IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 26, 2021, accepted June 10, 2021, date of publication June 15, 2021, date of current version July 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3089565

Full-Band LPCNet: A Real-Time Neural Vocoder for 48 kHz Audio With a CPU

KEISUKE MATSUBARA[1,2], TAKUMA OKAMOTO[2], (Member, IEEE), RYOICHI TAKASHIMA[1], (Member, IEEE), TETSUYA TAKIGUCHI[1], (Member, IEEE), TOMOKI TODA[2,3], (Senior Member, IEEE), YOSHINORI SHIGA[2,4], 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
[4]School of Engineering, Tokyo Denki University, Tokyo 120-8551, Japan

Corresponding author: Keisuke Matsubara (kmatsubara@stu.kobe-u.ac.jp)

ABSTRACT This paper investigates a real-time neural speech synthesis system on CPUs that can synthesize high-fidelity 48 kHz speech waveforms to cover the entire frequency range audible by human beings. Although most previous studies on 48 kHz speech synthesis have used traditional source-filter vocoders or a WaveNet vocoder for waveform generation, they have some drawbacks regarding synthesis quality or inference speed. LPCNet was proposed as a real-time neural vocoder with a mobile CPU but its sampling frequency is still only 16 kHz. In this paper, we propose a Full-band LPCNet to synthesize high-fidelity 48 kHz speech waveforms with a CPU by introducing some simple but effective modifications to the conventional LPCNet. We then evaluate the synthesis quality using both normal speech and a singing voice. The results of these experiments demonstrate that the proposed Full-band LPCNet is the only neural vocoder that can synthesize high-quality 48 kHz speech waveforms while maintaining real-time capability with a CPU.

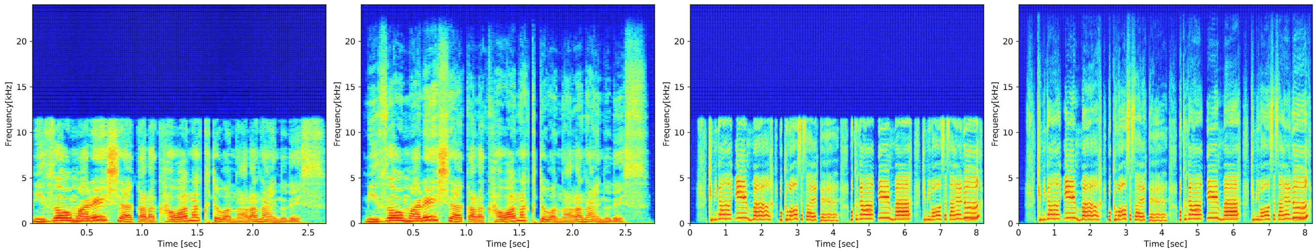
INDEX TERMS Speech synthesis, neural vocoder, LPCNet, text-to-speech, singing voice synthesis.

I. INTRODUCTION

Text-to-speech (TTS) and singing voice synthesis are important speech technologies for creating a more accessible society, and have therefore long been a subject of research. In recent years, a succession of TTS techniques using deep neural networks have been developed, and the quality of synthetic speech has improved significantly [1], [2]. Most neural TTS architectures consists of two modules: a neural acoustic model and a neural vocoder model. A neural acoustic model receives text and infers the acoustic features, such as mel-spectrograms, that correspond to the input text. A neural vocoder model receives acoustic features from acoustic models to generate raw speech waveforms. Notably, neural vocoders, such as the WaveNet vocoder [3] can synthesize more higher-quality speech waveforms than conventional source-filter vocoders [4]–[6], and they have greatly contributed to the improvement of neural TTS.

Because of its autoregressive architecture, which uses very large convolutional layers, the WaveNet vocoder suffers from the problem of low inference speed. To solve this problem, many neural vocoder models have been proposed [7]–[25], and they can synthesize high-quality speech waveforms in real time. Although most real-time neural vocoders require a GPU or multiple CPU cores for real-time synthesis, WaveRNN [8] and LPCNet [11] are neural vocoders that can perform real-time synthesis even on a CPU. WaveRNN is also used for singing voice synthesis [26], [27]. Additionally, they can be simply trained with the (dual) softmax loss functions in the time domain [1], [8] in contrast to other neural vocoders, which require multiple loss functions, such as short-time Fourier transform (STFT) loss [7] or adversarial loss [28]. However, the sampling frequency used in these vocoder is 24 kHz at the most. As shown in Fig. 1, human speech waveforms include frequency components higher than 12 kHz, and these high frequencies cannot be represented if the sampling frequency is less than (or equal to) 24 kHz. In particular, the harmonic structures are also included above 12 kHz in the singing voice, as shown in Fig. 1(d). Additionally,

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



(a) Normal speech at 24 kHz (b) Normal speech at 48 kHz (c) Singing voice at 24 kHz (d) Singing voice at 48 kHz

FIGURE 1. Spectrograms of normal speech and singing voice.

the audible frequency range for humans is considered to be around 20 kHz. In [29], the perceptual quality of synthesized speech was improved by increasing the sampling frequency from 16 kHz to 48 kHz. Although the ability to perform full-band synthesis has little merit in normal speech synthesis, we believe that it will be effective in tasks where expressiveness is more important such as singing voice synthesis and emotional speech synthesis. Therefore, neural vocoders that can synthesize audio with a higher sampling frequency are important for representing human speech more accurately.

Some previous studies have performed 48 kHz neural speech synthesis [6], [29]–[39], but almost all of them have used traditional source-filter vocoders [6], [29]–[31], [35], [38] or a WaveNet vocoder [32], [33], [36] for waveform generation. The performance of traditional source-filter vocoders is clearly inferior to that of neural vocoders, and the WaveNet vocoder has the problem of low inference speed as noted above. Some previous studies conducted real-time synthesis of full-band speech using a neural vocoder [34], [40], but they required complicated training processes, such as adversarial training [28]. Reference [41] proposed WG-WaveNet which comprises non-autoregressive WaveNet and WaveGlow and realizes full-band audio synthesis in real-time using a CPU.

In this paper, we propose a Full-band LPCNet to synthesize 48 kHz speech waveforms with a CPU by introducing some modifications to the conventional LPCNet. The original LPCNet was proposed as a neural vocoder to synthesize 16 kHz speech in real time with a mobile CPU. Although some improvements in the performance of LPCNet have been studied [42]–[53], they have still used sampling frequencies of 16 kHz or 24 kHz. Modifications of LPCNet to synthesize 48 kHz speech waveforms cause some problems: a decrease in the inference speed and an increase in the difficulty of inference. LPCNet has an autoregressive architecture; therefore, its inference speed may decrease in inverse proportion to the increase in sampling frequency. Moreover, inference may become more difficult because the time resolution of the output speech increases while the time resolution of the input acoustic features remains unchanged. It is possible to increase the time resolution of the input features, but this is not desirable because it makes inference difficult in the former acoustic model. With regard to the specific modifications, we increased the number of model parameters and the number of dimensions of the input acoustic features. When modifying the input features, we designed a novel filter bank based on the Bark scale [54] for full-band speech synthesis. In singing voice synthesis, we found it necessary to adjust the batch length of the input features appropriately. Although we performed only a simple extension for full-band synthesis in this study, acceleration methods such as subband [20], [33], [49], [51], sample bunching [50], and tensor decomposition [47] methods can be directly applied to Full-band LPCNet to further improve the synthesis speed.

We conducted experiments for both normal speech synthesis and singing voice synthesis conditions to evaluate Full-band LPCNet in comparison with the WORLD [5], WaveNet, Parallel WaveGAN [15], WG-WaveNet [41], and PeriodNet [40] vocoders with a sampling frequency of 48 kHz. Additionally, we investigated the inference speeds by increasing the number of model parameters of Full-band LPCNet. The results of these experiments demonstrate that full-band LPCNet is the only neural vocoder that can synthesize higher-quality 48 kHz speech waveforms in real-time with a CPU, as a result of some simple but effective modifications.

The rest of this paper is organized as follows. Section II briefly introduces the conventional LPCNet. Full-band LPCNet is then proposed in Section III. In Section IV, experiments are described and the results are discussed. Section V concludes the paper.

II. LPCNet

LPCNet is a WaveRNN-based neural vocoder model with a recurrent neural network architecture. LPCNet predicts residual signals between natural speech and predicted speech calculated by linear predictive coding (LPC) [55]. Whereas WaveRNN infers 16 bit audio samples using dual-softmax [8], LPCNet synthesizes residual samples that are compressed using 8-bit μ-law coding [56], which can suppress quantization errors. Therefore, LPCNet can synthesize high quality speech waveforms while reducing the network model size [11].

Fig. 2 shows an overview of the LPCNet architecture. LPCNet comprises two neural network blocks the frame rate network and the sample rate network. The input features

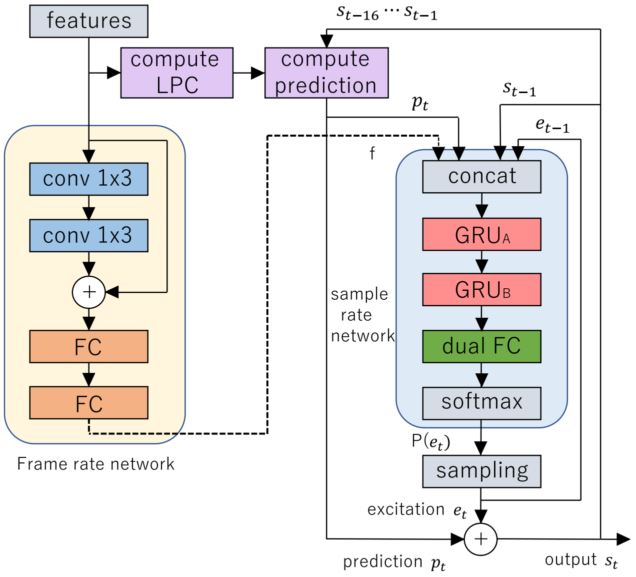


FIGURE 2. Overview of LPCNet.

consist of 18-dimensional Bark-Frequency Cepstrum Coefficients (BFCCs) [54] and two pitch parameters (period and correlation) for a sampling frequency of 16 kHz.

Samples predicted by LPC are computed as follows:

$$p_t = \sum_{k=1}^{M} a_k s_{t-k}, \tag{1}$$

where p_t and s_t denote the predicted samples and natural samples, respectively at time t. a_k is the k-th linear prediction coefficient.

The linear prediction coefficients a_k are calculated from the input BFCCs. Specifically, the BFCCs are first converted to a linear-frequency power spectral density, which is then converted to an autocorrelation by applying the inverse FFT. Finally, the prediction coefficients are computed from the autocorrelation using the Levinson-Durbin algorithm.

The frame rate network extracts intermediate features from the input acoustic features. The sample rate network receives a previous one-step natural sample, a predicted sample using LPC and the output from the frame rate network to infer a current residual sample.

The sample rate network is also an autoregressive model that comprises two gated recurrent unit (GRU [57]) layers (The first and second layers are referred to as $\mathrm{GRU_A}$ and $\mathrm{GRU_B}$, respectively). Therefore, it originally requires a long time for inference as is the case in WaveNet. However, sparse coding, which forces the lowest value of a weight matrix to zero is applied to accelerate inference, as with sparse WaveRNN [8]. LPCNet can also synthesize speech waveforms in real-time, even in the restricted environment of a mobile CPU [11].

III. FULL-BAND LPCNet

The original LPCNet was proposed as a neural vocoder that can synthesize speech waveforms for a sampling frequency of 16 kHz and some subsequent research has produced a 24 kHz LPCNet with higher fidelity synthesis [43], [46], [52], [53]. As described in Section I, we propose a Full-band LPCNet by introducing the following simple but effective modifications to synthesize high-fidelity speech waveforms with a sampling frequency of 48 kHz, which can cover the entire speech waveform and human auditory frequency ranges, using a CPU.

A. PROPOSED INPUT FEATURES

We expand the 18-dimensional BFCCs to 50-dimensional BFCCs. The original LPCNet uses a voice compression method called Opus [58] for the integration of frequency filter banks. With Opus, the frequency bands are divided at regular intervals at low frequencies and then divided following the Bark scale at high frequencies. The frequencies from 0 to 8 kHz are then integrated into 18 filter banks. In this study, we applied more subdivided Bark scale filter banks instead of the Opus scale.

The conversion from a linear frequency scale to the Bark scale is given by

$$B = 13 \arctan(0.00076f) + 3.5 \arctan((f/7500)^2), \tag{2}$$

where f and B are the linear and Bark frequencies, respectively.

The original Bark scale integrates the frequencies from 0 to 15.5 kHz into 24-dimensional filter banks using Eq. (2). Although the Bark scale above 15.5 kHz is not defined, we assume that (2) is still applicable when using 48 kHz audio, up to 24 kHz. We then integrate the frequencies from 0 to 24 kHz into 50-dimensional filter banks; this scale is more specific than the original Bark scale as shown in Fig. 3.

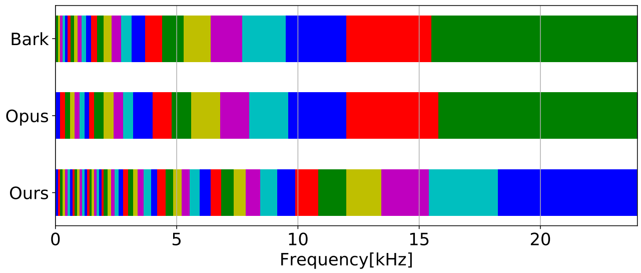


FIGURE 3. Band layouts of Bark, Opus, and proposed (Ours).

B. INCREASING NUMBER OF MODEL PARAMETERS

For high-fidelity speech synthesis with a sampling frequency of 48 kHz, we investigated the effect of the number of model parameters in LPCNet. Increasing the number of parameters may cause a decrease in the inference speed, and improve the quality of the synthesized speech. As the result of preliminary experiments, we found that the number of $\mathrm{GRU_A}$ parameters drastically affects the synthesis quality. As shown in Table 1, $\mathrm{GRU_A}$ is an important layer that accounts for most of the model parameters. Therefore, we increased the number of

TABLE 1. Number of parameters in Full-band LPCNet.

GRU_A Channels	Overall	GRU_A	GRU_B
384	1.26M	1.03M	3.77K
512	1.81M	1.58M	3.77K
640	2.47M	2.21M	3.77K

GRU_A parameters from original 384 to 512 and 640, to investigate the effect on synthesis quality and inference speeds, leaving the number of GRU_B parameters fixed to 16, as in the original LPCNet.

C. APPROPRIATE BATCH LENGTH IN TRAINING

In singing voice synthesis, training under the same conditions as normal speech synthesis did not proceed sufficiently. As a result of investigation, we found that adjusting the input batch length appropriately had a great effect on quality in singing voice synthesis. In particular, the input batch length, which is 16 frames (7680 audio samples) in normal speech synthesis, and it was reduced to 3 frames (1440 voice sample) in singing voice synthesis.

IV. EXPERIMENTS

A. EXPERIMENTAL CONDITIONS

We conducted two subjective experiments to evaluate Full-band LPCNet in comparison with the conventional WORLD [5], WaveNet (8 bit μ-law) [3], Parallel WaveGAN [15] and WG-WaveNet [41] vocoders, for a sampling frequency of 48 kHz, regarding inference speed and synthesis quality.[1] One experiment was conditioned by normal speech, and another was conditioned by singing audio. We believe that the experiment conditioned by singing audio can more effectively evaluate quality with 48 kHz audio because singing audio includes high-frequency components such as harmonic structures than normal speech, as shown in Fig. 1. In the experiments, we modified the open-source code of LPCNet,[2] WaveNet vocoder,[3] Parallel WaveGAN,[4] and WG-WaveNet[5] for training and synthesis of 48 kHz audio to ensure reproducibility. We implemented PeriodNet by modifying the source code of Parallel WaveGAN.

Dataset: For the experiment with normal speech, we used 7697 sentences (about 10 hours of recorded speech) uttered by a Japanese female speaker, from the JSUT corpus [59] to ensure reproducibility. JSUT is a free speech corpus with a sampling frequency of 48 kHz. We removed the silent parts of the speech samples by applying forced alignment using the Julius speech recognition toolkit [60]. To train the neural vocoder models, 7497 utterances (all except Basic5000-0001 to Basic5000-0200) were used. To train acoustic models in the TTS condition, 4800 sentences (Basic5000-0201 to Basic5000-5000) were used because HTS-style context labels based on manual annotation were available.[6] One hundred utterances (Basic5000-0101 to Basic5000-0200) were used for validation, and the remaining 100 utterances (Basic5000-0001 to Basic5000-0100) were used for evaluation.

For the experiment with singing audio, we used 50 acapella songs (about 1 hour) uttered by a Japanese female speaker, from the Tohoku Kiritan corpus [61] to ensure reproducibility. Kiritan an open singing voice corpus with a sampling frequency of 96 kHz. We downsampled the audio to 48 kHz and clipped it into segments of appropriate length. We separated all 50 songs into phrases by using the provided labels and used two songs (05.wav and 30.wav), each of which includes 10 phrases, for evaluation; the remaining 48 songs constructed from 376 phrases, were used for training. Therefore, 376 phrases were used for training and 20 phrases were used for evaluation.

LPCNet: The network structure of Full-band LPCNet was the same as that in [11] except for the number of GRU_A units. The number of GRU_A parameters was set to 384, 512, and 640. The input features comprised 50-dimensional BFCCs, pitch period, and pitch correlation. The number of LPC filter coefficients was the same as in the original implementation. To calculate the BFCCs, spectrum analysis was performed with a window length of 20 ms and a frame shift of 10 ms, and the Bark-scale filter bank was applied. Pitch calculation was based on an open-loop cross-correlation search. Additionally, Full-band LPCNet was compared with LPCNet for a sampling frequency of 24 kHz with 30-dimensional BFCCs has investigated in [46], [53]. An Adam optimizer was used for parameter updating [62].

WORLD: In the WORLD vocoder for a sampling frequency of 48 kHz, 50-dimensional mel-cepstra with warping coefficient $\alpha = 0.55$, and five-dimensional parameters for the smooth vocal tract spectrum and aperiodicity components were obtained from the original WORLD spectrum with 2,048 dimensions. The vocoded waveforms were synthesized using the compressed acoustic features [5]. The fundamental frequency (F0) was analyzed by the Harvest algorithm [63].

WaveNet vocoder: The network structure of the WaveNet vocoder was the same as that in [3]. We applied time-invariant noise shaping [64] to suppress the perceptual noise components caused by the prediction error where 35-dimensional mel-cepstra were used and a parameter to control noise energy in the formant regions was set to 0.5. The input features were 80-band log-mel spectrograms with a band-limited frequency range (80 to 7,600 Hz). The window and shift lengths were set to 42.7 ms and 10 ms, respectively.

[1] Although Multi-band MelGAN [22] can realize real-time synthesis with multiple CPU cores, it was not included in the experiments because the synthesis quality of Multi-band MelGAN was significantly worse than that of Parallel WaveGAN in preliminary experiments with a sampling frequency of 24 kHz. Although LVCNet [25] can realize real-time synthesis with a CPU for 24 kHz audio, it was also not included in the experiments because its synthesis quality was almost the same as that of Parallel WaveGAN [25].

[2] https://github.com/mozilla/LPCNet

[3] https://github.com/kan-bayashi/PytorchWaveNetVocoder

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

[5] https://github.com/BogiHsu/WG-WaveNet

[6] https://github.com/sarulab-speech/jsut-label

WG-WaveNet: The network structure of the WG-WaveNet vocoder was the same as that in [41]. The input features were also 80-band log-mel spectrograms with a band-limited frequency range (80 to 7,600 Hz) used in the WaveNet vocoder.

Parallel WaveGAN: The network structure of Parallel WaveGAN was the same as that in [15]. In the conditions of Parallel WaveGAN, we used two types of input features: 50-dimensional BFCCs and pitch features, as used in Full-band LPCNet (lpcn), and 80-dimensional log-mel spectrograms, as used for WaveNet and WG-WaveNet (melspc).[7] To extend WG-WaveNet and Parallel WaveGAN from 24 kHz to 48 kHz, we modified the parameters of the STFT loss used in them by doubling all the FFT sizes, window lengths, and frame shift lengths.

PeriodNet: The network structure of the PeriodNet vocoder was the same as that of the Non-AR series model in [40]. The implementation was based on Parallel WaveGAN, and two generators that generate periodic and aperiodic signals as well as discriminators that operate at multiple sampling frequencies were added. As input features, we used the same features as those used in the WORLD vocoder described above. As excitation signals, we used sine waves calculated from the F0 values.

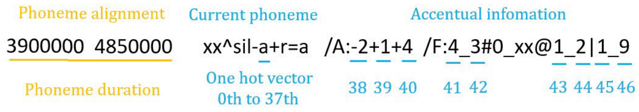


FIGURE 4. Phoneme alignment and 47-dimensional input feature vectors of acoustic models for TTS obtained from context labels.

Acoustic models: For the TTS condition, we used a FastSpeech-based acoustic model [65] with full-context label input to estimate a proper accent as in [66]–[68]. Although 575-dimensional context vectors were used as input features with the JSUT corpus [69], simple 47-dimensional vectors constructed from 38-dimensional phoneme one-hot vectors and nine-dimensional accentual label vectors, were used for the acoustic models, as shown in Fig. 4. To train of the duration predictor, the phoneme durations were obtained from the context labels, as shown in Fig. 4. The output features of the acoustic models were 52-dimensional features for Full-band LPCNet. The output features were normalized to have a zero-mean and unit-variance. The network structure of the acoustic model was based on the implementation of ESPnet-TTS [70] and some modifications were applied to input full-context labels. In contrast to that used for LPCNet, a RAdam optimizer was used for the acoustic model, WaveNet vocoder, WG-WaveNet, Parallel WaveGAN, and PeriodNet [71].

B. REAL-TIME FACTOR EVALUATION

We measured the real-time factors (RTFs) of the neural vocoders and acoustic model to evaluate the synthesis speeds. Although simple PyTorch-based implementations were used in the WaveNet vocoder, Parallel WaveGAN, WG-WaveNet, PeriodNet and acoustic model, a C-based implementation was used in Full-band LPCNet for inference.

Table 2 shows the RTFs for inference using an NVIDIA GeForce RTX2060 GPU or Intel Core i9-10900X. We found that Full-band LPCNet achieved real-time synthesis with only one CPU core, as in [52], even for 48 kHz audio synthesis. Comparing to all conditions of neural vocoders that use a CPU, WG-WaveNet achieved the fastest RTFs. However, as described below, the perceptual quality of WG-WaveNet was not sufficient. The RTF of Parallel WaveGAN and PeriodNet did not fall below 1.0 in this experiment. To improve the RTF of Parallel WaveGAN and PeriodNet, a C-based implementation is required instead of PyTorch, as used in Full-band LPCNet. These results suggest that a full-band real-time neural TTS can be realized by Full-band LPCNet combined with the acoustic models based on FastSpeech.

TABLE 2. Real-time factors for inference using an NVIDIA GeForce RTX2060 GPU or Intel Core i9-10900X (maximum 20 cores).

Model	GPU	CPU
FastSpeech (24 kHz)	0.01	0.05 (20 cores)
FastSpeech (48 kHz)	0.01	0.05 (20 cores)
WaveNet	641	6322 (20 cores)
Parallel WaveGAN	0.05	1.92 (20 cores)
WG-WaveNet	0.004	0.63 (20 cores)
PeriodNet	0.12	3.86 (20 cores)
Full-band LPCNet[384]	-	0.44 (1 core)
Full-band LPCNet[512]	-	0.62 (1 core)
Full-band LPCNet[640]	-	0.85 (1 core)

C. EVALUATION OF NORMAL SPEECH SYNTHESIS

1) OBJECTIVE EVALUATION

We evaluate the distortion between the original and synthetic speech. The metrics used for evaluation are signal-to-noise ratio (SNR), root mean square error (RMSE) of the spectrogram (Spec-RMSE), mel-cepstrum distortion (MCD), and F0 RMSE, as in [3] where F0 was also analized by the Harvest algorithm [63]. Table 3 shows the result of the objective evaluation. With the evaluation metrics related to the frequency domain, such as Spec-RMSE and MCD, the results of neural vocoders dare inferior to those of WORLD. However, with the metrics related to the time domain, such as SNR, these vocoders achieved better scores than WORLD. These results show that neural vocoders were able to directly model the audio signal, including phase matching, whereas the phase of audio synthesized by the WORLD vocoder did not match that of the original audio [72]. Compared with the results of neural vocoders, PeriodNet achieved the highest SNR score. Because PeriodNet receives the excitation signal explicitly, we believe that fidelity to periodic signals tended to be prioritized, and as a result, the SNR was improved. Among

[7] As described in [9], the WaveNet vocoder, WG-WaveNet and Parallel WaveGAN for 48 kHz audio with full-band mel-spectrograms as used by WaveGrad [23] were also investigated. However, they could not outperform those with band-limited mel-spectrograms in preliminary experiments. The frequency range and number of dimensions for these vocoders should be further investigated as future work.

TABLE 3. Objective evaluation results in normal speech synthesis.

Model	Spec-RMSE (dB)	Spec-RMSE (12 kHz~) (dB)	MCD (dB)	SNR (dB)	F0 RMSE (cent)
WORLD	6.76	4.14	3.31	−0.34	1.11
WaveNet	7.23	5.07	5.13	2.32	1.22
Parallel WaveGAN (melspc)	7.17	4.36	4.51	2.25	1.44
WG-WaveNet	8.81	5.91	4.51	1.22	1.45
PeriodNet	6.81	4.21	3.37	5.11	1.25
Full-band LPCNet[384]	7.14	4.38	4.90	3.37	1.41
Full-band LPCNet[512]	7.16	4.42	4.82	3.27	1.51
Full-band LPCNet[640]	7.09	4.38	4.71	3.35	1.36

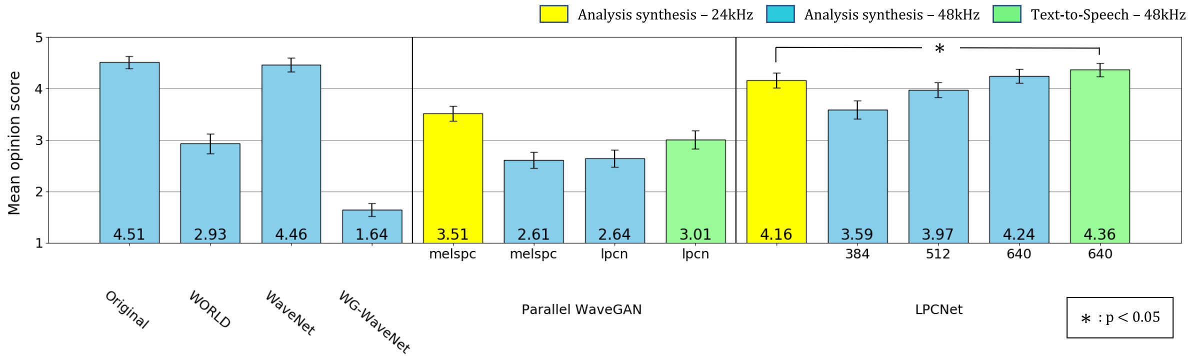


FIGURE 5. Result of MOS test using normal speech with 10 listening subjects. Confidence level of the error bars is 95 %.

the other neural vocoders, Full-band LPCNet achieved the second highest SNR score; this is because WaveNet and LPCNet use the loss function in the time domain which can maximize SNR whereas WG-WaveNet and Parallel WaveGAN use multiple STFT losses which do not explicitly consider phase components.

2) SUBJECTIVE EVALUATION

We conducted mean opinion score (MOS) tests with a five-point scale: 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad to evaluate the subjective perceptual quality of the synthesized speech waveforms [73]. Ten native Japanese speakers (undergraduate and graduate students) without hearing loss listened to the synthesized speech samples using headphones (20 utterances × 13 conditions including the ground-truth condition = 260 utterances).

Fig. 5 shows the results of the MOS test using normal speech. The WaveNet vocoder achieved the best score, which was comparable to the score of the original audio. Although WaveNet slightly outperformed Full-band LPCNet with 640 GRU_A units, WaveNet suffers from a fundamental inference speed problem. Although Full-band LPCNet with 384 GRU_A units achieved a lower score than with 512 and 640 units, it still outperformed WORLD, Parallel WaveGAN, and WG-WaveNet. In this experiment, Parallel WaveGAN and WG-WaveNet were inferior to WORLD, which is a conventional vocoder. This may be because we simply used the same parameters as those used in the original models. The synthesis quality of these vocoders may be improved by adjusting the hyperparameters to match the corpus and sampling frequency. Although we evaluated two conditions using different acoustic features in Parallel WaveGAN, there was no significant difference between these conditions. This means that these two features do not differ in expressiveness. Comparing the conditions of 24 kHz and 48 kHz audio, there was no significant difference between 24 kHz LPCNet and 48 kHz LPCNet with 640 GRU_A units. We believe that the accuracy of the high-frequency components was not sufficiently evaluated because the corpus comprises normal speech.

For TTS conditions, Full-band LPCNet with 640 GRU_A units achieved the best score surpassing the results of analysis-synthesis and it was comparable to the original score.[8] Although the acoustic features estimated by TTS are usually inferior to those extracted from natural speech, this inferiority was not confirmed by this experiment. In [15], it was been reported that a full-band mel-scale spectrogram inferred by TTS causes over-smoothing of high-frequency components and deterioration of quality. However, the proposed acoustic features for Full-band LPCNet comprises the Bark cepstrum equivalent to the vocal tract filter and the fundamental frequency equivalent to the vocal

[8] Because the speaker in the JSUT corpus is not a professional speaker and the pronunciation is often inaccurate, we believe that the smoothed speech inferred by TTS was often more highly evaluated than the natural audio. Therefore, the MOS values of Full-band LPCNet and Parallel WaveGAN with estimated acoustic features by TTS were higher than those of the analysis-synthesis condition. Evaluation using a professional speaker with a sampling frequency of 48 kHz is a subject for future work.

TABLE 4. Objective evaluation results in singing voice synthesis.

Model	Spec-RMSE (dB)	Spec-RMSE (12 kHz~) (dB)	MCD (dB)	SNR (dB)	F0 RMSE (cent)
WORLD	5.73	2.86	3.06	1.57	1.14
WaveNet	6.51	3.40	5.23	2.88	1.25
Parallel WaveGAN (melspc)	6.36	3.09	4.34	0.37	1.21
WG-WaveNet	8.71	4.46	8.59	0.73	1.21
PeriodNet	5.62	2.87	3.13	6.42	0.98
Full-band LPCNet[384]	6.64	3.25	4.63	1.39	1.62
Full-band LPCNet[512]	6.68	3.26	4.68	2.42	1.62
Full-band LPCNet[640]	6.51	3.24	4.51	2.36	1.72

cord vibration. We assume that these features are easier to predict than mel-spectrograms and they achieved sufficient quality for 48 kHz speech waveform synthesis. Therefore, the proposed acoustic features of Full-band LPCNet are effective for neural TTS.

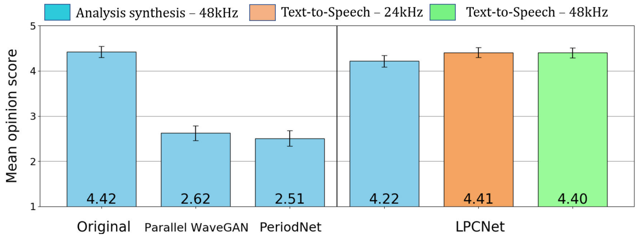


FIGURE 6. Result of MOS test including PeriodNet using normal speech with 10 listening subjects. Confidence level of the error bars is 95 %.

Figure 6 shows the results of the MOS test including the conditions of PeriodNet and 24 kHz TTS. Although PeriodNet achieved an excellent score in the objective evaluation, the MOS score of PeriodNet was insufficient and comparable with that of Parallel WaveGAN. We believe that PeriodNet has insufficient fidelity to aperiodic signals, and this deteriorates the quality. We also believe that the objective evaluation did not sufficiently evaluate the quality of the aperiodic signal. In the original paper of PeriodNet, because the quality of speech was only evaluated for singing voice synthesis, we think that an appropriate adjustment of the hyperparameters for normal speech synthesis may be needed. In addition, there was no significant difference between 24 kHz TTS and 48 kHz TTS. Because normal speech contains few components at high frequencies, it may have been difficult to perceive the difference between these qualities.

D. EVALUATION OF SINGING VOICE SYNTHESIS

1) THE EFFECT OF BATCH LENGTH

As mentioned in Section III-B, it was necessary to adjust the input batch lengths appropriately when training LPCNet using a singing voice. Fig. 7 shows the transition of the loss function on with batch lengths of 16 and 3. When the value of the loss function spikes, sparse coding is applied to simplify the network. In the case of the batch length of 16, the loss value did not decrease and the network was not trained sufficiently. Conversely, with a batch length of 3, it generally converged to a loss value during training, in the same manner as normal speech.

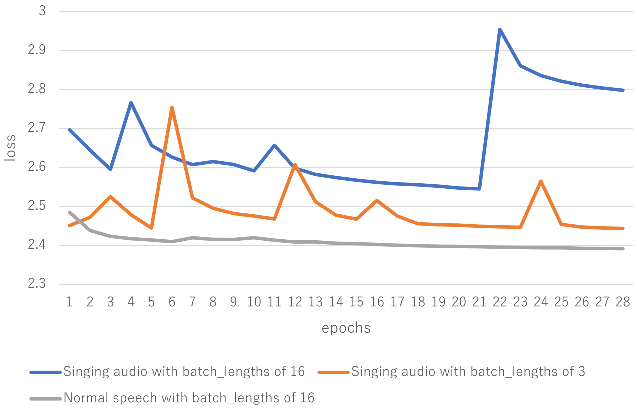


FIGURE 7. Transition of loss value for each batch length in training of Full-band LPCNet with 640 GRU_A units.

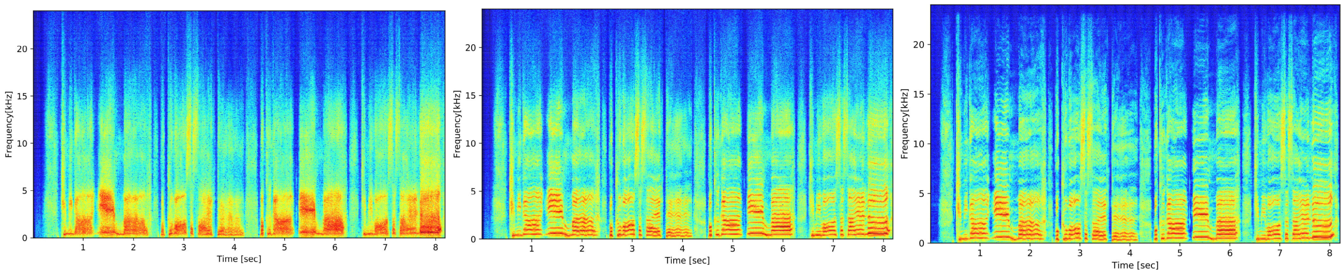
TABLE 5. Properties of F0 on each corpus.

Corpus	Mean (Hz)	Standard deviation (Hz)
JSUT	216	57
Kiritan	367	93

Fig. 8 shows the spectrograms of singing audio synthesized by Full-band LPCNet trained with batch lengths of 16 and 3, and original speech. with a batch length of 16, the spectrogram looks like a blur, whereas with a batch length of 3, the harmonic components were stably synthesized and high-fidelity synthesis was achieved. As shown in Fig. 8(b) the harmonic components above 12 kHz can be slightly synthesized by the proposed LPCNet with a batch length of 3. We believe that the reason why the processing is required is that singing audio contains a wide variety of F0. Table 5 shows the properties of F0 on the JSUT corpus and Tohoku Kiritan database. Comparing these corpora, the Kiritan database has a high average of F0 and large fluctuations. We believe that it is difficult to infer audio samples by increasing the batch length to expand the receptive field when F0 is high and the fluctuations are large.

2) OBJECTIVE EVALUATION

We evaluated the distortion between the original and synthetic speech, in the same manner as in the experiment evaluating normal speech synthesis. Table 4 shows the results of the



(a) batch length of 16 (b) batch length of 3 (c) original

FIGURE 8. Spectrograms of singing audio synthesized by Full-band LPCNet for each batch length.

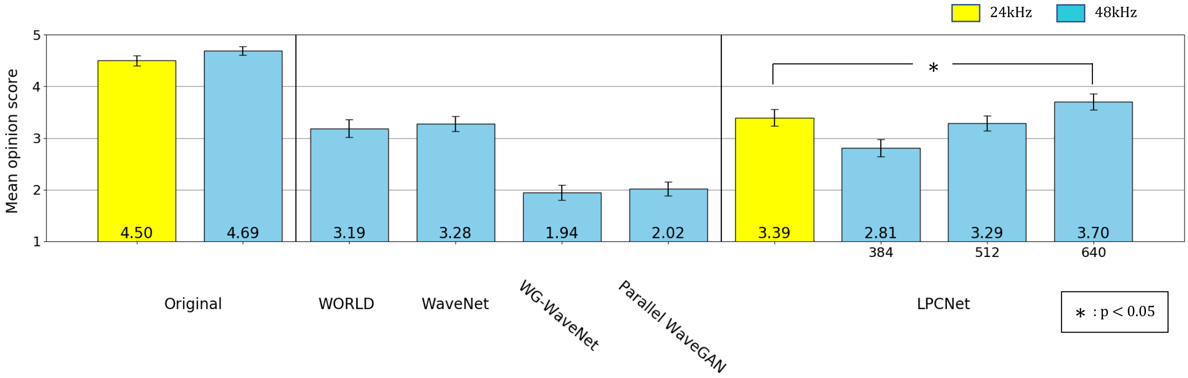


FIGURE 9. Result of MOS test using singing audio with 10 listening subjects. Confidence level of the error bars is 95 %.

objective evaluation of singing voice synthesis. With the evaluation metrics related to the frequency domain, WORLD achieved the highest score, as in the experiment with normal speech synthesis. Parallel WaveGAN, WG-WaveNet, and LPCNet with 384 $\mathrm{GRU_A}$ units had lower SNR than WORLD. The results show that modeling of the singing voice is a more difficult task than modeling normal speech.

3) SUBJECTIVE EVALUATION

To investigate the perceptual quality of speech that contains high-frequency components, we conducted a MOS test using singing audio with 20 test set phrases. The listening subjects and experimental environment were the same as with the experiment using normal speech. Fig. 9 shows the results of the MOS test using singing audio. In the results using vocoders, Full-band LPCNet with 640 $\mathrm{GRU_A}$ units achieved the best score. Remarkably, there was a significant difference between the conventional LPCNet for 24 kHz and full-band LPCNet with 640 $\mathrm{GRU_A}$ units. This means that Full-band LPCNet can synthesize high-fidelity audio including high-frequency components. However, the synthesis speech quality for singing audio was not as high as that for normal speech. We believe that this is a consequence of the lack of training data. Although [52] shows that the amount of data required for training neural vocoders for 24 kHz synthesis is about one hour, singing audio contains a wide variety of F0 and the sampling frequency is 48 kHz. Therefore, we believe that it requires more data for sufficient training. These detailed investigations are a topic of future work.

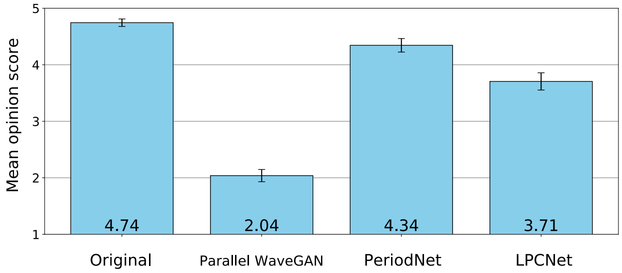


FIGURE 10. Result of MOS test including PeriodNet using singing audio with 10 listening subjects. Confidence level of the error bars is 95 %.

Figure 10 shows the results of the MOS test including the condition of PeriodNet. Compared with the proposed Full-band LPCNet, PeriodNet realized higher synthesis quality for singing voice synthesis, as in [40]. This is because a singing voice includes fewer aperiodic components than normal speech and the F0-based excitation signals were explicitly given into PeriodNet. Therefore, PeriodNet effectively achieved high quality synthesis with fewer training data in singing voice synthesis. In contrast, as mentioned above, the proposed Full-band LPCNet requires more training data for higher fidelity synthesis. However, PeriodNet cannot realize real-time inference with a CPU.

Consequently, Full-band LPCNet is the only neural vocoder that can realize real-time and high-fidelity speech synthesis with a sampling frequency of 48 kHz using a CPU. As future work, Full-band LPCNet can be made much faster by applying acceleration methods, such as the subband [20], [33], [49], [51], sample bunching [50] and tensor decomposition [47] methods. Additionally, Full-band LPCNet can be extended to multi-speaker neural vocoder to synthesize the speech waveforms of many and unspecified speakers that were not included in training [74].

V. CONCLUSION

This paper proposed Full-band LPCNet which can synthesize high-fidelity 48 kHz speech waveforms in real-time using a CPU, by introducing simple but effective modifications to the conventional LPCNet. The input feature was extended to 50-dimensional BFCC and the number of model parameters was increased. Experiments, using both normal speech and a singing voice, were conducted to compare Full-band LPCNet with conventional source-filter and neural vocoders. The results of the RTF evaluation indicate that Full-band LPCNet can realize speech synthesis and neural TTS for 48 kHz speech waveforms in real time using a CPU. The results of the subjective evaluations suggest that Full-band LPCNet can realize higher-fidelity synthesis than other real-time neural vocoders in normal speech synthesis. In particular, the effectiveness of Full-band LPCNet was validated in singing voice synthesis and compared with LPCNet for 24 kHz audio, although PeriodNet realized higher synthesis quality than the proposed Full-band LPCNet. Additionally, the proposed acoustic features with 50-dimensional BFCC were effective for neural TTS. Consequently, the results of the experiments demonstrated Full-band LPCNet is the only neural vocoder that can realize real-time and high-fidelity speech synthesis for 48 kHz speech waveforms using a CPU.

ACKNOWLEDGMENT

This work was performed while Keisuke Matsubara was interning at NICT.

REFERENCES

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

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. ICASSP*, Apr. 2018, pp. 4749–4783.

[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] 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.

[5] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[6] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "A comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1658–1670, Sep. 2018.

[7] A. van den Oord *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, Jul. 2018, pp. 3915–3923.

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

[9] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR*, May 2019. [Online]. Available: https://openreview.net/forum?id=HklY120cYm

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

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

[12] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 402–415, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8915761/footnotes#footnotes

[13] S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet: A generative flow for raw audio," in *Proc. ICML*, Jun. 2019, pp. 3370–3378.

[14] 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.

[15] 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.

[16] K. Peng, W. Ping, Z. Song, and K. Zhao, "Non-autoregressive neural text-to-speech," in *Proc. ICML*, Jul. 2020, pp. 7586–7598.

[17] W. Ping, K. Peng, K. Zhao, and Z. Song, "WaveFlow: A compact flow-based model for raw audio," in *Proc. ICML*, Jul. 2020, pp. 7706–7716.

[18] 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. [Online]. Available: https://openreview.net/forum?id=r1gfQgSFDr

[19] J. Yang, J. Lee, Y. Kim, H.-Y. 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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. [Online]. Available: https://openreview.net/forum?id=NsMLjcFaO8O

[24] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *Proc. ICLR*, May 2021. [Online]. Available: https://openreview.net/forum?id=a-xFK8Ymz5J

[25] Z. Zeng, J. Wang, N. Cheng, and J. Xiao, "LVCNet: Efficient condition-dependent modeling network for waveform generation," in *Proc. ICASSP*, Jun. 2021, pp. 6054–6058.

[26] Y.-H. Yi, Y. Ai, Z.-H. Ling, and L.-R. Dai, "Singing voice synthesis using deep autoregressive neural networks for acoustic modeling," in *Proc. Interspeech*, Sep. 2019, pp. 2593–2597.

[27] Y. Gu, X. Yin, Y. Rao, Y. Wan, B. Tang, Y. Zhang, J. Chen, Y. Wang, and Z. Ma, "ByteSing: A Chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and WaveRNN vocoders," in *Proc. ISCSLP*, Jan. 2021, pp. 1–5.

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

[29] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *Proc. ICASSP*, Apr. 2018, pp. 4804–4808.

[30] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "GlottDNN—A full-band glottal Vocoder for statistical parametric speech synthesis," in *Proc. Interspeech*, Sep. 2016, pp. 2473–2477.

[31] X. Wang, S. Takaki, and J. Yamagishi, "A comparative study of the performance of HMM, DNN, and RNN based speech synthesis systems trained on very large speaker-dependent corpora," in *Proc. SSW*, Sep. 2016, pp. 125–128.

[32] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. ICLR*, Apr. 2018. [Online]. Available: https://openreview.net/forum?id=HJtEm4p6Z

[33] 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.

[34] K. Oura, K. Nakamura, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Deep neural network based real-time speech vocoder with periodic and aperiodic inputs," in *Proc. SSW*, Sep. 2019, pp. 13–18.

[35] I. Himawan, S. Aryal, I. Ouyang, S. Kang, P. Lanchantin, and S. King, "Speaker adaptation of a multilingual acoustic model for cross-language synthesis," in *Proc. ICASSP*, May 2020, pp. 7629–7633.

[36] T. Fujimoto, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Semi-supervised learning based on hierarchical generative models for end-to-end speech synthesis," in *Proc. ICASSP*, May 2020, pp. 7644–7648.

[37] T. Saeki, Y. Saito, S. Takamichi, and H. Saruwatari, "Lifter training and sub-band modeling for computationally efficient and high-quality voice conversion using spectral differentials," in *Proc. ICASSP*, May 2020, pp. 7784–7788.

[38] J. Koguchi, S. Takamichi, M. Morise, H. Saruwatari, and S. Sagayama, "DNN-based full-band speech synthesis using GMM approximation of spectral envelope," *IEICE Trans. Inf. Syst.*, vol. E103.D, no. 12, pp. 2673–2681, 2020.

[39] J. Chen, X. Tan, J. Luan, T. Qin, and T.-Y. Liu, "HiFiSinger: Towards high-fidelity neural singing voice synthesis," 2020, *arXiv:2009.01776*. [Online]. Available: http://arxiv.org/abs/2009.01776

[40] Y. Hono, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Periodnet: A non-autoregressive waveform generation model with a structure separating periodic and aperiodic components," in *Proc. ICASSP*, Jun. 2021, pp. 6049–6053.

[41] P.-C. Hsu and H.-Y. Lee, "WG-WaveNet: Real-time high-fidelity speech synthesis without GPU," in *Proc. Interspeech*, Oct. 2020, pp. 210–214.

[42] Z. Kons, S. Shechtman, A. Sorin, C. Rabinovitz, and R. Hoory, "High quality, lightweight and adaptable TTS using LPCNet," in *Proc. Interspeech*, Sep. 2019, pp. 176–180.

[43] M.-J. Hwang, E. Song, R. Yamamoto, F. Soong, and H.-G. Kang, "Improving LPCNET-based text-to-speech with linear prediction-structured mixture density network," in *Proc. ICASSP*, May 2020, pp. 7219–7223.

[44] J.-M. Valin and J. Skoglund, "A real-time wideband neural Vocoder at 1.6kb/s using LPCNet," in *Proc. Interspeech*, Sep. 2019, pp. 3406–3410.

[45] V. Popov, M. Kudinov, and T. Sadekova, "Gaussian LPCNet for multisample speech synthesis," in *Proc. ICASSP*, May 2020, pp. 6204–6208.

[46] Y. Zheng, X. Li, F. Xie, and L. Lu, "Improving end-to-end speech synthesis with local recurrent neural network enhanced transformer," in *Proc. ICASSP*, May 2020, pp. 6734–6738.

[47] H. Kanagawa and Y. Ijima, "Lightweight LPCNet-based neural vocoder with tensor decomposition," in *Proc. Interspeech*, Oct. 2020, pp. 205–209.

[48] V. Popov, S. Kamenev, M. Kudinov, S. Repyevsky, T. Sadekova, V. Bushaev, V. Kryzhanovskiy, and D. Parkhomenko, "Fast and lightweight on-device TTS with Tacotron2 and LPCNet," in *Proc. Interspeech*, Oct. 2020, pp. 220–224.

[49] 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.

[50] R. Vipperla, S. Park, K. Choo, S. Ishtiaq, K. Min, S. Bhattacharya, A. Mehrotra, A. G. C. P. Ramos, and N. D. Lane, "Bunched LPCNet: Vocoder for low-cost neural text-to-speech systems," in *Proc. Interspeech*, Oct. 2020, pp. 3565–3569.

[51] 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.

[52] 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. Technol.*, vol. 42, no. 1, pp. 65–68, Jan. 2021.

[53] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga, and H. Kawai, "High-intelligibility speech synthesis for dysarthric speakers with LPCNet-based TTS and CycleVAE-based VC," in *Proc. ICASSP*, Jun. 2021, pp. 7058–7062.

[54] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Leiden, The Netherlands: Brill Academic Pub, 2013.

[55] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[56] *Pulse Code Modulation (PCM) Voice Frequencies*, document ITU-T Recommendation G. 711, 1988.

[57] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, Oct. 2014, pp. 1724–1734.

[58] J. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-quality, low-delay music coding in the opus codec," in *Proc. 135th AES Conv.*, Oct. 2013. [Online]. Available: https://arxiv.org/abs/1602.04845

[59] 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.

[60] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *Proc. APSIPAASC*, Oct. 2009, pp. 131–137.

[61] I. Ogawa and M. Morise, "Tohoku Kiritan singing database: A singing database for statistical parametric singing synthesis using Japanese pop songs," *Acoust. Sci. Technol.*, vol. 42, no. 3, pp. 140–145, May 2021.

[62] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, May 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

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

[64] K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "An investigation of noise shaping with perceptual weighting for wavenet-based speech generation," in *Proc. ICASSP*, Apr. 2018, pp. 5664–5668.

[65] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, Dec. 2019, pp. 3165–3174.

[66] 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*, Sep. 2019, pp. 1308–1312.

[67] 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. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 214–221.

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

[69] T. Koriyama and H. Saruwatari, "Utterance-level sequential modeling for deep Gaussian process based speech synthesis using simple recurrent unit," in *Proc. ICASSP*, May 2020, pp. 7249–7253.

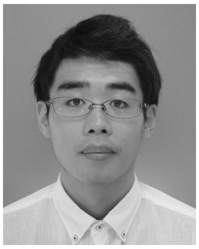
[70] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proc. ICASSP*, May 2020, pp. 7654–7658.

[71] 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. [Online]. Available: https://openreview.net/forum?id=rkgz2aEKDr

[72] T. Raitio, L. Juvela, A. Suni, M. Vainio, and P. Alku, "Phase perception of the glottal excitation and its relevance in statistical parametric speech synthesis," *Speech Commun.*, vol. 81, pp. 104–119, Jul. 2016.

[73] *Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs*, document ITU-T Recommendation P. 830, 1990.

[74] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, and V. Aggarwal, "Towards achieving robust universal neural vocoding," in *Proc. Interspeech*, Sep. 2019, pp. 181–185.



KEISUKE MATSUBARA received the B.E. degree from Kobe University, Japan, in 2020, where he is currently pursuing the master's degree. Since 2019, 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 21st Student Presentation Award from the Acoustical Society of Japan (ASJ), in 2021.

TAKUMA OKAMOTO (Member, IEEE) received the B.E., M.S., and Ph.D. degrees from Tohoku University, Japan, in 2004, 2006, and 2009, respectively. From 2009, he was 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 based on acoustic signal processing and speech synthesis based on neural networks. From 2007 to 2009, he was a Research Fellow (DC2) of the Japan Society for the Promotion of Science. He is a member of the Audio Engineering Society (AES) and ASJ. He received the 32nd Awaya Prize Young Researcher Award and the 57th Sato Prize Paper Award from the Acoustical Society of Japan (ASJ), in 2012 and 2017, 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 had been 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.

TETSUYA TAKIGUCHI (Member, IEEE) received the M.Eng. and Dr.Eng. degrees in information science, Nara. He was a Researcher with the Tokyo Research Laboratory, IBM Research. From 2004 to 2016, he was an Associate Professor with Kobe University. Since 2016, he has been a Professor with Kobe University. 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 the Laboratoire d'InfoRmatique en Image et Systèmes d'information, INSA Lyon. His research interests include speech, image, and brain processing, and multimodal assistive technologies for people with articulation disorders. He is a member of IEICE, IPSJ, and ASJ.

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. His research interests include statistical approaches to speech and audio processing. From 2003 to 2005, he was a Research Fellow of the Japan Society for the Promotion of Science. 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* journal).

YOSHINORI SHIGA received the B.Eng. and M.Eng. degrees from the Tokyo University of Science and the Ph.D. degree from the University of Edinburgh. He has been involved in speech technology research at various institutes, including the Centre for Speech Technology Research, the University of Edinburgh; the Centre for Vision, Speech and Signal Processing, the University of Surrey; the Spoken Language Communication Research Laboratories, Advanced Telecommunications Research Institute International; and the Advanced Speech Technology Laboratory, the National Institute of Information and Communications Technology, since 1987. He is currently with the School of Engineering, Tokyo Denki University, where he is a Professor of speech and audio signal processing. He received the 2002 Information and Systems Society Excellent Paper Award and the 2014 Best Paper Award from the Institute of Electronics, Information and Communication Engineers, and the 2015 TELECOM System Technology Award from the Telecommunications Advancement Foundation.

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 worked for 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 worked for 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 working 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).