS-iSTFT-HiFi-GAN (Fig. 2(e)) is used for the neural vocoder instead of vanilla HiFi-GAN [53].

2) JETS [8]

Compared with VITS, which efficiently introduces three kinds of deep generative models, Flow [56], VAE [57] and GAN [29], JETS is a simpler E2E TTS model while realizing higher synthesis quality than VITS [8]. JETS is realized by joint training of FastSpeech 2 [4]-based acoustic model and HiFi-GAN-based neural vocoder with the same discriminators for HiFi-GAN without intermediate mel-spectrograms nor external aligners although FastSpeech 2 [4] requires an external aligner, such as Montreal Forced Aligner [58]. In JETS, an alignment training framework proposed in [59] with MAS is introduced, 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 as VITS.

III. ISSUES FOR ISTFT LAYER-BASED UPSAMPLING

In this section, we first show that the iSTFT-based upsampling used in iSTFTNet and MS-iSTFT-HiFi-GAN does not work as expected. As described in Sec. II-A3, the amplitude and phase components of the STFT spectra are inferred by the 1D convolutional layer before the iSTFT layer, and high-fidelity speech waveforms can be synthesized by the final iSTFT layer-based fast upsampling in iSTFTNet. To explain the actual behavior of iSTFTNet, Figure 4 shows the magnitude and phase components of the STFT spectrum of an original female speech waveform used in the experiments of analysis-synthesis condition conducted in Sec. V, those estimated by iSTFTNet, and those reanalyzed from the speech waveform synthesized by using the estimated STFT spectrum, respectively. The estimated magnitude and phase components (Fig. 4(b)) differ from those of the orig-

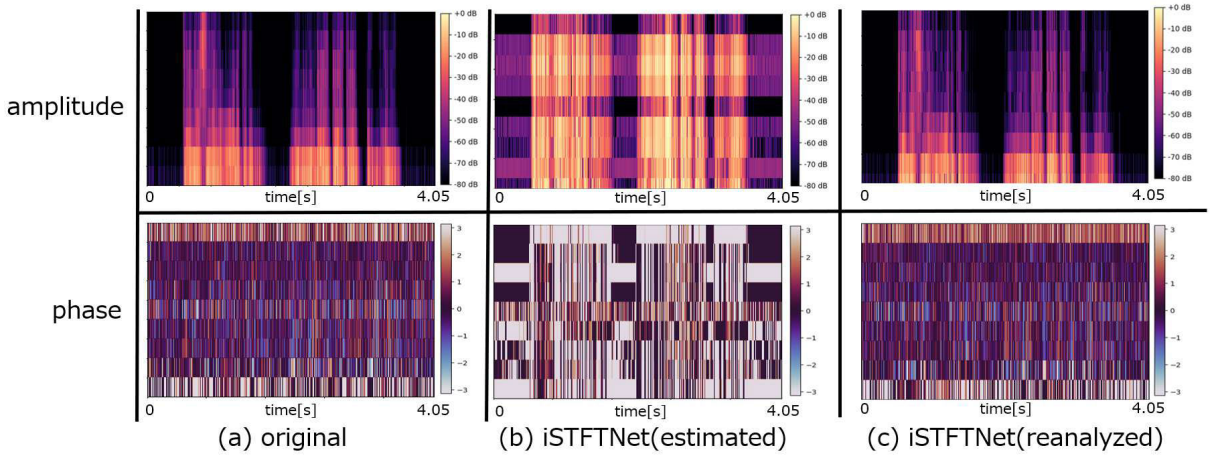


FIGURE 4. (a) amplitude and phase components of STFT spectrum of an original speech waveform (jvs001-BASIC5000-0025), (b) those estimated by iSTFTNet trained using JVS corpus, (c) those reanalyzed from the speech waveform synthesized by using (b).

inal (Fig. 4(a)). This result indicates that iSTFTNet cannot perfectly predict the magnitude and phase components of the STFT spectra. However, the reanalyzed magnitude and phase components (Fig. 4(c)) are indistinguishable from those of the original Fig. 4(a). When the fast Fourier transform (FFT) length and shift length of acoustic feature analysis in STFT are M and N, $M/N = Q$ samples are summed for each sample in iSTFT by the overlap-add operation (Fig. 5(a)). Therefore, the overlap-add operation in iSTFT has the "redundancy" for $Q \geq 2$. By the redundancy of the overlap-add operation and the GAN-based training in the time domain, the magnitude and phase components estimated by iSTFTNet, that differ from those of the original, can still synthesize high-fidelity speech waveforms. Conversely, iSTFTNet is trained to estimate STFT spectra for synthesizing high-quality speech waveforms through 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 introduced in iSTFTNet. Additionally, there is a room for improvement in the iSTFT layer-based upsampling with untrainable fixed weights based on the Fourier basis compared with MS-HiFi-GAN with trainable fast upsampling [50].

IV. PROPOSED FULLY-CONNECTED LAYER-BASED TRAINABLE UPSAMPLING WITHOUT OVERLAP-ADD OPERATION: FC-HiFi-GAN AND MS-FC-HiFi-GAN

As described in Sec. III, the iSTFT layer-based upsampling has the following issues.

- The intermediate features inferred by the 1D convolutional layer in iSTFTNet are completely different from the original STFT spectra.
- The iSTFT-based upsampling in iSTFTNet and MS-iSTFT-HiFi-GAN is not suitable for GAN-based training in time domain because GAN-based training in the time domain has no restriction in the STFT domain.
- The iSTFT layer introduces untrainable fixed weights based on the Fourier basis.

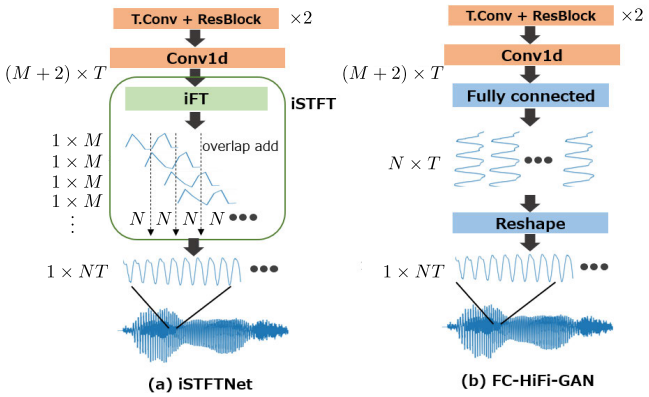


FIGURE 5. (a) iSTFTNet with overlap-add operarion, and (b) proposed FC-HiFi-GAN without overlap-add operation. T is the number of frames, and M and N are the FFT length and shift length of acoustic feature analysis.

Inspired by MS-HiFi-GAN with trainable fast upsampling [50], we propose a simple but efficient FC layer-based fast upsampling to solve the above issues of the iSTFT-based upsampling. Then, we proposed FC-HiFi-GAN and MS-FC-HiFi-GAN by replacing the iSTFT layer-based upsampling in iSTFTNet and MS-iSTFT-HiFi-GAN with the FC layer-based trainable fast upsampling as shown in Figs. 1(d) and 2(f), respectively.

In the proposed FC layer-based upsampling, $N\times$ upsampling is simply realized by a trainable FC layer with output channels of N and reshaping the output tensor shape from (B, N, T) to (B, 1, NT) as shown in Figs. 1(d), 2(f) and 5(b), where B and T are the batch size and number of frames, respectively. The proposed FC-layer-based upsampling is equivalent to sub-pixel convolution (pixel shuffler)-based

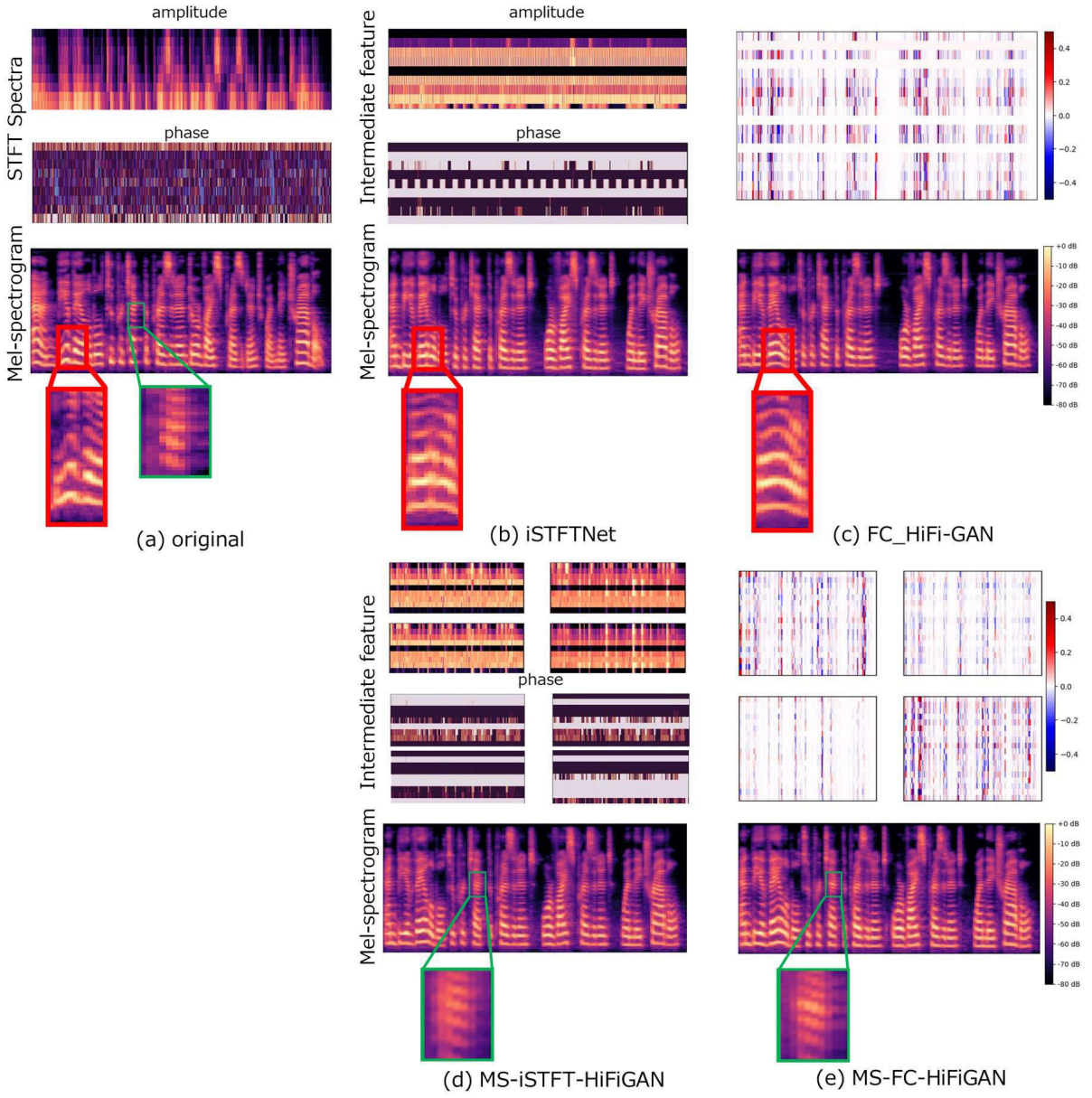


FIGURE 6. (a) mel-spectrogram of original speech waveform and (b) to (e) mel-spectrograms and intermediate features of iSTFTNet, proposed FC-HiFi-GAN, MS-iSTFT-HiFi-GAN and proposed MS-FC-HiFi-GAN input to iSTFT layer or FC layer for JETS-based E2E TTS with LJSpeech corpus (LJ050-0225).

upsampling [60] with 1×1 convolutional layer. As MS-HiFi-GAN [50], the trainable FC layer without bias is introduced.

Additionally, FC can be realized with fewer calculations than iSTFT. When M is a power of 2, inverse FFT is applyed. M audio samples are calculated by $2M \log_2 M$ real number multiplications and $3M \log_2 M$ real number additions in the iFFT. Then, $2MN \log_2 M$ real number multiplications and $3MN \log_2 M + N(Q-1) = 3MN \log_2 M + M - N$ real number additions are required to synthesize N audio samples in the iSTFT because iSTFT is calculated by shifting frame and overlap-add with shift length N. Conversely, FC without bias is calculated as $\mathbf{x} = \mathbf{W}\mathbf{h}$, where $\mathbf{x} \in \mathbb{R}^{N\times 1}$, $\mathbf{W} \in \mathbb{R}^{N\times(M+2)}$, and $\mathbf{h} \in \mathbb{R}^{(M+2)\times 1}$ are the vector of N audio samples, trainable weight matrix of the fully-connected layer, and vector of hidden features, respectively. Then, $(M+2)N$ real number multiplications and $(M+1)N$ real number addictions are required to synthesize N audio samples in the FC. In iSTFTNet and FC-HiFi-GAN, $M = 16$ and $N = 4$. Therefore, FC-based upsampling can realize faster inference than iSTFT-based upsampling.

The FC layer-based upsampling differs from the iSTFT layer-based upsampling in the following important points.

- The weights of FC layer-based upsampling are trainable.
- The FC layer-based upsampling can directly predict speech waveform samples without overlap-add operation (Fig. 5(b)), and it is more suitable for GAN-based

TABLE 1. Results of objective and subjective evaluations of analysis-synthesis condition for unseen speaker synthesis with multi-speaker models.

	RTF	MOS	MCD [dB]	$\log f_0$ RMSE
HiFi-GAN [10]	0.53	3.82 ± 0.11	2.51 ± 0.20	0.19 ± 0.07
iSTFTNet [51]	0.30	3.85 ± 0.10	2.52 ± 0.20	0.19 ± 0.06
MS-HiFi-GAN [50]	0.28	3.96 ± 0.09	2.47 ± 0.19	$\mathbf{0.18 \pm 0.07}$
FC-HiFi-GAN (Proposed)	0.28	3.83 ± 0.11	2.56 ± 0.19	0.18 ± 0.06
MS-iSTFT-HiFi-GAN [53]	0.11	4.00 ± 0.09	$\mathbf{2.43 \pm 0.17}$	0.19 ± 0.07
MS-FC-HiFi-GAN (Proposed)	**0.10**	$\mathbf{4.01 \pm 0.10}$	2.60 ± 0.17	0.20 ± 0.07
Ground truth	N/A	4.19 ± 0.08	N/A	N/A

training in time domain than the indirect estimation of STFT spectra by the iSTFT-based upsampling.

With these features, the proposed FC layer-based upsampling with trainable weights without overlap-add operation is expected to further improve the inference speed and synthesis quality compared to the iSTFT layer-based upsampling with fixed weights based on the Fourier basis and overlap-add operation.

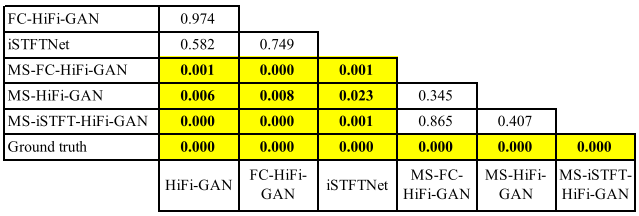


FIGURE 7. Result of T-test for MOS tests in Table 1. Values for $p < 0.05$ (statistically significant) are bold with yellow highlighting.

V. EXPERIMENTS

A. EXPERIMENTS OF ANALYSIS-SYNTHESIS CONDITION FOR UNSEEN SPEAKER SYNTHESIS WITH MULTI-SPEAKER MODELS

To evaluate the proposed FC-HiFi-GAN and MS-FC-HiFi-GAN and to compare them with the conventional HiFi-GAN, MS-HiFi-GAN, iSTFTNet and MS-iSTFT-HiFi-GAN in fundamental analysis-synthesis condition, experiments of analysis-synthesis condition for unseen speaker synthesis with multi-speaker models were first conducted. Some of the speech samples used in the experiments are available online.[3]

1) EXPERIMENTAL CONDITIONS

a: DATASET

We used JVS corpus [61] of parallel 100 and non-parallel 30 sentences read by 100 Japanese speakers with a sampling frequency of 24 kHz. The utterances of 90 speakers (jvs011 to jvs100) were used for the training set, and the non-parallel 30 sentences of the remaining 10 speakers (jvs001 to jvs010) not included in the training set were used for the test set. The input acoustic features were 80-dimensional mel-spectrograms bandlimited to 7600 Hz where the FFT and hop sizes were 1,024 and 256 samples, respectively.

b: MODEL SETTING

In the experiments, HiFi-GAN-based models were trained and inferred by modifying a PyTorch [62]-based open source implementation,[4] and each model was trained up to 2.5 million iterations by using an NVIDIA Tesla V100 GPU. As shown in Figs 1 and 2, the upsampling rates and kernel sizes of the transposed convolutional layers for HiFi-GAN were [8, 8, 2, 2] and [16, 16, 4, 4], those for MS-HiFi-GAN, iSTFTNet and FC-HiFi-GAN were [8, 8] and [16, 16], and those for MS-iSTFT-HiFi-GAN and MS-FC-HiFi-GAN were [4, 4] and [8, 8], respectively. The initial channel of all the models was 512 as HiFi-GAN V1 model [10]. The model configuration of HiFi-GAN was the default setting[5] where only the sampling frequency was changed from 22,050 Hz to 24 kHz. The model configurations of the other models were modified from that of HiFi-GAN.

c: EVALUATION CRITERIA

As objective evaluation criteria, the mel-cepstral distortion (MCD) and $\log f_0$ root mean square error ($\log f_0$ RMSE) between the original and synthesized speech waveforms were evaluated. These values were calculated by using ESPNet2-TTS [63].[6] To measure RTFs, we used an Intel Xeon 6152 CPU (with one core). A mean opinion score (MOS) test with a five-point scale (5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad) [64] was conducted to evaluate the subjective perceptual quality of the ground truth and synthesized speech waveforms. For the MOS test, non-parallel 30 utterances of two female (jvs004 and jvs008) and two male (jvs001 and jvs003) speakers were used. In the MOS test, twenty adult native Japanese speakers without hearing loss listened to the original and synthesized speech samples using headphones and evaluated 140 sentences in total, consisting of 20 sentences of each model and ground truth samples ($6 \times 20 + 20 = 140$).

2) RESULTS OF ANALYSIS-SYNTHESIS EXPERIMENTS

The results of the objective and subjective evaluations for unseen speaker synthesis with multi-speaker models are

[3] Please download the zip file from https://www.okamotocamera.com/UduQZiJzCw3f1Iw.zip in peer review. The page will be published online when the submission is accepted.

[4] https://github.com/kan-bayashi/ParallelWaveGAN

[5] https://github.com/kan-bayashi/ParallelWaveGAN/blob/master/egs/ljspeech/voc1/conf/hifigan.v1.yaml

[6] https://github.com/espnet/espnet/tree/master/egs2/TEMPLATE/asr1/pyscripts/utils/

shown in Table 1. Additionally, Figure 7 shows the result of the T-test for the MOS tests in Table 1. First of all, the proposed MS-FC-HiFi-GAN realized the fastest inference speed and highest synthesis quality compared with the other models although there was no significant difference between the MOS value of MS-FC-HiFi-GAN and those of the other models. As expected, FC-HiFi-GAN and MS-FC-HiFi-GAN without overlap-add operation realized slightly faster inference than iSTFTNet and MS-iSTFT-HiFi-GAN with overlap-add operation while maintaining the synthesis quality by the trainable FC layer.

B. EXPERIMENTS OF E2E TTS CONDITION

The analysis-synthesis condition is a simpler problem because the inputs were ground truth mel-spectrograms. For this reason, there was no significant difference between the MOS values of iSTFTNet and FC-HiFi-GAN, or between those of MS-iSTFT-HiFi-GAN and MS-FC-HiFi-GAN.

Therefore, we evaluated the performance of each neural vocoder in E2E TTS condition, which is a more complex problem than analysis-synthesis condition. For E2E TTS models, we introduced VITS [7], which was used in MS-iSTFT-HiFi-GAN. Additionally, we introduced JETS, which is expected to realize higher quality and more stable synthesis than VITS [8]. The neural vocoder part of each E2E TTS model was changed to iSTFTNet, MS-HiFi-GAN, MS-iSTFT-HiFi-GAN, and MS-FC-HiFi-GAN, and the inference speed and synthesis quality of these models for E2E TTS condition were compared.

1) EXPERIMENTAL CONDITIONS

a: DATASET

In the TTS experiments, we first used LJspeech [65] with a sampling frequency of 22.05 kHz. As the default setting of ESPnet2-TTS [63], 12,600 utterances, 250 utterances and 250 utterances were used for the training, validation and test sets, respectively. Although LJSpeech is widely used in TTS experiments as [7], [8], [51], and [53], the original recordings were distributed as 128 kbps MP3 files and they may contain artifacts introduced by the MP3 encoding [65]. To evaluate the neural vocoder models using a higher quality corpus, we introduced Hi-Fi TTS dataset [66] with a sampling frequency of 44.1 kHz. In the experiments, a clean female speaker corpus (Reader ID: 92) was selected, and normal-band (24 kHz) and full-band (44.1 kHz) E2E TTS models were trained combined with these neural vocoders. As the default setting of Hi-Fi TTS dataset [66], 35,146 utterances, 50 utterances and 100 utterances were used for the training, validation and test sets, respectively.

b: MODEL SETTING

In the experiments, VITS- and JETS-based E2E TTS models were trained and inferred by the modifying PyTorch-based open source implementation provided in ESPnet2-TTS [63]. Each model was trained up to 1.0 million iterations by using four NVIDIA Tesla V100 GPUs.

The FFT and hop sizes of acoustic feature extraction for sampling frequencies of 22.05 kHz and 24 kHz were also 1,024 and 256 samples, respectively. Then, the upsampling rates and kernel sizes of the transposed convolutional layers in the neural vocoder part for 22.05 kHz and 24 kHz were the same as those used in the experiments for analysis-synthesis condition. The model configurations of VITS and JETS with HiFi-GAN for LJSpeech (22.05 kHz) were the default settings.[7][8] The model configurations of the other models for 22.05 kHz and 24 kHz were modified from the default settings.

The FFT and hop sizes of acoustic feature extraction for full-band VITS and JETS with a sampling frequency of 44.1 kHz were 2,048 and 512 samples, respectively. Then, the upsampling rates and kernel sizes of the transposed convolutional layers for HiFi-GAN were [8, 8, 2, 2, 2] and [16, 16, 4, 4, 4] [63], those for MS-HiFi-GAN, iSTFTNet and FC-HiFi-GAN were [8, 8, 2] and [16, 16, 4], and those for MS-iSTFT-HiFi-GAN and MS-FC-HiFi-GAN were [4, 4, 2] and [8, 8, 4], respectively. The model configuration of full-band VITS with HiFi-GAN for 44.1 kHz was the default setting.[9] The model configurations of full-band VITS with the other models were modified from the default setting of full-band VITS. The model configurations of full-band JETS with these models were modified from the default setting of JETS for 22.05 kHz.

c: EVALUATION CRITERIA

As objective evaluation criteria, the MCD, $\log f_0$ RMSE and RTF were also evaluated as the analysis-synthesis condition. Additionlly, the character error rate (CER) of automatic speech recognition (ASR) were measured as in [8] and [63] to evaluate the stability of E2E TTS models. The CER was calculated by a Conformer-based ASR trained using LibriSpeech corpus [67] by ESPnet [68]. A MOS test with a five-point scale was also conducted to evaluate the subjective perceptual quality of the ground truth and synthesized speech waveforms. In the MOS test, twenty adult native English speakers without hearing loss listened to the original and synthesized speech samples using headphones and evaluated 390 sentences in total, consisting of 10 sentences of each model and ground truth samples ($(12 \times 10 + 10) \times 3$ [LJSpeech, Hi-Fi TTS (24 kHz) and Hi-Fi TTS (44.1 kHz)] $= 390$).

2) RESULTS OF E2E TTS EXPERIMENTS

The results of the subjective and objective evaluations for normal-band and full-band E2E TTS conditions are pre-

[7]https://github.com/espnet/espnet/blob/master/egs2/ljspeech/tts1/conf/tuning/train_vits.yaml

[8]https://github.com/espnet/espnet/blob/master/egs2/ljspeech/tts1/conf/tuning/train_jets.yaml

[9]https://github.com/espnet/espnet/blob/master/egs2/jsut/tts1/conf/tuning/train_full_band_vits.yaml

TABLE 2. Results of objective and subjective evaluations of normal-band E2E TTS conditions using LJSpeech corpus and Hi-Fi TTS dataset.

			LJSpeech (22.05 kHz)				Hi-Fi TTS [Reader ID: 92 (female)] (24 kHz)			
E2E TTS model	vocoder	RTF	MOS	MCD [dB]	$\log f_0$ RMSE	CER	MOS	MCD [dB]	$\log f_0$ RMSE	CER
VITS	HiFi-GAN [7]	0.57	3.20 ± 0.15	7.05 ± 0.72	0.28 ± 0.07	4.9	3.79 ± 0.14	7.17 ± 1.45	0.30 ± 0.13	4.2
	iSTFTNet	0.37	3.16 ± 0.15	6.83 ± 0.64	0.28 ± 0.06	4.8	3.50 ± 0.14	7.15 ± 1.19	0.29 ± 0.13	4.1
	MS-HiFi-GAN	0.35	3.35 ± 0.16	6.90 ± 0.64	0.28 ± 0.06	4.5	3.47 ± 0.13	6.88 ± 1.23	0.29 ± 0.12	3.8
	FC-HiFi-GAN	0.35	3.14 ± 0.16	7.00 ± 0.64	0.28 ± 0.07	4.8	3.35 ± 0.17	7.15 ± 1.42	0.30 ± 0.13	3.8
	MS-iSTFT-HiFi-GAN [53]	0.18	3.14 ± 0.16	6.90 ± 0.63	0.28 ± 0.05	5.0	3.34 ± 0.15	7.18 ± 1.41	0.28 ± 0.12	4.2
	MS-FC-HiFi-GAN	**0.17**	3.32 ± 0.16	6.92 ± 0.60	0.29 ± 0.07	5.2	3.53 ± 0.15	7.02 ± 1.16	0.29 ± 0.13	3.6
JETS	HiFi-GAN [8]	0.54	3.73 ± 0.12	6.76 ± 0.54	0.28 ± 0.07	**3.6**	$\mathbf{4.24 \pm 0.11}$	7.02 ± 1.05	0.28 ± 0.12	3.5
	iSTFTNet	0.34	3.11 ± 0.16	$\mathbf{6.54 \pm 0.57}$	0.28 ± 0.07	**3.6**	3.72 ± 0.13	6.99 ± 0.88	0.28 ± 0.11	**3.2**
	MS-HiFi-GAN	0.32	$\mathbf{3.87 \pm 0.13}$	6.69 ± 0.55	0.28 ± 0.07	**3.6**	3.96 ± 0.13	$\mathbf{6.72 \pm 1.00}$	0.28 ± 0.12	**3.2**
	FC-HiFi-GAN	0.32	3.58 ± 0.14	6.60 ± 0.55	$\mathbf{0.26 \pm 0.06}$	**3.6**	3.87 ± 0.13	6.81 ± 0.95	0.28 ± 0.13	**3.2**
	MS-iSTFT-HiFi-GAN	0.15	3.60 ± 0.13	6.62 ± 0.55	0.27 ± 0.06	**3.6**	4.01 ± 0.13	6.89 ± 0.99	$\mathbf{0.27 \pm 0.12}$	3.5
	MS-FC-HiFi-GAN	**0.14**	3.83 ± 0.14	6.64 ± 0.57	0.28 ± 0.07	**3.6**	3.97 ± 0.12	6.91 ± 1.02	$\mathbf{0.27 \pm 0.13}$	**3.2**
	ground truth	N/A	4.09 ± 0.12	N/A	N/A	3.4	4.20 ± 0.11	N/A	N/A	3.0

TABLE 3. Results of objective and subjective evaluations of full-band E2E TTS conditions using Hi-Fi TTS dataset.

			Hi-Fi TTS [Reader ID: 92 (female)] (44.1 kHz)			
E2E TTS model	vocoder	RTF	MOS	MCD [dB]	$\log f_0$ RMSE	CER
Full-band VITS	HiFi-GAN	0.64	3.59 ± 0.13	7.19 ± 1.17	0.29 ± 0.11	3.7
	iSTFTNet	0.41	3.78 ± 0.14	7.16 ± 1.29	0.29 ± 0.12	4.2
	MS-HiFi-GAN	0.40	3.66 ± 0.13	7.26 ± 1.42	0.31 ± 0.14	4.5
	FC-HiFi-GAN	0.40	3.56 ± 0.14	7.43 ± 1.19	0.30 ± 0.11	4.0
	MS-iSTFT-HiFi-GAN	0.19	3.63 ± 0.14	7.17 ± 1.20	0.31 ± 0.15	4.6
	MS-FC-HiFi-GAN	**0.18**	3.57 ± 0.13	7.34 ± 1.30	0.29 ± 0.11	4.5
Full-band JETS	HiFi-GAN	0.63	4.03 ± 0.12	7.08 ± 0.95	$\mathbf{0.26 \pm 0.11}$	**3.3**
	iSTFTNet	0.41	3.77 ± 0.14	6.99 ± 0.86	0.28 ± 0.11	**3.3**
	MS-HiFi-GAN	0.39	$\mathbf{4.17 \pm 0.12}$	7.07 ± 0.93	0.27 ± 0.12	3.4
	FC-HiFi-GAN	0.39	4.08 ± 0.11	7.29 ± 0.98	0.29 ± 0.12	3.4
	MS-iSTFT-HiFi-GAN	0.17	4.05 ± 0.11	7.06 ± 0.89	0.28 ± 0.12	3.4
	MS-FC-HiFi-GAN	**0.16**	4.09 ± 0.11	$\mathbf{6.93 \pm 0.87}$	0.28 ± 0.12	**3.3**
	ground truth	N/A	4.20 ± 0.12	N/A	N/A	3.0

sented in Tables 2 and 3. Additionally, Figure 8 shows the results of the T-test for the MOS tests in Tables 2 and 3. First, JETS-based models significantly realized higher quality synthesis than VITS-based models and outperformed VITS-based models in terms of the MCD, $\log f_0$ RMSE, and CER for both normal-band and full-band E2E TTS conditions as [8]. Although the proposed JETS-based MS-FC-HiFi-GAN could not realize the highest synthesis quality compared with MS-HiFi-GAN (LJSpeech and full-band Hi-Fi TTS) or HiFi-GAN (normal-band Hi-Fi TTS), it significantly realized higher synthesis quality than the conventional MS-iSTFT-VITS [53] (VITS-based E2E TTS with MS-iSTFT-HiFi-GAN), and realized the fastest inference and lowest CER compared with the other models for both normal and full-band E2E TTS conditions. Additionally, there were significant differences between the MOS values of JETS with FC-HiFi-GAN and JETS with iSTFTNet for both normal-band and full-band conditions, and those of JETS with MC-FC-HiFi-GAN and JETS with MS-iSTFT-HiFi-GAN for LJSpeech corpus.

Fig. 6 shows the STFT spectra of a speech waveform in the test set and intermediate features of JETS-based E2E TTS models with iSTFTNet, FC-HiFi-GAN, MS-iSTFT-HiFi-GAN, and MS-FC-HiFi-GAN, and mel-spectrograms of the original speech waveform and those synthesized by JETS-based E2E TTS models with iSTFTNet, FC-HiFi-GAN, MS-iSTFT-HiFi-GAN, and MS-FC-HiFi-GAN. The intermediate features of iSTFTNet are also completely different from the STFT spectra of the original speech waveform as shown in Fig. 4 in analysis-synthesis condition. As the intermediate features of iSTFTNet, those of MS-iSTFT-HiFi-GAN have the same tendency. Compared with the intermediate features of iSTFTNet and MS-iSTFT-HiFi-GAN, those of the proposed FC-HiFi-GAN and MS-FC-HiFi-GAN input to the trainable FC layer-based fast upsampling layer are optimally trained to synthesize high-fidelity speech waveforms as shown in Fig. 6(c) and (e). Additionally, the harmonic structures of the proposed (c) FC-HiFi-GAN and (e) MS-FC-HiFi-GAN are clearer than those of (b) iSTFTNet and (d) MS-iSTFT-HiFi-GAN as showin in the red and green boxes in Fig. 6.

Consequently, the proposed FC-HiFi-GAN and MS-FC-HiFi-GAN with trainable FC layer-based fast upsampling layer without overlap-add operation can realize slightly faster inference and significantly improve the synthesis quality for JETS-based E2E TTS than iSTFTNet and MS-iSTFT-HiFi-GAN with iSTFT layer-based upsampling using fixed weights and overlap-add operation. Therefore, the iSTFT layer-based upsampling can be replaced by the proposed FC layer-based upsampling in HiFi-GAN-based neural vocoders. The summary of the results of the experiments are as follows:

- The proposed JETS-based models can significantly improve the synthesis quality with lower CER compared to the VITS-based models.
- JETS with the proposed MS-FC-HiFi-GAN can realize higher MOS values than the conventional MS-iSTFT-VITS in all the conditions.
- In many conditions, the proposed FC-HiFi-GAN can realize higher MOS values than the conventional iSTFTNet.
- The proposed MS-FC-HiFi-GAN can realized significantly higher MOS values than MS-iSTFT-HiFi-GAN in many conditions.

		HiFi-GAN	FC-HiFi-GAN	iSTFTNet	MS-FC-HiFi-GAN	MS-HiFi-GAN	MS-iSTFT-HiFi-GAN	Ground truth	HiFi-GAN	FC-HiFi-GAN	iSTFTNet	MS-FC-HiFi-GAN	MS-HiFi-GAN
JETS	FC-HiFi-GAN	**0.030**											
	iSTFTNet	**0.000**	**0.000**										
	MS-FC-HiFi-GAN	0.137	**0.001**	**0.000**									
	MS-HiFi-GAN	**0.032**	**0.000**	**0.000**	0.485								
	MS-iSTFT-HiFi-GAN	**0.030**	0.823	**0.000**	**0.000**	**0.000**							
	Ground truth	**0.000**	**0.000**	**0.000**	**0.001**	**0.007**	**0.000**						
VITS	HiFi-GAN	**0.000**	**0.000**	0.271	**0.000**	**0.000**	**0.000**	**0.000**					
	FC-HiFi-GAN	**0.000**	**0.000**	0.683	**0.000**	**0.000**	**0.000**	**0.000**	0.444				
	iSTFTNet	**0.000**	**0.000**	0.478	**0.000**	**0.000**	**0.000**	**0.000**	0.637	0.794			
	MS-FC-HiFi-GAN	**0.000**	**0.003**	**0.023**	**0.000**	**0.000**	**0.001**	**0.000**	0.119	**0.049**	**0.047**		
	MS-HiFi-GAN	**0.000**	**0.008**	**0.004**	**0.000**	**0.000**	**0.002**	**0.000**	**0.041**	**0.010**	**0.015**	0.655	
	MS-iSTFT-HiFi-GAN	**0.000**	**0.000**	0.708	**0.000**	**0.000**	**0.000**	**0.000**	0.388	0.948	0.749	**0.026**	**0.006**

(a) LJSpeech (22.05 kHz)

		HiFi-GAN	FC-HiFi-GAN	iSTFTNet	MS-FC-HiFi-GAN	MS-HiFi-GAN	MS-iSTFT-HiFi-GAN	Ground truth	HiFi-GAN	FC-HiFi-GAN	iSTFTNet	MS-FC-HiFi-GAN	MS-HiFi-GAN
JETS	FC-HiFi-GAN	**0.000**											
	iSTFTNet	**0.000**	**0.038**										
	MS-FC-HiFi-GAN	**0.000**	0.077	**0.000**									
	MS-HiFi-GAN	**0.000**	0.092	**0.001**	0.872								
	MS-iSTFT-HiFi-GAN	**0.000**	**0.022**	**0.000**	0.543	0.439							
	Ground truth	0.487	**0.000**	**0.000**	**0.003**	**0.002**	**0.022**						
VITS	HiFi-GAN	**0.000**	0.275	0.430	**0.009**	**0.024**	**0.003**	**0.000**					
	FC-HiFi-GAN	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**				
	iSTFTNet	**0.000**	**0.000**	**0.002**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.056			
	MS-FC-HiFi-GAN	**0.000**	**0.000**	**0.008**	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**	**0.039**	0.663		
	MS-HiFi-GAN	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.142	0.725	0.476	
	MS-iSTFT-HiFi-GAN	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.846	**0.022**	**0.010**	**0.043**

(b) Hi-Fi TTS [Reader ID: 92 (female)] (24 kHz)

		HiFi-GAN	FC-HiFi-GAN	iSTFTNet	MS-FC-HiFi-GAN	MS-HiFi-GAN	MS-iSTFT-HiFi-GAN	Ground truth	HiFi-GAN	FC-HiFi-GAN	iSTFTNet	MS-FC-HiFi-GAN	MS-HiFi-GAN
JETS	FC-HiFi-GAN	0.310											
	iSTFTNet	**0.000**	**0.000**										
	MS-FC-HiFi-GAN	0.278	0.801	**0.000**									
	MS-HiFi-GAN	**0.022**	0.155	**0.000**	0.204								
	MS-iSTFT-HiFi-GAN	0.611	0.658	**0.000**	0.515	0.072							
	Ground truth	**0.013**	0.067	**0.000**	0.116	0.631	**0.030**						
VITS	HiFi-GAN	**0.000**	**0.000**	**0.021**	**0.000**	**0.000**	**0.000**	**0.000**					
	FC-HiFi-GAN	**0.000**	**0.000**	**0.008**	**0.000**	**0.000**	**0.000**	**0.000**	0.678				
	iSTFTNet	**0.002**	**0.000**	0.844	**0.000**	**0.000**	**0.000**	**0.000**	**0.008**	**0.001**			
	MS-FC-HiFi-GAN	**0.000**	**0.000**	**0.006**	**0.000**	**0.000**	**0.000**	**0.000**	0.782	0.893	**0.004**		
	MS-HiFi-GAN	**0.000**	**0.000**	0.171	**0.000**	**0.000**	**0.000**	**0.000**	0.351	0.166	0.125	0.237	
	MS-iSTFT-HiFi-GAN	**0.000**	**0.000**	0.066	**0.000**	**0.000**	**0.000**	**0.000**	0.481	0.311	0.067	0.306	0.760

(c) Hi-Fi TTS [Reader ID: 92 (female)] (44.1 kHz)

FIGURE 8. Results of T-test for MOS tests in Tables 2 and 3. Values for $p < 0.05$ (statistically significant) are bold with yellow highlighting.

VI. CONCLUSION

HiFi-GAN is widely used not only for TTS but also for many speech and audio applications. Although iSTFTNet and MS-iSTFT-HiFi-GAN have been proposed to accelerate HiFi-GAN while maintaining the synthesis quality, we first pointed out that the predicted intermediate features input to the iSTFT layer are completely different from the original STFT spectra due to the redundancy of the overlap-add operation in iSTFT. To further improve the synthesis quality and inference speed of HiFi-GAN based neural vocoder, we proposed FC-HiFi-GAN and MS-FC-HiFi-GAN by introducing trainable FC layer-based fast upsampling without overlap-add operation instead of the iSTFT layer. The results of experiments for unseen speaker synthesis with multi-speaker models and E2E TTS with VITS- and JETS-based normal-band and full-band models demonstrated that the proposed methods with trainable FC layer-based fast upsampling without overlap-add operation can slightly accelerate the inference speed and significantly improve the synthesis quality in JETS-based E2E TTS than iSTFTNet and MS-iSTFT-HiFi-GAN with iSTFT-based upsampling using fixed weights based on the Fourier basis and overlap-add

operation. Consequently, the iSTFT layer-based upsampling can be replaced by the proposed FC layer-based upsampling in HiFi-GAN-based neural vocoders.

ACKNOWLEDGMENT

This work was performed while Haruki Yamashita was interning with NICT.

REFERENCES

[1] J. Shen, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4779–4783.

[2] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, Aug. 2017, pp. 1118–1122.

[3] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *Proc. ICLR*, May 2021, pp. 1–23.

[4] 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, pp. 1–15.

[5] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 5679–5683.

[6] 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.

[7] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5530–5540.

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

[9] B. Nguyen, F. Cardinaux, and S. Uhlich, "Autotts: End-to-end text-to-speech synthesis through differentiable duration modeling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[10] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17022–17033.

[11] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 8067–8077.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, nos. 3–4, pp. 187–207, Apr. 1999.

[13] 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. 99, no. 7, pp. 1877–1884, 2016.

[14] 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*, Jul. 2018, pp. 2415–2424.

[15] J.-M. Valin and J. Skoglund, "LPCNET: Improving neural speech synthesis through linear prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5891–5895.

[16] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, D. Su, and D. Yu, "DurIAN: Duration informed attention network for speech synthesis," in *Proc. Interspeech*, Oct. 2020, pp. 2027–2031.

[17] Q. Tian, Z. Zhang, H. Lu, L.-H. Chen, and S. Liu, "FeatherWave: An efficient high-fidelity neural vocoder with multi-band linear prediction," in *Proc. Interspeech*, Oct. 2020, pp. 195–199.

[18] Y. Cui, X. Wang, L. He, and F. K. Soong, "An efficient subband linear prediction for LPCNet-based neural synthesis," in *Proc. Interspeech*, Oct. 2020, pp. 3555–3559.

[19] 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.

[20] P. L. Tobing and T. Toda, "High-fidelity and low-latency universal neural vocoder based on multiband WaveRNN with data-driven linear prediction for discrete waveform modeling," in *Proc. Interspeech*, Aug. 2021, pp. 2217–2221.

[21] A. van den Oord et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, vol. 80, J. G. Dy and A. Krause, Eds. Stockholm, Sweden, Jul. 2018, pp. 3915–3923.

[22] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3617–3621.

[23] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6199–6203.

[24] 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, pp. 1–15.

[25] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *Proc. ICLR*, May 2021, pp. 1–17.

[26] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Noise level limited sub-modeling for diffusion probabilistic vocoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6029–6033.

[27] 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.

[28] 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. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 492–498.

[29] I. J. Goodfellow, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[30] Y. Koizumi, K. Yatabe, H. Zen, and M. Bacchiani, "Wavefit: An iterative and non-autoregressive neural vocoder based on fixed-point iteration," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2023, pp. 884–891.

[31] 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*, Jul. 2022, pp. 2709–2720.

[32] H. Guo, F. Xie, X. Wu, F. K. Soong, and H. Meng, "MSMC-TTS: Multi-stage multi-codebook VQ-VAE based neural TTS," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1811–1824, 2023.

[33] B. Nguyen and F. Cardinaux, "NVC-Net: End-to-end adversarial voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7012–7016.

[34] S. Kovela, R. Valle, A. Dantrey, and B. Catanzaro, "Any-to-any voice conversion with f0 and timbre disentanglement and novel timbre conditioning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[35] Z. Zhang, Y. Zheng, X. Li, and L. Lu, "WeSinger 2: Fully parallel singing voice synthesis via multi-singer conditional adversarial training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[36] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "HIFI++: A unified framework for bandwidth extension and speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[37] 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. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[38] Y.-C. Wu, I. D. Gebru, D. Markovic, and A. Richard, "Audiodec: An open-source streaming high-fidelity neural audio codec," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[39] 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. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[40] I. R. Titze et al., "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *J. Acoust. Soc. Amer.*, vol. 137, no. 5, pp. 3005–3007, May 2015.

[41] R. Yoneyama, Y.-C. Wu, and T. Toda, "Source-filter HiFi-GAN: Fast and pitch controllable high-fidelity neural vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[42] Y. Shirahata, R. Yamamoto, E. Song, R. Terashima, J.-M. Kim, and K. Tachibana, "Period VITS: Variational inference with explicit pitch modeling for end-to-end emotional speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[43] 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, Language Process.*, vol. 31, pp. 1902–1915, 2023.
[44] E. Cohen, F. Kreuk, and J. Keshet, "Speech time-scale modification with GANs," *IEEE Signal Process. Lett.*, vol. 29, pp. 1067–1071, 2022.
[45] E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstoft, "Generative models for sound field reconstruction," *J. Acoust. Soc. Amer.*, vol. 153, no. 2, pp. 1179–1190, Feb. 2023.
[46] 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.
[47] 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.
[48] 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, pp. 1–19.
[49] 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. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[50] T. Okamoto, T. Toda, and H. Kawai, "Multi-stream HiFi-GAN with data-driven waveform decomposition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 610–617.
[51] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "ISTFTNET: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6207–6211.
[52] T. Kaneko, H. Kameoka, K. Tanaka, and S. Seki, "MISRNet: Lightweight neural vocoder using multi-input single shared residual blocks," in *Proc. Interspeech*, Sep. 2022, pp. 1631–1635.
[53] 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. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[54] 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.
[55] T. Q. Nguyen, "Near-perfect-reconstruction pseudo-QMF banks," *IEEE Trans. Signal Process.*, vol. 42, no. 1, pp. 65–76, Jan. 1994.
[56] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. 32nd Int. Conf. Mach. Learn. (PMLR)*, 2015, pp. 1530–1538.
[57] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, Apr. 2024, pp. 1–14.
[58] 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.
[59] R. Badlani, A. Lancucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, "One TTS alignment to rule them all," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6092–6096.
[60] 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. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[61] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoust. Sci. Technol.*, vol. 41, no. 5, pp. 761–768, Sep. 2020.
[62] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NIPS*, Dec. 2019, pp. 8024–8035.
[63] 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," 2021, *arXiv:2110.07840*.
[64] *Methods for Subjective Determination of Transmission Quality*, document ITU-T Recommendation P. 800, 1996.
[65] K. Ito and L. Johnson. (2017). *The LJ Speech Dataset*. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/
[66] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-Fi multi-speaker English TTS dataset," in *Proc. Interspeech*, Aug. 2021, pp. 2776–2780.
[67] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[68] 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. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 5874–5878.

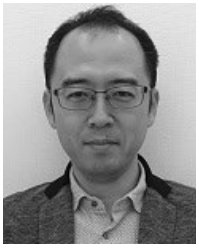
HARUKI YAMASHITA received the B.E. degree from Kobe University, Japan, in 2022, where he is currently pursuing the master's degree. Since 2022, he has been an Internship Student with the National Institute of Information and Communications Technology (NICT), Japan. His research interests include speech synthesis and voice conversion. He is a Student Member of ASJ. He received the 25th Student Presentation Award from the Acoustical Society of Japan (ASJ), in 2022.

TAKUMA OKAMOTO (Member, IEEE) received the B.E., M.S., and Ph.D. degrees from Tohoku University, Japan, in 2004, 2006, and 2009, respectively. Since 2009, he has been a Postdoctoral Research Fellow with Tohoku University. From 2012 to 2020, he was a Researcher with the National Institute of Information and Communications Technology (NICT), Japan, where he is currently a Senior Researcher. His main research interests include sound field synthesis and speech synthesis. He is a member of the Audio Engineering Society (AES) and the Acoustical Society of Japan (ASJ). He received the 32nd Awaya Prize Young Researcher Award, the 57th Sato Prize Paper Award, and the Ninth Society Activity Contribution Award from ASJ, in 2012, 2017, and 2022, respectively.

RYOICHI TAKASHIMA (Member, IEEE) received the B.E., M.E., and Dr.Eng. degrees in computer science from Kobe University, in 2008, 2010, and 2013, respectively. From 2013 to 2018, he was a Researcher with Hitachi Ltd., Tokyo, Japan. From 2016 to 2018, he was on loan to the National Institute of Information and Communication Technology (NICT), Kyoto, Japan. He is currently an Associate Professor with Kobe University. His research interests include machine learning and signal processing. He is a member of ASJ.

YAMATO OHTANI (Member, IEEE) received the B.E. degree in engineering from Osaka University, Japan, in 2005, and the M.E. and D.E. degrees from the Nara Institute of Science and Technology (NAIST), in 2007 and 2010, respectively. He was an Intern Researcher with the ATR Spoken Language Communication Research Laboratory, Kyoto, Japan, from 2006 to 2009. He was an Engineer, from 2010 to 2015, and a Research Scientist, from 2015 to 2017, with Toshiba Corporation, Japan. He was a Senior Researcher, from 2017 to 2019, the Head of the Research and Development Department, from 2019 to 2022, and the Director, from 2021 to 2023, with AI Inc., Japan. He is currently a Senior Researcher with the National Institute of Information and Communications Technology (NICT), Japan. His research interests include statistical approaches to speech signal processing, such as speech synthesis and voice conversion. He received the 39th Awaya Prize Young Researcher Award from the Acoustical Society of Japan (ASJ), in 2015.

TETSUYA TAKIGUCHI (Member, IEEE) received the M.Eng. and Dr.Eng. degrees from the Nara Institute of Science and Technology, Japan, in 1996 and 1999, respectively. From 1999 to 2004, he was a Researcher with the Tokyo Research Laboratory, IBM Research. From 2004 to 2016, he was an Associate Professor with Kobe University, where he has been a Professor, since 2016. From May 2008 to September 2008, he was a Visiting Scholar with the Department of Electrical Engineering, University of Washington. From March 2010 to September 2010, he was a Visiting Scholar with the Institute for Learning and Brain Sciences, University of Washington. From April 2013 to October 2013, he was a Visiting Scholar with Laboratoire d'InfoRmatique en Image et Systémes d'information, INSA Lyon. His research interests include speech, image, and brain processing, and multimodal assistive technologies for people with articulation disorders. He received the Best Paper Award from IEEE ICME 2008.



TOMOKI TODA (Senior Member, IEEE) received the B.E. degree from Nagoya University, Japan, in 1999, and the M.E. and D.E. degrees from the Nara Institute of Science and Technology (NAIST), Japan, in 2001 and 2003, respectively. He was an Assistant Professor, from 2005 to 2011, and an Associate Professor, from 2011 to 2015, with NAIST. Since 2015, he has been a Professor with the Information Technology Center, Nagoya University. From 2003 to 2005, he was a Research Fellow of the Japan Society for the Promotion of Science. His research interests include statistical approaches to speech and audio processing. He received more than ten article/achievement awards, including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (*Speech Communication*).

HISASHI KAWAI (Member, IEEE) received the B.E., M.E., and D.E. degrees in electronic engineering from The University of Tokyo, in 1984, 1986, and 1989, respectively. In 1989, he joined Kokusai Denshin Denwa Company Ltd. From 2000 to 2004, he was with ATR Spoken Language Translation Research Laboratories, where he engaged in the development of text-to-speech synthesis system. From October 2004 to March 2009 and from April 2012 to September 2014, he was with KDDI Research and Development Laboratories, where he was engaged in the research and development of speech information processing, speech quality control for telephone, speech signal processing, acoustic signal processing, and communication robots. From April 2009 to March 2012 and since October 2014, he has been with the National Institute of Information and Communications Technology (NICT), where he is engaged in development of speech technology for spoken language translation. He is a member of the Acoustical Society of Japan (ASJ) and the Institute of Electronics, Information and Communication Engineers (IEICE).