# MULTI-STREAM HIFI-GAN WITH DATA-DRIVEN WAVEFORM DECOMPOSITION

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

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

## ABSTRACT

Although a HiFi-GAN vocoder can synthesize high-fidelity speech waveforms in real time on CPUs, there is a tradeoff between synthesis quality and inference speed. To increase inference speed while maintaining synthesis quality, a multi-band structure is introduced to HiFi-GAN. However, it cannot be trained well because of the strong constraint imposed by the fixed multi-band structure. As an alternative approach, Multi-stream MelGAN and HiFi-GAN are proposed, in which the fixed synthesis filter in Multi-band MelGAN is replaced by a trainable convolutional layer with the same structure. In contrast to Multi-band MelGAN, the proposed methods use the trainable synthesis filter to decompose speech waveforms in a data-driven manner. To evaluate the proposed Multi-stream HiFi-GAN as an entire real-time neural text-to-speech system on CPUs, a fast acoustic model, based on Parallel Tacotron 2 with forced alignment and accentual label input, was implemented. The results of experiments—using Japanese male, female, and multi-speaker corpora—indicate that Multi-stream HiFi-GAN can increase synthesis speed while improving or maintaining synthesis quality in analysis–synthesis and text-to-speech conditions for single-speaker models and unseen speaker synthesis for multi-speaker models, compared with the original HiFi-GAN.

***Index Terms***— Speech synthesis, neural vocoder, HiFi-GAN, data-driven waveform decomposition, Parallel Tacotron 2

## 1. INTRODUCTION

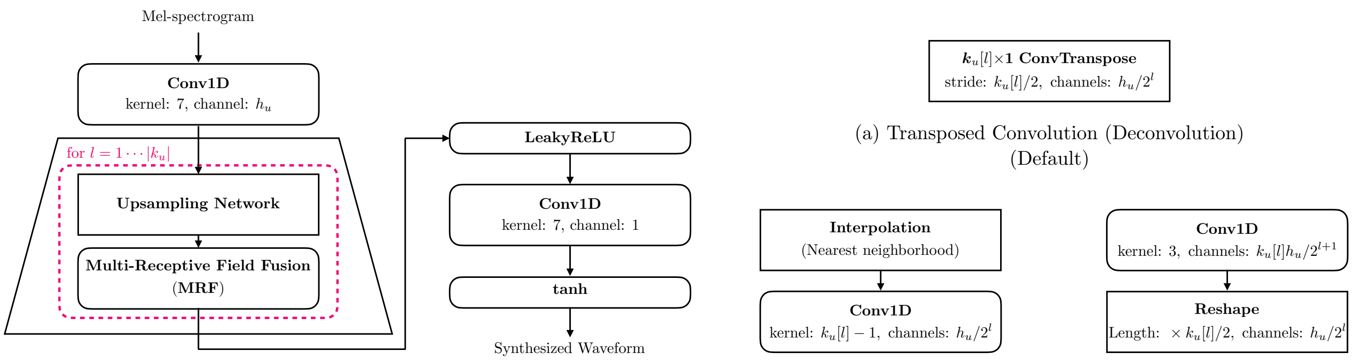
Recent advances in neural speech synthesis have made it possible to synthesize high-fidelity speech waveforms with the same quality as natural human speech, using Tacotron 2 [1] combined with the autoregressive WaveNet-based [2] neural vocoder [3]. Additionally, entire end-to-end text-to-speech (TTS) models, which can directly synthesize speech waveforms from character or phoneme sequences with a single neural network, have also been investigated, such as EATS [4], FastSpeech 2+ [5], Wave-Tacotron [6], VITS [7], and Reinforce-Aligner [8].

To realize high-fidelity and real-time neural speech synthesis, many types of real-time neural vocoders, based on both autoregressive and non-autoregressive structures, have been investigated. Compared with real-time autoregressive models, such as WaveRNN [9], LPCNet [10], DurIAN [11], FeatherWave [12], Subband-LPCNet [13], Fullband-LPCNet [14], and MWDLP [15], non-autoregressive models, which can simultaneously synthesize all waveform samples, can be more easily implemented, and many models have been proposed. Non-autoregressive models are broadly classified into the following three types. The first type comprises flow-based approaches [16], such as Parallel WaveNet [17, 18] and WaveGlow [19]. The second type comprises generative adversarial network (GAN)-based models [20], such as WaveGAN [21], MelGAN [22], Parallel WaveGAN [23], GAN-TTS [24], VocGAN [25], HiFi-GAN [26], Multi-band MelGAN [27], Quasi-Periodic Parallel WaveGAN [28], Fre-GAN [29], Glow-WaveGAN [30], UnivNet [31], and Basis-MelGAN [32]. The final type comprises diffusion probabilistic-based models [33], such as WaveGrad [34–36] and DiffWave [35, 37]. Although these models can synthesize high-fidelity speech waveforms, most of them require a GPU for real-time inference. However, for actual implementations, the development of real-time neural vocoders on CPUs is important.

MelGAN and HiFi-GAN are GAN-based non-autoregressive neural vocoders that can realize real-time inference on CPUs. In particular, HiFi-GAN can realize higher-fidelity synthesis than MelGAN, using sophisticated generator and discriminators. To further improve the synthesis quality, Fre-GAN, with modified generator and discriminators [29], was recently proposed. Additionally, VITS [7], a HiFi-GAN-based end-to-end neural TTS combined with Glow-TTS [38], was recently proposed. However, there is a tradeoff between the synthesis quality and inference speed in HiFi-GAN. To increase the inference speed and improve the synthesis quality of MelGAN, Multi-band MelGAN was proposed [27], in which multi-band waveforms are synthesized by a MelGAN generator and integrated to a fullband waveform, using multi-rate signal processing [39] instead of neural-network-based upsampling.

To increase the inference speed of HiFi-GAN while maintaining the synthesis quality, we first simply introduce a multi-band structure into HiFi-GAN, as Multi-band MelGAN [27]. However, this cannot be trained well because of the strong constraint imposed by the fixed multi-band structure. As a simple but effective alternative approach, we then propose Multi-stream MelGAN and HiFi-GAN. The fixed synthesis filter in Multi-band MelGAN is realized by a mixture of FIR filters, which is equivalent to a layer of a convolutional neural network (CNN). In the proposed methods, the fixed synthesis filter in Multi-band MelGAN is replaced by a trainable CNN layer without bias. In contrast to Multi-band MelGAN, the MelGAN and HiFi-GAN generators in the proposed methods use the trainable synthesis filter to synthesize multi-stream waveforms decomposed in a data-driven manner, to optimally synthesize the final output waveforms. By introducing the proposed trainable filter, both Multi-stream MelGAN and HiFi-GAN can be trained well using the same discriminators and loss functions as used for the original Multi-band MelGAN and HiFi-GAN without multi-band waveforms.

A similar approach, Basis-MelGAN [32], was recently proposed, in which speech waveforms are decomposed with a trainable basis and their associated weights by Conv-TasNet [40], and the associated weights are predicted by inference, to simplify the upsampling layers. As a result, the inference speed can be increased while realizing high-fidelity synthesis. Compared with Basis-MelGAN, the proposed methods are much simpler and no pre-training of Conv-TasNet is required. An alternative GAN-based vocoder, UnivNet,




(A) HiFi-GAN Generator Network; (a) Transposed Convolution (Deconvolution) (Default); (b) Interpolation + Convolution; (c) Sub-pixel Convolution

**Fig. 1**. (A) HiFi-GAN generator network and three types of upsampling layers: (a) transposed convolution used in the original HiFi-GAN [26], (b) interpolation and convolution, and (c) sub-pixel convolution. $k_u$, $l$, and $h_u$ are the kernel size, number of upsampling layers, and number of hidden channels, respectively.

was also recently proposed, which can realize higher-quality and faster synthesis than HiFi-GAN [31]. However, only its inference speed on a GPU was measured; speed on CPUs was not investigated.

To evaluate the proposed Multi-stream HiFi-GAN as an entire real-time neural TTS system on CPUs, a non-autoregressive fast acoustic model—based on Parallel Tacotron 2 [41] with forced alignment and accentual label input—was implemented. This is also important because HiFi-GAN was only evaluated with autoregressive Tacotron 2 [1] in [26]. The results of experiments, reported in Section 5, indicate that Multi-stream HiFi-GAN can increase the synthesis speed while improving or maintaining the synthesis quality in the analysis–synthesis and TTS conditions (for single-speaker models) and unseen speaker synthesis (for multi-speaker models), compared with the original HiFi-GAN.

## 2. HIFI-GAN AND MULTI-BAND MELGAN VOCODERS

### 2.1. HiFi-GAN

As depicted in Fig. 1(A), HiFi-GAN [26] converts input mel-spectrograms to speech waveforms with multiple upsampling layers, without white noise input. HiFi-GAN uses a multi-period discriminator for modeling periodic patterns and a multi-scale discriminator for capturing consecutive patterns and long-term dependencies, in the same manner as MelGAN [22]. As a consequence of the sophisticated generator and discriminators, HiFi-GAN can synthesize high-fidelity speech waveforms in real time on CPUs. In a large model as HiFi-GAN V1, the initial number of hidden channels $h_n$ is 512 and the kernel sizes of the transposed convolution layers (Fig. 1(a)) are $k_u = [16, 16, 4, 4]$. For a small model as HiFi-GAN V2, $h_n$ is 128 and the other parameters are the same as those of HiFi-GAN V1. Although HiFi-GAN V2 realizes high-quality synthesis with fast inference on CPUs [26], the synthesis quality of HiFi-GAN V2 is lower than that of HiFi-GAN V1, and there is a tradeoff between the synthesis quality and inference speed, according to the results of experiments reported in Section 5.

### 2.2. Investigation of upsampling layers in HiFi-GAN

In [42], three types of upsampling layers for neural audio synthesis were introduced and compared. In contrast to the original HiFi-GAN, which uses transposed convolution layers (Fig. 1(a)) for upsampling, this paper investigates interpolation-based upsampling layers [43] (Fig. 1(b)), as used in [44], and sub-pixel convolution (pixel shuffler) layers [45], as used in [46, 47]. These upsamplers are compared with transposed convolution layers (Fig. 1(a)), with respect to synthesis accuracy and inference speed, in Section 5.

### 2.3. Multi-band MelGAN

To increase the inference speed and improve the synthesis quality of MelGAN [22], Multi-band MelGAN [27] was proposed [27], in which four subband waveforms are synthesized by a MelGAN generator and integrated to a fullband waveform by a synthesis filter bank based on multi-rate signal processing [39][1], instead of neural-network-based upsampling (Fig. 2(a)). In Multi-band MelGAN, Pseudo-Quadrature Mirror Filter Bank [51] is used to calculate the analysis and synthesis filter banks [11–13, 15]. In the training, subband waveforms are calculated from target fullband waveforms by the analysis filter bank and decimation-based downsampling, and both the fullband and subband STFT losses are used to update model parameters. Because zero-padding-based upsampling with a synthesis filter bank is much faster than neural-network-based upsampling, but still effective, Multi-band MelGAN can realize high-fidelity synthesis faster than the original MelGAN [27].

### 2.4. Investigation of Multi-band HiFi-GAN

Following the success of Multi-band MelGAN, a multi-band structure was first applied to HiFi-GAN to increase the inference speed while maintaining the synthesis quality. In Multi-band HiFi-GAN, $k_u$ in the HiFi-GAN V1 generator is $[16, 16]$ and the number of output channels in the final CNN is 4. Four subband waveforms are synthesized by the HiFi-GAN generator and integrated to a fullband waveform by zero-padding-based upsampling with a synthesis filter, as in Multi-band MelGAN, in which the length of the synthesis filter bank is 63 [11, 27]. However, in preliminary experiments, it could not be trained well, despite the use of pre-training using both the fullband and subband STFT losses, as in Multi-band MelGAN. The quality of speech synthesized by Multi-band HiFi-GAN was lower than that of speech synthesized by the HiFi-GAN V2 and V3 models [26]. This may be because the constraint imposed by the fixed

[1] Multi-rate signal processing was first introduced to autoregressive neural vocoders by the authors, as Subband WaveNet and FFTNet [48–50].

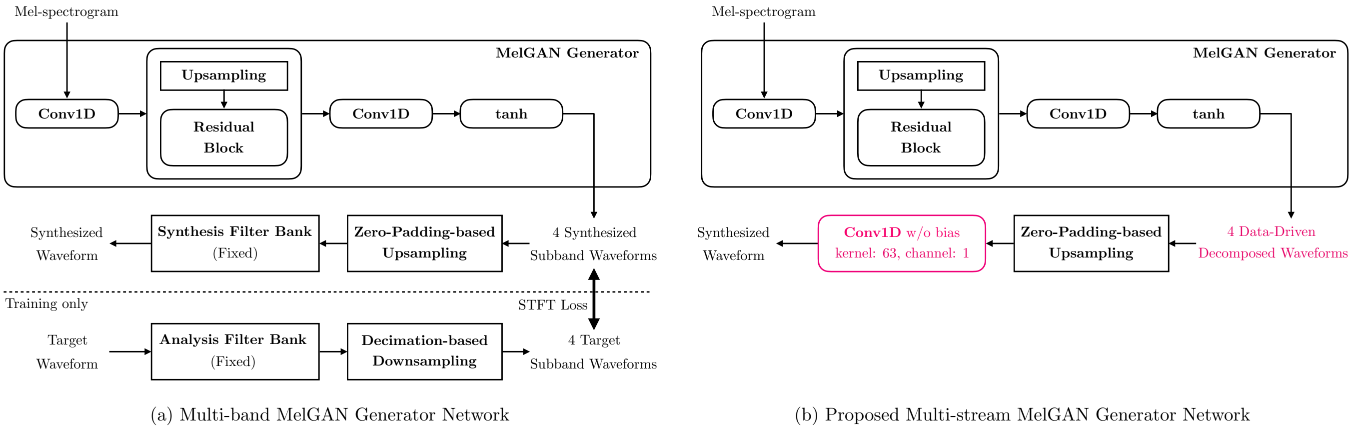



**Fig. 2**. (a) Multi-band MelGAN generator network and (b) proposed Multi-stream MelGAN generator network.

multi-band structure is too strong for HiFi-GAN, and the sophisticated HiFi-GAN discriminators can easily distinguish between real and synthetic speech.

## 3. MULTI-STREAM MELGAN AND HIFI-GAN

### 3.1. Multi-stream MelGAN

This subsection describes the proposed Multi-stream MelGAN, to ease the understanding of Multi-stream HiFi-GAN, which is introduced in the next subsection. The fixed synthesis filter bank in Multi-band MelGAN is realized by a mixture of FIR filters, which is equivalent to a CNN layer in a neural network. In Multi-stream MelGAN, the fixed synthesis filter in Multi-band MelGAN is simply replaced by a trainable CNN layer without bias, with the same number of channels as the fixed synthesis filter (Fig. 2(b)). The Multi-stream MelGAN generator can also be successfully trained with only the fullband STFT losses, without subband waveforms. Because the only difference between Multi-stream MelGAN and Multi-band MelGAN is the trainability of the final CNN layer, their inference speeds are the same. Therefore, Multi-stream MelGAN is simpler than Multi-band MelGAN because no analysis or synthesis filter banks are required. A similar approach, in which fixed frontend filters are replaced by trainable filters, was also recently investigated for audio classification [52].

### 3.2. Multi-stream HiFi-GAN

The proposed Multi-stream HiFi-GAN (Fig. 3) is described in this subsection. Because of the higher inference speed, Multi-stream HiFi-GAN introduces sub-pixel CNN layers for upsampling. As in Multi-stream MelGAN, the final fourfold upsampling is realized by zero-padding-based upsampling with a trainable CNN layer without bias. Compared with Multi-band HiFi-GAN, which cannot be trained well (as explained in Section 2.4), Multi-stream HiFi-GAN can be successfully trained with the same discriminators and loss functions as the original HiFi-GAN. This is because the final CNN layer is trainable without constraint and it can also be optimally and jointly trained with the HiFi-GAN generator network, which outputs four-stream waveforms.

In contrast to Multi-band MelGAN, the MelGAN and HiFi-GAN generators in the proposed methods synthesize four-stream waveforms decomposed in a data-driven manner by the trainable

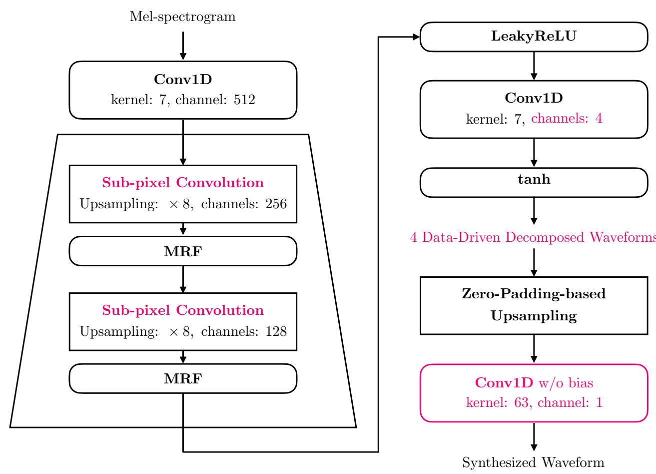



**Fig. 3**. Proposed Multi-stream HiFi-GAN generator network.

CNN layer, to optimally synthesize the final output speech waveforms. Therefore, Multi-stream HiFi-GAN is expected to increase the inference speed while maintaining the synthesis quality.

## 4. FAST ACOUSTIC MODEL BASED ON PARALLEL TACOTRON 2 WITH FORCED ALIGNMENT

To evaluate the proposed Multi-stream HiFi-GAN vocoder as an entire real-time neural TTS system on CPUs, a non-autoregressive fast acoustic model, based on Parallel Tacotron 2 [41], was designed. Parallel Tacotron 2 is a state-of-the-art non-autoregressive acoustic model for multi-speaker neural TTS; it is an extension of Parallel Tacotron [53] using lightweight convolutions (LConv) [54]. Although Parallel Tacotron requires phoneme alignment, obtained from an external model, Parallel Tacotron 2 introduced a trainable upsampling layer and soft-DTW [55] to jointly optimize output mel-spectrograms and phoneme alignment without forced alignment.

Although phoneme alignment could be successfully trained in single-speaker Parallel Tacotron 2 with soft-DTW[2] by using single-speaker corpora, as explained in Section 5, output mel-spectrograms

[2]https://github.com/Maghoumi/pytorch-softdtw-cuda

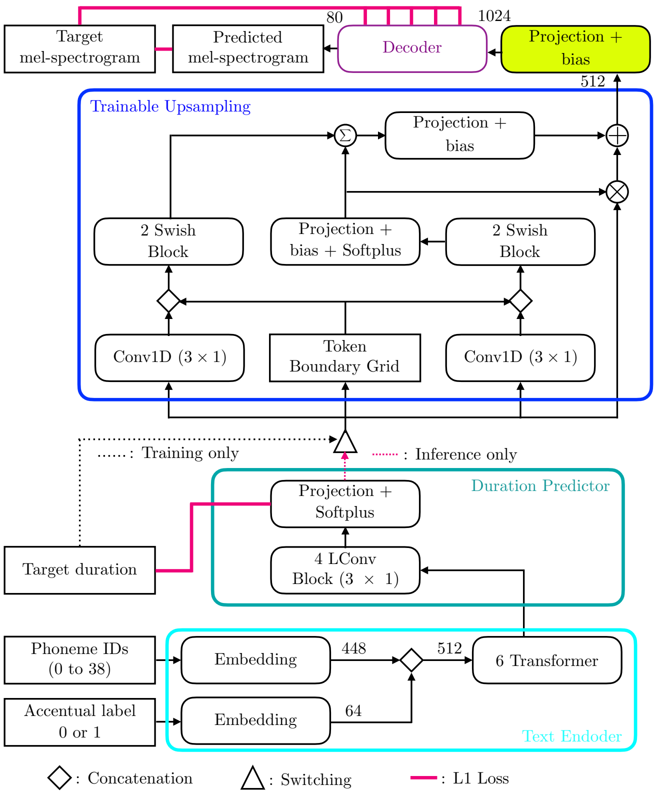



**Fig. 4**. Single-speaker Parallel Tacotron 2 with forced alignment and accentual label input for pitch accent languages.

were degraded and high-fidelity synthesis could not be realized[3]. Therefore, in this study, forced alignment—calculated by Montreal Forced Aligner [56], as used in FastSpeech 2 [5]—was performed, and single-speaker Parallel Tacotron 2 with forced alignment was implemented, as shown in Fig. 4. In this model, both phoneme sequences and accentual label sequences are input to the network for pitch accent languages [57–59]. The decoder shown in Fig. 4 is also constructed from six LConv blocks (17 × 1) and each block outputs mel-spectrograms to calculate the $L_1$ losses, in the same manner as [41, 53]. Although the number of channels used in the network was not described for either Parallel Tacotron [53] or Parallel Tacotron 2 [41], it was set to 512 in this study. However, the results of preliminary experiments suggested that it is better to introduce 1,024 channels in the decoder. Therefore, to increase the number of channels, an additional projection layer was introduced between the trainable upsampling layer and the decoder. Eight attention heads are used in all the feedforward Transformer and LConv blocks. The loss function for training is defined as:

$$\mathcal{L} = \frac{1}{6KT}\sum_{i=1}^{6}\mathcal{L}_{\mathrm{spec}_i} + \frac{1}{N}\lambda\mathcal{L}_{\mathrm{dur}}, \quad (1)$$

where $\mathcal{L}_{\mathrm{spec}_i}$ is the $L_1$ mel-spectrogram loss for the $i$-th LConv block output, $\mathcal{L}_{\mathrm{dur}}$ is the $L_1$ duration loss, $\lambda$ is the weight coefficient, $K$ is the size of the mel-spectrogram, $T$ is the number of frames, and $N$ is the number of phonemes. In this study, $K$ and $\lambda$ were set to 80 and 1, respectively. In contrast to the model trained with soft-DTW, the model with forced alignment using the loss function $\mathcal{L}$ was successfully trained and high-fidelity synthesis could be achieved.

## 5. EXPERIMENTS

### 5.1. Experimental conditions

Experiments were conducted to evaluate the proposed Multi-stream MelGAN and HiFi-GAN and compare them with Multi-band MelGAN and the original HiFi-GAN. All the neural network models were implemented by PyTorch and trained using NVIDIA Tesla V100 GPUs. Both the analysis–synthesis and TTS conditions were evaluated for single-speaker models. Additionally, analysis–synthesis with unseen speaker features was evaluated for multi-speaker models, as in [26].

**Speech corpora:**

The experiments were conducted with Japanese female and male speech corpora of professional speakers (JAF001 and JAM017) for single-speaker models, and the JVS corpus [60] for multi-speaker models, with a sampling frequency of 24 kHz. The JAF001, JAM017, and JVS corpora included 19,644 (21.8 hours), 19,584 (20.7 hours), and 12,737 (25.7 hours; JVS001 and JVS004 were not included as the validation set) utterances, respectively, for the training set[4]. One hundred utterances from each of JAF001 and JAM017 were used for the test set. Because the JAF001 and JAM017 speakers were not included in the JVS corpus, unseen speaker synthesis could be evaluated for multi-speaker models and compared with single-speaker models. As in [26], band-limited 80-dimensional mel-spectrograms were analyzed. The FFT, window, and hop sizes were 1024, 1024, and 256, respectively. These were used in all the neural vocoders and acoustic models for TTS.

**Multi-band and Multi-stream MelGAN vocoders:**

Multi-band MelGAN used the network structures (of the generator and discriminator) and training condition reported in [27]. Both the fullband and subband STFT losses were used for the generator. In Multi-stream MelGAN, the fixed synthesis filter of Multi-band MelGAN was replaced by a trainable CNN layer without bias (Fig. 2(b)), and only the fullband STFT losse was used for the generator. These models were implemented by using an unofficial implementation[5] with simple modifications, and the Adam optimizer [61] was used to update parameters. The number of parameter updates was 2M.

**HiFi-GAN and Multi-stream HiFi-GAN vocoders:**

As the baselines, the original HiFi-GAN V1 (a) and V2 (a) models from [26], with transposed convolution-based upsampling (Fig. 1(a)), were introduced; the upsampling rates were $[8, 8, 2, 2]$. As described in Section 2.2, to compare the upsampling methods, the HiFi-GAN V1 (b) model with interpolation-based upsampling (Fig. 1(b))—with CNN kernel sizes of $[15, 15, 3, 3]$—and the HiFi-GAN V1 (c) and V2 (c) models with sub-pixel convolution-based upsampling (Fig. 1(c)) were also investigated. In Multi-stream HiFi-GAN, the upsampling rates of the HiFi-GAN network were $[8, 8]$ and the final fourfold upsampling was realized by zero-padding-based upsampling and a trainable CNN layer without bias (Fig. 3).

[3] Compared to the original model with 405 hours of multi-speaker speech data and a batch size of 2,048 [41], only about 20 hours of single-speaker speech data and a batch size of 128 were used in preliminary experiments. Additionally, there are some hyperparameters, such as the warp penalty in soft-DTW and the weight coefficients of loss functions. Therefore, further work is required to successfully train Parallel Tacotron 2 with soft-DTW.

[4] Japanese speech corpora JAF001 and JAM017 will be released by NICT for open innovation in speech synthesis research. Therefore, all the corpora used in the experiments will be available.

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

**Table 1**. Number of model parameters (#param) and real-time factor (RTF) in inference using an NVIDIA Tesla V100 GPU and Intel Xeon 6152 CPU with 16 cores. "MS" is Multi-stream. (a), (b), and (c) are the types of upsampling layers depicted in Fig. 1.

| Model | #param | RTF (1GPU) | RTF (CPU) |
|---|---|---|---|
| MS MelGAN [27] | 2.54M | 0.0033 | 0.034 |
| HiFi-GAN V1 (a) [26] | 13.9M | 0.011 | 0.095 |
| HiFi-GAN V2 (a) [26] | 0.93M | 0.0079 | 0.050 |
| HiFi-GAN V1 (b) | 13.8M | 0.013 | 0.094 |
| HiFi-GAN V1 (c) | 15.3M | 0.011 | 0.084 |
| HiFi-GAN V2 (c) | 1.01M | 0.0082 | 0.049 |
| MS HiFi-GAN | 14.6M | 0.0067 | 0.050 |
| Sub-DiffWave [35] | 14.3M | 0.31 | 10.25 |
| Transformer | 53.0M | 0.55 | 3.2 |
| Parallel Tacotron 2 | 99.7M | 0.0044 | 0.020 |

The kernel size in the sub-pixel convolution layers was 3, as in [42]. The other parameters and discriminators were the same as in the original HiFi-GAN V1 (a) and V2 (a), for all the models. All the models were implemented by an official implementation[6] with simple modifications, and the AdamW optimizer [62], with the same learning rate and schedule as in [26], was used. The number of parameter updates was 1M. No fine-tuning with mel-spectrograms predicted by Parallel Tacotron 2 was performed for any of the models. In all the MelGAN-based and HiFi-GAN-based models, the batch size and batch length were 16 and 8,192, respectively.

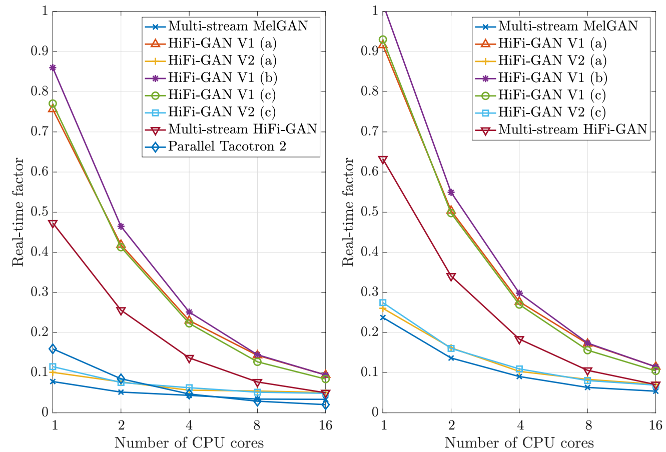
**DiffWave vocoder with noise-level-limited sub-modeling**

To compare MelGAN-based and HiFi-GAN-based models with other non-autoregressive neural vocoders, DiffWave [37]—conditioned on continuous noise level, as in WaveGrad [34] with noise-level-limited sub-modeling [35]—was used. As in [34, 35], the noise schedule for inference was Fibonacci-based 25 iterations. To efficiently use all 10 sub-models, the noise level range for training with a logarithmic scale was divided into 10 equal parts for the 10 sub-models, although only six sub-models were used for Fibonacci-based 25 iterations in the previous division criterion, with a linear scale [35]. The network structure and training condition were the same as those of [35], and only the upsampling rates in the acoustic feature conditioning network were changed, from [30, 10] to [16, 16]. The spectral enhancement postfilter used in [35] was not used in the experiments. The DiffWave model was implemented using an unofficial implementation[7] with some modifications for continuous noise level conditioning, as in WaveGrad [34][8].

**Parallel Tacotron 2 with forced alignment:**

Parallel Tacotron 2, a real-time neural TTS system for Japanese on CPUs, with forced alignment and accentual label input, was implemented (Fig. 4). The phoneme sequence and accentual label sequence were obtained by a text analyzer developed in NICT [57–59], and they were embedded and concatenated, to form 512-dimensional features. The feedforward Transformer and LConv blocks were implemented using ESPnet-TTS [63] and an official implementation[9], respectively. The loss function for training was the function defined in (1). The numbers of parameter updates for JAF001 and JAM017 were 500K and 650K, respectively. Additionally, Parallel Tacotron 2 models with simple duplication-based upsampling (used in [5]) and Gaussian upsampling (used in [4, 8, 36]) were investigated, instead of the trainable upsampling. The final training loss scores of trainable upsampling, duplication-based upsampling, and Gaussian upsampling were approximately 0.49, 0.51, and 0.50 for JAF001, and 0.48, 0.49, and 0.49 for JAM017, respectively. Therefore, the trainable upsampling was also effective for the models trained by using forced alignment.

[6] https://github.com/jik876/hifi-gan
[7] https://github.com/lmnt-com/diffwave
[8] https://github.com/lmnt-com/wavegrad
[9] https://github.com/pytorch/fairseq




(A) Real-time neural vocoders and acoustic model (B) Entire TTS system with Parallel Tacotron 2

**Fig. 5**. Results of real-time factors for different numbers of CPU cores on Intel Xeon 6152 for (A) real-time neural vocoders and acoustic model, and (B) entire text-to-speech (TTS) system with Parallel Tacotron 2. (a), (b), and (c) are the types of upsampling layers depicted in Fig. 1.

**Transformer-based acoustic model with accentual label input:**

To compare Parallel Tacotron 2 with an autoregressive acoustic model, a Transformer-based acoustic model [64] was used. The model structure was the same as that used in [59]. The first CNN layer in [59] was replaced by two embedding layers used in Parallel Tacotron 2 with accentual label input. The model was also implemented using ESPnet-TTS [63]. The number of parameter updates for both JAF001 and JAM017 was 500K. In the Parallel Tacotron 2 and Transformer models, the RAdam optimizer [65] was used with a learning rate of 0.0001, and the batch size was 32.

### 5.2. Real-time factor evaluation

The real-time factors (RTFs) of all the models for inference were measured by executing them on a NVIDIA Tesla V100 GPU and Intel Xeon 6152 CPU, where the number of CPU cores was increased from 1 to 16, as in [66]. The numbers of model parameters and the results of RTFs using a GPU and 16 CPU cores are shown in Table 1. Additionally, the results of RTFs with different numbers of CPU cores are plotted in Fig. 5. Multi-stream HiFi-GAN successfully increased the inference speed, compared with all the HiFi-GAN V1 models, even though the model size of Multi-stream HiFi-GAN introducing sub-pixel CNN layers was slightly larger than the sizes of the HiFi-GAN V1 (a) and (b) models. Multi-stream HiFi-GAN was also faster than the HiFi-GAN V2 models when using a GPU and the same speed as the HiFi-GAN V2 models when using 16 CPU cores. A fast neural TTS system with an RTF of 0.1 using eight CPU cores could then be realized by Multi-stream HiFi-GAN

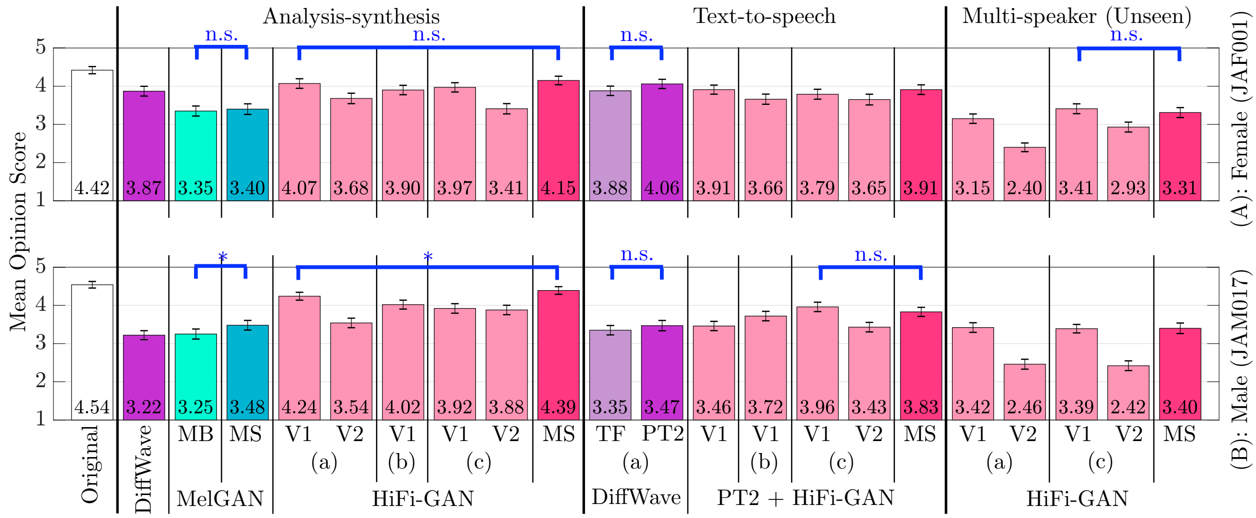



**Fig. 6**. Results of the MOS test for (A) female speaker (JAF001) and (B) male speaker (JAM017) with 14 listening subjects. The confidence level of the error bars is 95%. "MB," "MS," "TF," and "PT2" are Multi-band, Multi-stream, Transformer, and Parallel Tacotron 2, respectively. (a), (b), and (c) are the types of upsampling layers depicted in Fig. 1. In the multi-speaker condition, HiFi-GAN vocoders were trained using the JVS corpus and utterances of unseen speakers (JAF001 and JAM017) were synthesized.

and Parallel Tacotron 2. As discussed in [42], HiFi-GAN V1 (c) was slightly faster than the other HiFi-GAN V1 models for smaller kernel sizes when multiple CPU cores were used, even though the number of model parameters was slightly larger. HiFi-GAN V1 (b) was slightly slower than the other HiFi-GAN V1 models because of its double-layer structure.

### 5.3. Subjective evaluation

To subjectively evaluate the synthesized speech waveforms, mean opinion score (MOS) tests, on a five-point scale [67], were conducted. These were presented through headphones to 14 Japanese adult native speakers without hearing loss. There were 792 utterances: 18 utterances (out of 100 test set utterances) $\times$ 22 conditions $\times$ 2 (JAF001 and JAM017), as shown in Fig. 6, including the original test set waveforms. Some of the speech samples used in the experiments are available online[10]. The results of the MOS tests are plotted in Fig. 6. First, Multi-stream MelGAN and HiFi-GAN significantly outperformed the original Multi-band MelGAN and HiFi-GAN V1 models in the analysis–synthesis condition for the male speaker (Fig. 6). In the other conditions, including TTS and unseen speaker synthesis, there were no significant differences between the proposed Multi-stream HiFi-GAN and HiFi-GAN V1 models. These results validate the effectiveness of the proposed data-driven waveform decomposition, in comparison with the conventional multi-band decomposition approach [27]. Additionally, in the TTS conditions, the Parallel Tacotron 2 models achieved high-fidelity synthesis, similar to that of the Transformer models, for both JAF001 and JAM017[11]. If fine-tuning with mel-spectrograms predicted by Parallel Tacotron 2 is applied, the synthesis quality for the TTS conditions may be improved as much as that for the analysis–synthesis conditions [26]. The synthesis quality of multi-speaker models trained by the JVS corpus was lower than that of single-speaker models. There was no model that achieved a significantly higher synthesis quality than the others among HiFi-GAN V1 (a), (b), and (c). Therefore, sub-pixel CNN upsampling is better than the other upsampling models, with respect to inference speed.

In summary, Multi-stream HiFi-GAN can successfully increase the synthesis speed while improving or maintaining the synthesis quality, compared with the original HiFi-GAN.

[10]https://is.gd/XBmrMi

[11]The MOS values for the TTS conditions were sometimes higher than those for the analysis–synthesis conditions. This may be because the phoneme durations predicted by the acoustic models were suited to the listening subjects [58].

## 6. FUTURE WORK

Although the kernel size in the final CNN layer was set to 63, to allow a direct comparison with the synthesis filter used in Multi-band MelGAN, further investigation of the network structure and parameters is required to further improve the synthesis quality and increase the inference speed of the proposed method. Multi-stream HiFi-GAN should also be compared with other recent models, such as Fre-GAN [29], Basis-MelGAN [32], and UnivNet [31]. Furthermore, the proposed multi-stream structure will also be applied to an entire end-to-end TTS model, such as VITS [7].

## 7. CONCLUSIONS

To increase the inference speed while maintaining the synthesis quality of HiFi-GAN, this paper proposed Multi-stream MelGAN and HiFi-GAN. In the proposed methods, the fixed synthesis filter of Multi-band MelGAN is replaced by a trainable CNN layer with the same structure as the synthesis filter. In contrast to Multi-band MelGAN, the proposed methods use the trainable synthesis filter to decompose speech waveforms in a data-driven manner. Additionally, to evaluate Multi-stream HiFi-GAN as an entire real-time neural TTS system on CPUs, Parallel Tacotron 2, with forced alignment and accentual label input for pitch accent languages, was implemented. The results of experiments demonstrated that Multi-stream HiFi-GAN can increase the synthesis speed while improving or maintaining the synthesis quality, compared with the original HiFi-GAN.

## 8. REFERENCES


[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, RJ Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Apr. 2018, pp. 4779–4783.

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

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

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

[5] 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.

[6] R. J. Weiss, RJ Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *Proc. ICASSP*, June 2021, pp. 5664–5668.

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

[8] 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.

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

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. ICML*, July 2015, pp. 1530–1538.

[17] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, July 2018, pp. 3915–3923.

[18] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR*, May 2019.

[19] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, May 2019, pp. 3617–3621.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Dec. 2014, pp. 2672–2680.

[21] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Proc. ICLR*, May 2019.

[22] 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.

[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. ICASSP*, May 2020, pp. 6199–6203.

[24] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *Proc. ICLR*, Apr. 2020.

[25] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, "VocGAN: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network," in *Proc. Interspeech*, Oct. 2020, pp. 200–204.

[26] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, Dec. 2020, pp. 17022–17033.

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

[28] Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda, "Quasi-Periodic Parallel WaveGAN: A non-autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 792–806, 2021.

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

[30] J. Cong, S. Yang, L. Xie, and D. Su, "Glow-WaveGAN: Learning speech representations from GAN-based variational autoencoder for high fidelity flow-based speech synthesis," in *Proc. Interspeech*, Aug. 2021, pp. 2182–2186.

[31] 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.

[32] Z. Liu and Y. Qian, "Basis-MelGAN: Efficient neural vocoder based on audio decomposition," in *Proc. Interspeech*, Aug. 2021, pp. 2222–2226.

[33] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, Dec. 2020.

[34] 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.

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

[36] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, N. Dehak, and W. Chan, "WaveGrad 2: Iterative refinement for text-to-speech synthesis," in *Proc. Interspeech*, Aug. 2021, pp. 3765–3769.

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

[38] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *Proc. NeurIPS*, Dec. 2020, pp. 8067–8077.

[39] R. E. Crociere and L. R. Rabiner, *Multirate Digital Signal Processing*, Prentice Hall, Englewood Cliffs, 1983.

[40] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[41] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, RJ Skerry-Ryan, and Y. Wu, "Parallel Tacotron 2: A non-autoregressive neural TTS model with differentiable duration modeling," in *Proc. Interspeech*, Aug. 2021, pp. 141–145.

[42] J. Pons, S. Pascual, G. Cengarle, and J. Serrà, "Upsampling artifacts in neural audio synthesis," in *Proc. ICASSP*, June 2021, pp. 3000–3004.

[43] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, Oct. 2016.

[44] A. Gritsenko, T. Salimans, R. van den Berg, J. Snoek, and N. Kalchbrenner, "A spectral energy distance for parallel speech synthesis," in *Proc. NeurIPS*, Dec. 2020, pp. 13062–13072.

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

[46] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super resolution using neural networks," in *Proc. ICLR*, Apr. 2017.

[47] K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka, "Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks," in *Proc. SLT*, Dec. 2018, pp. 632–639.

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

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

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

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

[52] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," in *Proc. ICLR*, May 2021.

[53] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. Weiss, and Y. Wu, "Parallel Tacotron: Non-autoregressive and controllable TTS," in *Proc. ICASSP*, June 2021, pp. 5694–5698.

[54] F. Wu, A. Fan, A. Baevski, Y. Dauphin, and M. Auli, "Pay less attention with lightweight and dynamic convolutions," in *Proc. ICLR*, May 2019.

[55] M. Cuturi and M. Blondel, "Soft-DTW: a differentiable loss function for time-series," in *Proc. ICML*, Aug. 2017, pp. 894–903.

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

[57] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Real-time neural text-to-speech with sequence-to-sequence acoustic model and WaveGlow or single Gaussian WaveRNN vocoders," in *Proc. Interspeech*, Sept. 2019, pp. 1308–1312.

[58] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems," in *Proc. ASRU*, Dec. 2019, pp. 214–221.

[59] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Tranformer-based text-to-speech with weighted forced attention," in *Proc. ICASSP*, May 2020, pp. 6729–6733.

[60] 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. Tech.*, vol. 41, no. 5, pp. 761–768, Sept. 2020.

[61] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, May 2015.

[62] I. Loshchilov and F Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, May 2019.

[63] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proc. ICASSP*, May 2020, pp. 7654–7658.

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

[65] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. ICLR*, Apr. 2020.

[66] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga, and H. Kawai, "Investigation of training data size for real-time neural vocoders on CPUs," *Acoust. Sci. Tech.*, vol. 42, no. 1, pp. 65–68, Jan. 2021.

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