SUBBAND WAVENET WITH OVERLAPPED SINGLE-SIDEBAND FILTERBANKS

Takuma Okamoto[1]*, Kentaro Tachibana*[1]*, Tomoki Toda*[2,1]*, Yoshinori Shiga*[1]*, and Hisashi Kawai*[1]

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

ABSTRACT

Compared with conventional vocoders, deep neural network-based raw audio generative models, such as WaveNet and SampleRNN, can more naturally synthesize speech signals, although the synthesis speed is a problem, especially with high sampling frequency. This paper provides subband WaveNet based on multirate signal processing for high-speed and high-quality synthesis with raw audio generative models. In the training stage, speech waveforms are decomposed and decimated into subband short waveforms with a low sampling rate, and each subband WaveNet network is trained using each subband stream. In the synthesis stage, each generated signal is upsampled and integrated into a fullband speech signal. The results of objective and subjective experiments for unconditional WaveNet with a sampling frequency of 32 kHz indicate that the proposed subband WaveNet with a square-root Hann window-based overlapped 9-channel single-sideband filterbank can realize about four times the synthesis speed and improve the synthesized speech quality more than the conventional fullband WaveNet.

Index Terms— Speech synthesis, WaveNet, subband processing, multirate signal processing, single-sideband filterbank

1. INTRODUCTION

Text to speech synthesis (TTS) is an important technique in multilingual spoken language communications. Due to its flexibility and small footprint, statistical parametric speech synthesis (SPSS) has become the mainstream in TTS [1, 2]. In a conventional SPSS, texts are analyzed as linguistic features from which statistical acoustic models estimate the acoustic features. The estimated acoustic features are then converted into speech waveforms by vocoders. To improve the synthesized speech quality, deep learning-based acoustic models have been investigated [3–8] that outperform conventional hidden Markov model-based schemes [1, 2]. High-quality vocoders [9–12] have also been introduced in SPSS. Although these vocoders are corpus-independent, deep learning-based corpus-dependent data-driven approaches, such as acoustic feature extraction [13], glottal vocoder [14], and power spectrum reconstruction [15] have also been investigated. However, their synthesized speech quality cannot reach natural quality because of the over-smoothing problem in the acoustic models and analysis error and some approximations and assumptions in the vocoders.

To directly model raw speech waveforms from linguistic features without vocoders, mel-cepstral analysis-based approaches were first investigated [16, 17]. This method, which eliminates the assumption that speech waveforms are stationary in an analysis frame, models the phase component by introducing a complex cepstrum. However, unsolved problems remain, such as the modeling error caused by the Gaussian distribution assumption and restrictions on the source-filter model structure.

WaveNet [18], a deep neural network-based raw audio generative approach, was recently proposed. In TTS, WaveNet can directly synthesize raw speech waveforms from linguistic features without vocoders and outperforms state-of-the-art unit selection-based and long term-short memory (LSTM)-based speech synthesis systems [6, 19]. Another raw audio generative model, SampleRNN [20], has also been proposed. Such models also achieve end-to-end speech synthesis from texts to raw speech waveforms [21], and some end-to-end approaches have been investigated, including char2wav [22], Tacotron [23], and Deep Voice [24]. To drive conventional vocoders within a raw audio generative model framework, a WaveNet vocoder has been proposed [25], which directly synthesizes raw speech waveforms from acoustic features.

While vocoders require complicated signal processing for analysis and synthesis with some approximations and assumptions, no signal processing, not even Fourier transform, is employed in raw audio generative approaches, which can directly model the generative probabilities of raw speech waveforms from a speech corpus by neural networks. As a result, analysis error and approximation problems can be solved in conventional vocoders to more naturally synthesize speech waveforms than conventional vocoders.

Although conventional vocoders can synthesize speech waveforms in real time, one problem is the synthesis speed in raw audio generative models, even though parallel computing is available since they sequentially synthesize each sample and feed it back to the network to synthesize the next one. Crucially, TTS systems with high sampling frequencies over 16 kHz, such as 22.05 kHz [6], 24 kHz [23], and 48 kHz [8, 13, 14, 24], were recently investigated for high-quality synthesis. The synthesis speed problem in raw audio generative models is especially severe for such higher sampling frequencies. Deep Voice introduced smaller networks that quickly predict speech waveforms in real time. However, there is a tradeoff between the synthesis speed and the synthesized speech quality [24].

For rapid synthesis while maintaining synthesized speech quality for raw audio generative models, this paper proposes subband WaveNet and introduces multirate signal processing [26, 27]. During training, a speech waveform with length T is decomposed and decimated with factor M into subband short waveforms with length T/M and a low sampling rate, and each subband WaveNet network is trained using each subband stream. In the synthesis, each generated signal is upsampled and integrated into the fullband speech waveform. By introducing a square-root Hann window-based overlapped filterbank, the proposed subband WaveNet can accelerate the synthesis speed M times and improve the synthesized speech quality more than the conventional fullband WaveNet. To confirm the factors of the quality improvement, we conducted objective and subjective experiments. The experimental results clarified that the proposed subband WaveNet can improve WaveNet's prediction accuracy.

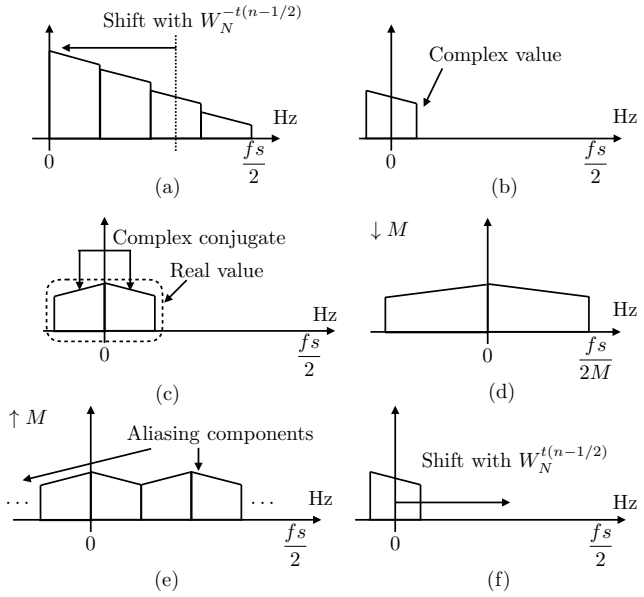


Fig. 1. Multirate signal processing flowchart of a spectrum with a single-sideband filterbank based on polyphase decomposition.

2. WAVENET

WaveNet models conditional probability distribution $p(\boldsymbol{x}|\boldsymbol{h})$ of raw audio waveform $\boldsymbol{x} = [x(1), \cdots, x(T)]$, given additional input $\boldsymbol{h}$, such as linguistic features for TTS [18] and acoustic features for vocoder [25], as

$$p(\boldsymbol{x}|\boldsymbol{h}) = \prod_{t=1}^{T} p(x(t)|x(1), \cdots, x(t-1), \boldsymbol{h}) \tag{1}$$

by a stack of dilated causal convolution layers, which efficiently inputs very long audio samples with a few layers.

In addition, the WaveNet model outputs a categorical distribution instead of a continuous one over next sample $x(t)$ with a softmax layer since it is more flexible and easily models arbitrary distributions, although raw waveform inputs are typically treated as continuous values. In WaveNet, a μ-law companding defined in G. 711 [28] is introduced and raw audio waveforms are quantized to 256 possible values.

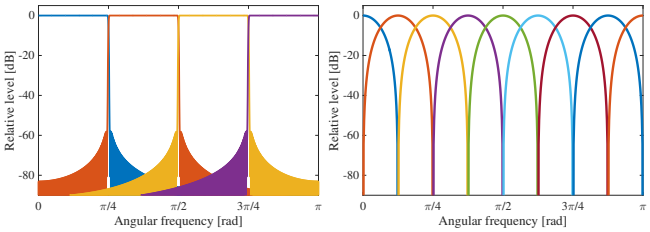
In the training stage, WaveNet can be trained in parallel since all of the $\boldsymbol{x}$ timestamps are known. In the synthesis stage, however, Eq. (1) indicates that WaveNet must sequentially synthesize each sample that is fed back to the network to synthesize the next one. Therefore, the synthesis speed problem in raw audio generative models is not solved, even though parallel computing is available.

3. PROPOSED SUBBAND WAVENET

3.1. Multirate signal processing with single-sideband filterbanks

To decompose original fullband waveform $x(t)$ into N subband streams, $x(t)$ is modulated by $W_N^{-t(n-1/2)}$ and shifted to the baseband frequency (Fig. 1(a)):

$$x_n(t) = x(t) W_N^{-t(n-1/2)}, \tag{2}$$



(a) Maximally decimated (b) Overlapped

Fig. 2. Division examples: (a) maximally decimated with $N = M = 4$ and (b) overlapped with $N = 9$ and $M = 4$ single-sideband filterbanks.

where $W_N = \exp(j2\pi/2N)$ and $n = 1, 2, , \cdots, N$. In standard polyphase decomposition, $x_m(t)$ is bandlimited by prototype lowpass analysis filter $h(t)$ whose cutoff frequency is set to $\pi/2N$:

$$x_{n,\mathrm{pp}}(t) = f(t) * x_n(t), \tag{3}$$

where $*$ denotes the convolution operator and $x_{n,\mathrm{pp}}(t)$ is obtained as a complex value (Fig. 1(b)). To treat real values in WaveNet, single-sideband (SSB) modulation [26] is introduced. Real-valued signal $x_{n,\mathrm{SSB}}(t)$ is then obtained:

$$x_{n,\mathrm{SSB}}(t) = x_{n,\mathrm{pp}}(t) W_N^{t/2} + x_{n,\mathrm{pp}}^*(t) W_N^{-t/2}, \tag{4}$$

where $x_{n,\mathrm{pp}}^*(t)$ is the complex conjugate of $x_{n,\mathrm{pp}}(t)$ (Fig. 1(c)). $x_{n,\mathrm{SSB}}(t)$ is decimated by factor M, and the n-th subband waveform is obtained:

$$x_n(k) = x_{n,\mathrm{SSB}}(kM). \tag{5}$$

By Eqs. (2) to (5), fullband waveform $x(t)$ with length T and sampling frequency fs is decomposed into N subband short waveforms with length T/M and sampling frequency fs/M (Fig. 1(d)). In multirate signal processing, each subband stream $x_n(k)$ with a short length and a low sampling rate can be separately and efficiently processed into $\hat{x}_n(k)$.

In the synthesis process, each subband waveform $\hat{x}_n(k)$ is upsampled by factor M and represented:

$$\hat{x}_{n,\mathrm{SSB}}(t) = \begin{cases} \hat{x}_n(t/M), & \text{for } t = 0,\ M,\ 2M, \cdots \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where the signal length and sampling frequency are respectively restored to T and fs. As upsampling artifact, undesired aliasing components additionally occur (Fig. 1(e)). To reduce them, $\hat{x}_{m,\mathrm{SSB}}(t)$ is shifted to the baseband and bandlimited by prototype lowpass synthesis filter $g(t)$ whose cutoff frequency is also set to $\pi/2N$:

$$\hat{x}_{n,\mathrm{pp}}(t) = g(t) * \hat{x}_{n,\mathrm{SSB}}(t) W_N^{-t/2}. \tag{7}$$

Fullband reconstructed waveform $\hat{x}(t)$ is finally integrated (Fig. 1(f)) and obtained:

$$\hat{x}(t) = \sum_{n=1}^{N} \hat{x}_{n,\mathrm{pp}}(t) W_N^{t(n-1/2)} + \hat{x}_{n,\mathrm{pp}}^*(t) W_N^{-t(n-1/2)}. \tag{8}$$

When $N = M$, the most efficient processing is achieved, which is called maximally decimated filterbank [27]. A division example of a maximally decimated SSB filterbank with $N = M = 4$ and a simple lowpass prototype analysis filter are plotted in Fig. 2(a).

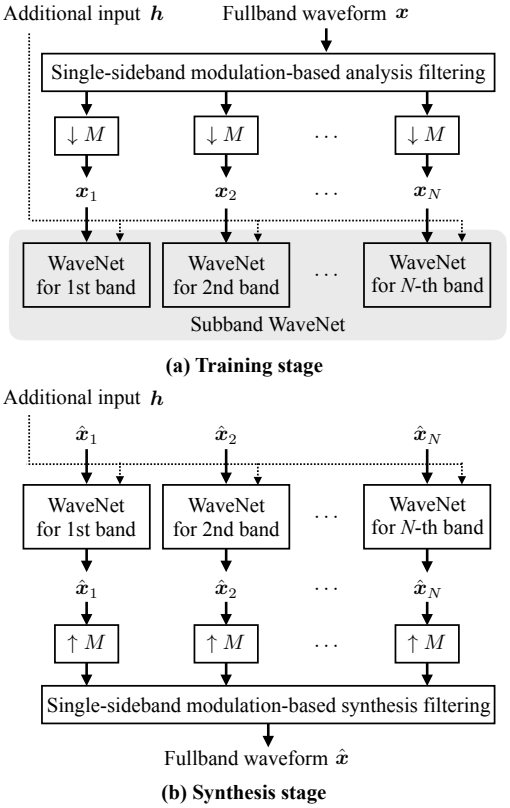


Fig. 3. Block diagram of proposed subband WaveNet.

3.2. Overlapped single-sideband filterbanks

Even though maximally decimated filterbanks can efficiently decompose fullband waveforms into subband short ones with a low sampling rate, the proposed subband WaveNet with a maximally decimated filterbank cannot synthesize high-quality speech waveforms. The reason is described in Section 4. To improve the synthesis quality of the proposed subband WaveNet, one-half overlapped SSB filterbanks are introduced where $N = 2M + 1$. In these filterbanks, $W_N = \exp(j2\pi/(N-1))$, the cutoff frequency of $h(t)$ and $g(t)$ is set to $\pi/(N-1)$, and Eqs. (2) and (8) are respectively superseded:

$$x_n(t) = x(t)W_N^{-tn/2}, \tag{9}$$

$$\hat{x}(t) = \sum_{n=0}^{N-1} \hat{x}_{n,\mathrm{pp}}(t)W_N^{tn/2} + \hat{x}^*_{n,\mathrm{pp}}(t)W_N^{-tn/2}, \tag{10}$$

where $n = 0,\ 1,\ 2,\ , \cdots,\ N-1$ and

$$x_{n,\mathrm{SSB}}(t) = x_{n,\mathrm{pp}}(t), \tag{11}$$

$$\hat{x}_{n,\mathrm{pp}}(t) = g(t) * \hat{x}_{n,\mathrm{SSB}}(t) \tag{12}$$

for $n = 0$ and $N-1$, which are respectively introduced instead of Eqs. (4) and (7), since $x_{0,\mathrm{pp}}(t)$ and $x_{N-1,\mathrm{pp}}(t)$ are the original real values. A division example of a one-half overlapped SSB filterbank with $M = 4$, $N = 9$, and a Hann window-based prototype analysis filter are plotted in Fig. 2(b).

3.3. Subband WaveNet

A block diagram of the proposed subband WaveNet is described in Fig. 3. In the training stage, fullband speech waveforms $\boldsymbol{x} = [x(1),\ \cdots,\ x(T)]$ in the training set are decomposed into N subband streams $\boldsymbol{x}_n = [x_n(1),\ \cdots,\ x_n(T/M)]$ with short length T/M and low sampling frequency fs/M by maximally decimated or overlapped SSB analysis filterbanks, given as Eqs. (2) to (5), (9) and (10). Each subband WaveNet network $P_n(\boldsymbol{x}_n|\boldsymbol{h})$ is then separately and efficiently trained by each subband waveform $\boldsymbol{x}_n$ with additional input $\boldsymbol{h}$. In the synthesis stage, each subband stream $\hat{\boldsymbol{x}}_n = [\hat{x}_n(1),\ \cdots,\ \hat{x}_n(T/M)]$ is simultaneously generated by the trained network and integrated into fullband waveform $\hat{\boldsymbol{x}} = [\hat{x}(1),\ \cdots,\ \hat{x}(T)]$ by the SSB synthesis filterbanks, given as Eqs. (6), (7), (11), and (12).

Compared with the conventional fullband WaveNet, the proposed subband WaveNet can synthesize N samples at one time and realize M times the synthesis speed with parallel computing.

Table 1. Results of synthesis time for one female and one male speech waveforms with lengths 3.85 and 3.78 s.

	Female		Male	
fs (kHz)	16	32	16	32
Fullband	5.37 m	11.40 m	5.37 m	11.21 m
Subband	1.40 m	2.68 m	1.47 m	2.65 m

4. EXPERIMENTS

4.1. Experimental conditions

4.1.1. Speech corpora

To evaluate the proposed subband WaveNet's effectiveness, we conducted objective and subjective experiments using Japanese female and male speech corpora recorded with a sampling frequency of 48 kHz and downsampled to 16 and 32 kHz. In the female speech synthesis, 7242 (about 4.8 hours) and 100 utterances were used as training and test sets. In the male speech synthesis, 5697 (about 3.7 hours) and 100 utterances were used as training and test sets.

4.1.2. Filterbank condition

As a first investigation of the proposed subband WaveNet, fixed decimation factor $M = 4$ and division numbers $N = 4$ and 9 were respectively set for maximally decimated and overlapped SSB filterbanks. In the experiments, the following maximally decimated (MD) and two overlapped (OL) SSB filterbanks were investigated.

LPF-MD: In the maximally decimated SSB filterbank, a simple lowpass FIR filter with a length of 1025 samples calculated from a sinc function with a Hamming window [29] was introduced as analysis and synthesis filters, $h(t)$ and $g(t)$. Its frequency response is plotted in Fig. 2(a).

LPF-OL: To examine the overlap effect, we introduced an overlapped 9-channel SSB filterbank with identical analysis-synthesis filters as LPF-MD where the reconstructed waveforms were divided by 2.

SQRT-Hann-OL: To investigate the analysis-synthesis filters with smoother frequency response than the above simple bandpass filter, a square-root Hann window-based filter was proposed and its frequency response $H(\omega)(= G(\omega))$ is given as

$$H(\omega) = \begin{cases} \sqrt{\cos(\frac{N-1}{2}\omega)} & \text{for } -\frac{\pi}{N-1} \le \omega \le \frac{\pi}{N-1} \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

Table 2. Results of objective evaluations of 100 test set utterances.

	Method	(A) SNR [dB] (analysis-synthesis)	(B) SNR [dB] (μ-law quantization)	(C) SNR [dB] (WaveNet)	(D) SD [dB] (WaveNet)	(E) MCD [dB] (WaveNet)
Female $fs = 16$ kHz	Fullband	-	30.0 ± 0.12	$\mathbf{20.4 \pm 0.28}$	7.58 ± 0.05	2.20 ± 0.03
	LPF-MD	46.3 ± 0.32	25.9 ± 0.16	7.0 ± 0.28	7.11 ± 0.07	2.59 ± 0.04
	LPF-OL	41.1 ± 0.35	26.8 ± 0.14	7.3 ± 0.20	7.01 ± 0.05	2.00 ± 0.03
	SQRT-Hann-OL	$\mathbf{76.2 \pm 0.11}$	$\mathbf{30.2 \pm 0.14}$	7.4 ± 0.20	$\mathbf{5.96 \pm 0.08}$	$\mathbf{1.62 \pm 0.03}$
Female $fs = 32$ kHz	Fullband	-	$\mathbf{30.1 \pm 0.13}$	$\mathbf{21.5 \pm 0.35}$	8.72 ± 0.07	2.45 ± 0.03
	LPF-MD	49.8 ± 0.27	26.7 ± 0.16	13.6 ± 0.28	7.68 ± 0.07	2.80 ± 0.04
	LPF-OL	46.6 ± 0.33	27.5 ± 0.13	13.1 ± 0.22	7.16 ± 0.05	1.95 ± 0.04
	SQRT-Hann-OL	$\mathbf{77.4 \pm 0.09}$	29.5 ± 0.13	13.0 ± 0.27	$\mathbf{6.45 \pm 0.09}$	$\mathbf{1.72 \pm 0.04}$
Male $fs = 16$ kHz	Fullband	-	$\mathbf{26.4 \pm 0.27}$	$\mathbf{20.5 \pm 0.21}$	8.80 ± 0.07	2.01 ± 0.03
	LPF-MD	43.1 ± 0.33	21.5 ± 0.33	6.80 ± 0.29	7.74 ± 0.09	2.60 ± 0.06
	LPF-OL	38.5 ± 0.33	23.2 ± 0.27	7.80 ± 0.07	7.47 ± 0.07	2.18 ± 0.24
	SQRT-Hann-OL	$\mathbf{74.0 \pm 0.19}$	25.1 ± 0.36	8.00 ± 0.22	$\mathbf{6.65 \pm 0.10}$	$\mathbf{1.77 \pm 0.05}$
Male $fs = 32$ kHz	Fullband	-	$\mathbf{26.6 \pm 0.27}$	$\mathbf{23.2 \pm 0.26}$	9.51 ± 0.08	2.75 ± 0.03
	LPF-MD	49.4 ± 0.30	22.5 ± 0.34	14.0 ± 0.25	8.17 ± 0.08	2.75 ± 0.05
	LPF-OL	43.4 ± 0.36	23.7 ± 0.28	13.4 ± 0.26	7.48 ± 0.06	2.10 ± 0.04
	SQRT-Hann-OL	$\mathbf{76.3 \pm 0.11}$	24.5 ± 0.37	13.6 ± 0.31	$\mathbf{7.26 \pm 0.10}$	$\mathbf{2.07 \pm 0.06}$

where ω is the angular frequency. $h(t)$ and $g(t)$ with a length of 1024 were numerically obtained from the inverse discrete Fourier transform. In the filterbank, the total frequency response of $h(t) * g(t)$ is equivalent to a Hann window (Fig. 2(b)), and fullband waveforms were perfectly reconstructed by Eq. (10) since the sum of the one-half overlapped Hann windows is just 1 at all the frequency components.

4.1.3. WaveNet condition

To investigate the fundamental performance of the proposed subband WaveNet, unconditional WaveNet training and synthesis without additional input $\boldsymbol{h}$ were conducted in this paper. To evaluate the test set speech waveforms generated by the unconditional WaveNet, each estimated sample $\hat{x}(t)$ was generated with original past samples $[x(1), \cdots, x(t-1)]$, and the estimated waveform was integrated as $\hat{\boldsymbol{x}} = [\hat{x}(1), \cdots, \hat{x}(T)]$.

In a conventional fullband WaveNet with a sampling frequency of 16 kHz, 30 dilated causal convolution layers were introduced as 1, 2, 4, $\cdots$, 512, 1, 2, 4, $\cdots$, 512, 1, 2, 4, $\cdots$, 512 [18, 25], whose receptive field length was $1024 \times 3/16000 = 0.192$ s. In a sampling frequency of 32 kHz, 33 layers as $\{1, 2, 4, \cdots, 1024\} \times 3$ were introduced to maintain the same receptive field length of 0.192 s.

In the subband WaveNet, the fullband speech waveforms with sampling frequencies of 16 and 32 kHz were decimated with $M = 4$, and the sampling frequencies of the decomposed waveforms were 4 and 8 kHz. In these cases, 24 causal convolution layers as $\{1, 2, 4, \cdots, 128\} \times 3$ for 4 kHz and 27 layers as $\{1, 2, 4, \cdots, 256\} \times 3$ for 8 kHz were respectively employed to cover the same receptive field length as the fullband WaveNet.

The mini-batch sizes with sampling frequencies of 32, 16, 8, and 4 kHz were respectively 80 k, 40 k, 20 k, and 10 k samples ($= 2.5$ s). In both the fullband and subband WaveNet, the dilation channels, the residual channels, and number of skip connections were set to 32, 32, and 512. An Adam optimization algorithm [30] updated the neural network parameters with a learning rate of 0.001 as an initial value that was multiplied by 0.5 at all 50 k parameter updates. The numbers of parameter updates for the fullband and subband WaveNet were 200 k and 100 k. Each WaveNet was trained using an Intel Xeon(R) CPU E5-2670 and a single GPU of NVIDIA GeForce GTX 1080.

4.2. Synthesis speed

To compare the synthesis speed of the fullband and subband WaveNet, the synthesis time results for one female and one male speech waveforms with lengths of 3.85 and 3.78 s synthesized by an Intel Xeon(R) CPU E7-8837 are presented in Table 1. In the experiments, a fast generation algorithm was introduced [31]. In the subband WaveNet, the synthesis speed of each subband was the same since the number of network parameters was identical to each band. The results indicate that the proposed subband WaveNet with a decimation factor of $N = 4$ successfully realized about four times the synthesis speed compared with the fullband WaveNet when parallel computing is available.

4.3. Objective evaluations

To objectively evaluate the test set speech waveforms synthesized by fullband and subband WaveNet, we introduced and defined the signal-to-noise ratio (SNR) and the spectral distortion (SD) between original waveform $x(t)$ and synthesized $\hat{x}(t)$:

$$SNR = 10 \log_{10} \left(\frac{\sum_{t=1}^{T} \hat{x}(t)^2}{\sum_{t=1}^{T} (x(t) - \hat{x}(t))^2} \right), \tag{14}$$

$$SD = \frac{1}{A} \sum_{a=1}^{A} \sqrt{\frac{1}{F} \sum_{f=1}^{F} \left(20 \log_{10} \frac{|\hat{X}(f,a)|}{|X(f,a)|} \right)^2}, \tag{15}$$

where $X(f,a)$ and $\hat{X}(f,a)$ are the short-time Fourier spectrums of $x(t)$ and $\hat{x}(t)$ in frame a for frequency bin f and A is the total frame number. The short-time Fourier transform analysis window function was a Hann window with a frame length of $16fs/1000$ samples ($=$ 16 ms) and a frameshift of $fs/1000$ samples ($=$ 1 ms). To consider

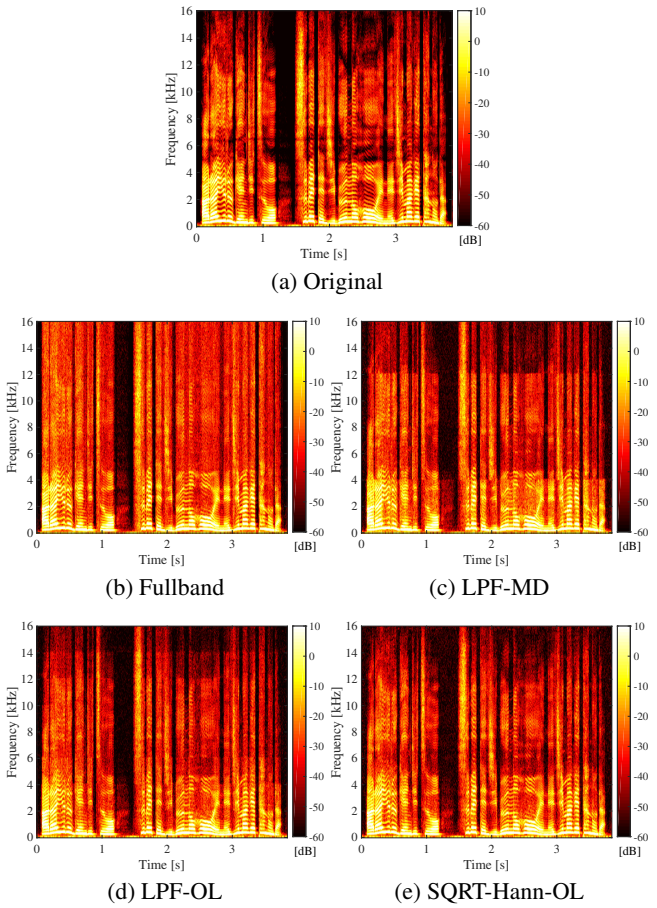


Fig. 4. Spectrograms: (a) test set female original speech waveform with a sampling frequency of 32 kHz, (b) estimated by fullband WaveNet, (c) estimated by subband WaveNet with LPF-MD, (d) estimated by subband WaveNet with LPF-OL, and (e) estimated by subband WaveNet with SQRT-Hann-OL.

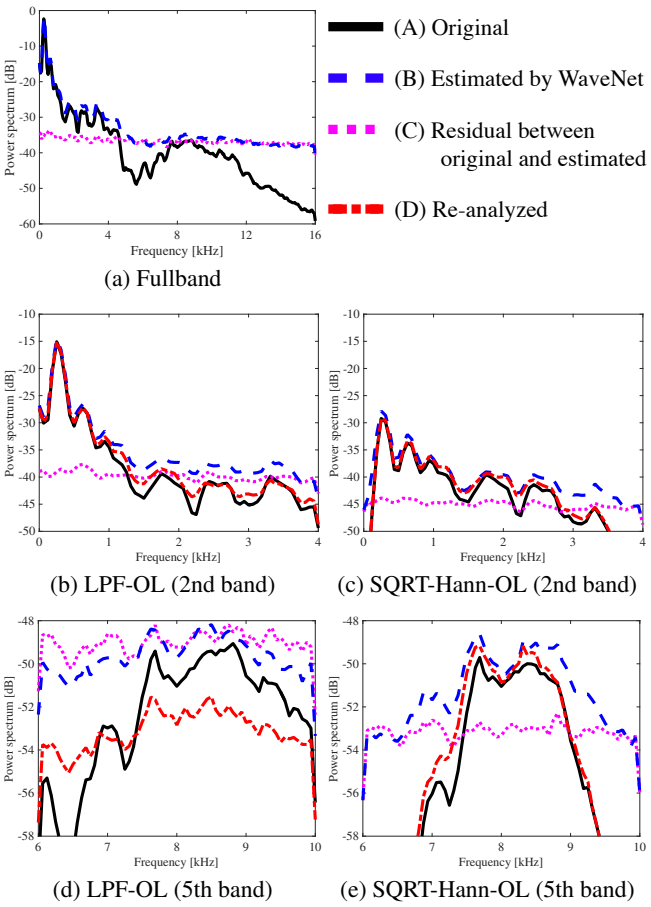


Fig. 5. Results of power spectrum: (A) test set of female original waveforms $x(t)$ and $x_n(k)$, (B) waveforms estimated by WaveNet $\hat{x}(t)$ and $\hat{x}_n(k)$, (C) residuals between original and estimated waveforms $x(t) - \hat{x}(t)$ and $x_n(k) - \hat{x}_n(k)$, and (D) subband waveforms obtained from re-analysis of $\hat{x}(t)$ by Eqs. (9) and (3) to (5).

the human auditory perception criterion in the objective evaluation, mel-cepstral distortion (MCD) was introduced and defined:

$$MCD = \frac{10}{\log 10}\sqrt{2\sum_{b=1}^{B}(c(b) - \hat{c}(b))^2}, \tag{16}$$

where $c(b)$ and $\hat{c}(b)$ are the b-th mel-cepstral coefficients [32] obtained from $X(f, a)$ and $\hat{X}(f, a)$ with a frame length of $fs/40$ samples ($= 25$ ms) and a frameshift of $fs/200$ samples ($= 5$ ms). In a sampling frequency of 16 kHz, warping coefficient α was 0.42, and $c(b)$ and $\hat{c}(b)$ were calculated to $B = 24$. In a sampling frequency of 32 kHz, the original and synthesized waveforms were upsampled to 48 kHz, and $\alpha = 0.55$ and $B = 60$ [8] were employed.

Table 2(A) shows the SNR results for evaluating the analysis-synthesis error with filterbanks. Although LPF-MD and LPF-OL with a simple lowpass filter reconstructed the original waveforms with an SNR of about 40 dB, the proposed square-root Hann window-based filter almost perfectly reconstructed the original waveforms with an SNR over 74 dB since the latter's frequency response is smoother and easier to implement with a finite length of digital filter. The SNR results for evaluating the μ-law quantization are described in Table 2(B). In the subband methods, the original waveforms were first decomposed into subband streams, and the reconstructed waveforms were calculated from the μ-law quantized subband streams. The result indicates that fullband processing can reconstruct original waveforms with lower error than subband processing.

Tables 2(C) to (E) show the SNR, SD, and MCD results defined in Eqs. (14) to (16). Although the fullband WaveNet outperformed the subband WaveNet in SNR, the proposed subband WaveNet with SQRT-Hann-OL achieved a better SD and MCD than the other methods. Eq. (1) suggests that WaveNet is trained to maximize SNR in the time domain, but SD and MCD cannot be optimized. The results in Table 2 suggest that the SNR in the proposed subband WaveNet is lower than that of the fullband WaveNet since each estimated subband waveform includes error and the phase component of the fullband waveform cannot be correctly reconstructed. The detailed analysis of the SD and MCD improvements in the proposed method is described in the next subsection. We compared the fullband Wavenet and the proposed subband WaveNet and evaluated the subjective evaluation results in Section 4.5.

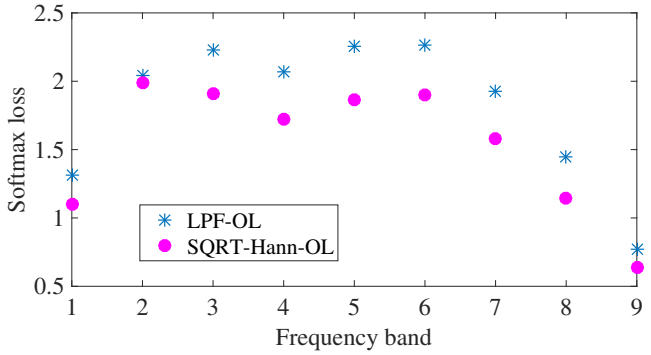


Fig. 6. Results of softmax loss averaged in last 1 k out of 100 k training times for female training set with a sampling frequency of 32 kHz.

Table 3. Results of paired comparison listening test with a sampling frequency of 32 kHz with 21 adult Japanese native speakers. C, P, and N are answers of conventional fullband WaveNet, proposed subband WaveNet with SQRT-Hann-OL, and neutral.

	C	P	N	p-value	Z-score
Female (%)	33 (6.3)	**413** (**78.7**)	79 (15.0)	$\ll 10^{-10}$	-18.0
Male (%)	31 (5.9)	**439** (**83.6**)	55 (10.5)	$\ll 10^{-10}$	-18.8

4.4. Analysis of effectiveness of proposed subband WaveNet

Figure 4 depicts the spectrograms of a female original speech waveform test set and those estimated by fullband and subband WaveNet with a sampling frequency of 32 kHz. Fig. 5(a) shows the following results of the averaged power spectrum: (A) test set of female original waveform $x(t)$, (B) that estimated by WaveNet $\hat{x}(t)$, and (C) residuals between original and estimated waveform $x(t) - \hat{x}(t)$ for fullband WaveNet. Furthermore, Figs. 5(b) to (e) plot the 2nd and 5th subbands for the subband WaveNet with overlapped filterbanks and (D) subband waveforms obtained from re-analysis of $\hat{x}(t)$ by Eqs. (9) and (3) to (5).

In the unconditional WaveNet, each sample $\hat{x}(t)$ (and $\hat{x}_n(k)$) is independently estimated at every t (and k), which causes random noise in $\hat{x}(t)$ and $\hat{x}_n(k)$, as found in the results of (C) in Figs. 5(a) to (e). As shown in Figs. 4(b) and 5(a), the random noise level is higher (over 4 kHz) than the original waveform, and this degrades SD and MCD in the fullband WaveNet.

In the LPF-OL, an overlap effect can be found in the lower bands, as depicted in Figs. 5(b)-(B) and (D); LPF-OL outperforms LPF-MD in SD and MCD. However, the result in Fig. 5(d) indicates that LPF-MD and LPF-OL cannot be adequately trained at higher bands, and the residual level is higher than the original subband waveform since its frequency response is almost "white." As a result, the overlap effect no longer works at higher bands in LPF-OL, and the reconstructed fullband spectrum has "striped" noise components (Figs. 4(c) and (d)).

Although the same frequency bands were used in the subband WaveNet with LPF-OL and SQRT-Hann-OL, the latter's spectrogram (Fig. 4(e)) is smoother, and its subband waveforms at higher bands were adequately estimated, as shown in Fig. 5(e)-(D). The difference between them is just analysis-synthesis filters $h(t)$ and $g(t)$. Fig. 6 plots the results of the softmax loss averaged in the last 1 k out of 100 k training times for the female training set with a sampling frequency of 32 kHz for LPF-OL and SQRT-Hann-OL. This result suggests that the proposed square-root Hann window-based analysis filter improved the prediction accuracy and simplified the estimation of the subband waveforms by WaveNet more than by the simple bandpass filter since the subband waveforms are forcibly modulated to the "colored" signals by the proposed filter (Fig. 5(e)-(A)). As a result, the proposed subband WaveNet with SQRT-Hann-OL generated speech waveforms with lower SD and MCD values throughout all the frequency bands than the other methods.

4.5. Subjective evaluations

For a subjective evaluation, we also conducted a paired comparison listening test between the conventional fullband WaveNet and the proposed subband WaveNet with a square-root Hann window-based overlapped single-sideband filterbank. 50 (female: 25 and male: 25) utterances out of the test set with a sampling frequency of 32 kHz were used as the evaluation speech set and presented by headphones. The listening subjects were 21 Japanese adult native speakers without hearing loss. The speech generated by the conventional fullband WaveNet and the proposed subband WaveNet of the test set utterances was continuously presented in a random order with a space of 0.25 s, and the subjects were allowed to freely re-listen many times. They compared and judged the quality of the two stimuli. To allow for cases where they were unable to judge between the two stimuli, we introduced an additional answer: *neutral*. The listening test result is depicted in Table 3. Statistical analysis of the result indicates that the proposed subband WaveNet significantly improved the synthesized speech quality.

Consequently, the proposed subband WaveNet with a square-root Hann window-based overlapped 9-channel SSB filterbank not only accelerates the synthesis speed by about four times but also improves the synthesized speech quality more than the conventional fullband WaveNet.

5. FUTURE WORK

We must find optimal decimation factors and analysis-synthesis filters for rapid and high-quality speech synthesis. The optimal quantization method and the number of network parameters for each subband are investigated since the frequency responses and signal levels are different for each subband. In addition, a method to simultaneously input and output all subband waveforms with a single network is investigated to reduce the computational cost and eliminate parallel computing. In the next step, the proposed subband WaveNet will be applied to TTS [18] and WaveNet vocoder [25] with the conditional WaveNet instead of the unconditional one with previous correct samples. Furthermore, the proposed method can be directly applied to other raw audio generative models, such as SampleRNN [20] and Tacotron [23].

6. CONCLUSIONS

This paper proposed subband WaveNet by introducing multirate signal processing for rapid and high-quality synthesis with raw audio generative models. The results of objective and subjective experiments for unconditional WaveNet suggested that the proposed subband WaveNet with a square-root Hann window-based overlapped 9-channel SSB filterbank increased the synthesis speed about four times and improved the synthesized speech quality more than the conventional fullband WaveNet.

7. REFERENCES

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[2] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, May 2013, pp. 7962–7966.

[4] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. Interspeech*, Sept. 2014, pp. 1964–1968.

[5] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, May 2015.

[6] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices," in *Proc. Interspeech*, Sept. 2016, pp. 2273–2277.

[7] K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "Model integration for HMM- and DNN-based speech synthesis using product-of-experts framework," in *Proc. Interspeech*, Sept. 2016, pp. 2288–2292.

[8] 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 9*, Sept. 2016, pp. 125–128.

[9] 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, no. 3–4, pp. 187–207, Apr. 1999.

[10] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE J. Sel. Topics Signal. Process.*, vol. 8, no. 2, pp. 184–194, Apr. 2014.

[11] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *Proc. ICASSP*, Apr. 2015, pp. 4230–4234.

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

[13] S. Takaki and J. Yamagishi, "A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis," in *Proc. ICASSP*, Mar. 2016, pp. 5535–5539.

[14] 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*, Sept. 2016, pp. 2473–2477.

[15] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "Deep neural network-based power spectrum reconstruction to improve quality of vocoded speech with limited acoustic parameters," *Acoust. Sci. Tech.*, (accepted, in press).

[16] K. Tokuda and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in *Proc. ICASSP*, Apr. 2015, pp. 4215–4219.

[17] K. Tokuda and H. Zen, "Directly modeling voiced and unvoiced components in speech waveforms by neural networks," in *Proc. ICASSP*, Mar. 2016, pp. 5640–5644.

[18] 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," *arXiv preprint arXiv:1609.03499*, Sept. 2016, (unreviewed manuscript).

[19] X. Gonzalvo, S. Tazari, C. an Chan, M. Becker, A. Gutkin, and H. Silen, "Recent advances in google real-time HMM-driven unit selection synthesizer," in *Proc. Interspeech*, Sept. 2016, pp. 2238–2242.

[20] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. ICLR*, Apr. 2017.

[21] W. Wang, S. Xu, and B. Xu, "First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention," in *Proc. Interspeech*, Sept. 2016, pp. 2243–2247.

[22] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *Proc. ICLR*, Apr. 2017.

[23] Y. Wang, RJ Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Aug. 2017, pp. 4006–4010.

[24] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," in *Proc. ICML*, Aug. 2017, pp. 195–204.

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

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

[27] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, Upper Saddle River, 1993.

[28] ITU-T. Recommendation G. 711, *Pulse Code Modulation (PCM) of voice frequencies*, 1988.

[29] Speech Digital Signal Processing Committee of the IEEE Acoustics and Signal Processing Society, Eds., *Programs for Digital Signal Processing*, New York, IEEE Press, 1979.

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

[31] P. Ramachandran, T. L. Paine, P. Khorrami, M. Babaeizadeh, S. Chang, Y. Zhang, M. Hasegawa-Johnson, R. Campbell, and T. Huang, "Fast generation for convolutional autoregressive models," in *Proc. ICLR*, Apr. 2017.

[32] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, Mar. 1992, vol. 1, pp. 137–140.