AN INVESTIGATION OF SUBBAND WAVENET VOCODER COVERING ENTIRE AUDIBLE FREQUENCY RANGE WITH LIMITED ACOUSTIC FEATURES

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

Although a WaveNet vocoder can synthesize more natural-sounding speech waveforms than conventional vocoders with sampling frequencies of 16 and 24 kHz, it is difficult to directly extend the sampling frequency to 48 kHz to cover the entire human audible frequency range for higher-quality synthesis because the model size becomes too large to train with a consumer GPU. For a WaveNet vocoder with a sampling frequency of 48 kHz with a consumer GPU, this paper introduces a subband WaveNet architecture to a speaker-dependent WaveNet vocoder and proposes a subband WaveNet vocoder. In experiments, each conditional subband WaveNet with a sampling frequency of 8 kHz was well trained using a consumer GPU. The results of subjective evaluations with a Japanese male speech corpus indicate that the proposed subband WaveNet vocoder with 36-dimensional simple acoustic features significantly outperformed the conventional source-filter model-based vocoders including STRAIGHT with 86-dimensional features.

Index Terms— Speech synthesis, vocoder, subband WaveNet, multirate signal processing, entire audible frequency range.

1. INTRODUCTION

In conventional statistical parametric speech synthesis (SPSS) [1] and voice conversion (VC) [2], source-filter model-based vocoders are typically introduced to synthesize speech waveforms from estimated and converted acoustic features that are mainly constructed from the fundamental frequency and vocal tract spectrums. To improve the synthesized speech quality in conventional SPSS and VC, sophisticated vocoders [3,4] have been introduced instead of a basic mel-log spectrum approximate (MLSA) filter with a simple pulse excitation and cepstrum [5]. Although these vocoders are corpus-independent, deep learning-based corpus-dependent data-driven approaches, such as acoustic feature extraction [6], glottal vocoder [7], and power spectrum reconstruction for vocoded speech [8] have also been investigated as in deep learning-based acoustic models for SPSS and VC [9]. However, their synthesized speech quality cannot reach natural quality because of analysis error and approximations and assumptions in the vocoders.

WaveNet [10,11], a deep neural network-based raw audio generative approach, was recently proposed. In text-to-speech synthesis, WaveNet can directly synthesize raw speech waveforms from linguistic features and outperform state-of-the-art unit selection-based and SPSS-based speech synthesis systems with sampling frequencies of 16 and 24 kHz [11]. Another raw audio generative model, SampleRNN [12], has also been proposed. Such models can realize end-to-end speech synthesis from texts to raw speech waveforms, such as Char2Wav [13], Deep Voice [14–16], and Tacotron 2 [17].

In addition, to fuel conventional source-filter model-based vocoders within a raw audio generative model framework, a WaveNet vocoder has been proposed [18], which directly synthesizes raw speech waveforms from acoustic features, and applied to a conventional VC framework [19]. By introducing a noise shaping method [20], it outperformed the conventional STRAIGHT vocoder [3] with a sampling frequency of 16 kHz [21].

For high-quality synthesis, SPSS systems with a sampling frequency of 48 kHz covering the entire human auditory frequency range were recently investigated [6,7,14,16,22–24]. Although Deep Voice first implemented a conditional raw audio generative model with a sampling frequency of 48 kHz by introducing smaller networks than vanilla WaveNet, there is a tradeoff between the model size and the synthesized speech quality [14]. To train larger models for high-quality synthesis, GPUs with quite large size memory are required. Therefore, a consumer GPU runs out of memory and no longer trains conditional WaveNet models for high-quality synthesis with a sampling frequency of 48 kHz. Although Deep Voice 3 implemented a WaveNet vocoder with a sampling frequency of 48 kHz, the network model size and parameters were not disclosed [16]. Therefore, directly extending the conventional WaveNet vocoder is difficult with sampling frequencies of 16 and 24 to 48 kHz.

For rapid synthesis and improving synthesized speech quality, a subband WaveNet architecture based on multirate signal processing [25] was proposed [26]. By introducing a square-root Hann window-based overlapped single-sideband (SSB) filterbank, the proposed subband WaveNet can accelerate the synthesis speed and improve the synthesized speech quality more than the fullband WaveNet since it can improve the prediction accuracy of WaveNet. Although the effectiveness of the subband WaveNet was only validated in experiments for unconditional WaveNet and the synthesis speed problem was solved by Parallel WaveNet [11], the subband architecture is expected to train conditional WaveNet models with a sampling frequency of 48 kHz using a consumer GPU.

To confirm the availability of the proposed subband architecture in conditional WaveNet and achieve a WaveNet vocoder with a sampling frequency of 48 kHz with a consumer GPU, this paper investigates a speaker-dependent subband WaveNet vocoder that covers the entire human audible frequency range for high-quality synthesis. To easily apply the proposed subband WaveNet vocoder to the existing SPSS and VC systems, we introduced lower-dimensional acoustic features constructed from the fundamental frequency and simple mel-cepstral coefficients rather than the mel-spectrograms used in Deep Voice 3 [16] and Tacotron 2 [17] and higher-dimensional STRAIGHT- and WORLD-based features [13, 22–24]. We also investigated how the proposed subband WaveNet vocoder with a sampling frequency of 48 kHz synthesizes speech waveforms us-

∗ K. Tachibana is currently with the DeNA Co., Ltd., Japan.

ing acoustic features with such lower sampling frequencies as 16 and 8 kHz to explore the possibility of bandwidth extension. This investigation will be useful when a speaker-independent subband WaveNet vocoder can be realized.

2. SPEAKER-DEPENDENT SUBBAND WAVENET VOCODER

2.1. WaveNet vocoder

A WaveNet vocoder [18,21] models conditional probability distribution $p(\boldsymbol{x}|\boldsymbol{h})$ of raw audio waveform $\boldsymbol{x} = [x(1), \cdots, x(T)]$, given acoustic features $\boldsymbol{h}$, 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. 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 a vanilla WaveNet, a μ-law companding defined in G. 711 [27] is introduced and raw audio waveforms are quantized to 256 possible values.

Acoustic features for vocoders in SPSS and VC are typically analyzed every 5 ms. The time resolution adjustment between speech waveform $\boldsymbol{x}$ and acoustic features $\boldsymbol{h}$ is then required. In WaveNet vocoder, a simple approach to match both sequence lengths of $\boldsymbol{x}$ and $\boldsymbol{h}$ is performed by copying $\boldsymbol{h}$ of each frame by the shift amount of the analysis window [18, 21].

2.2. Proposed subband WaveNet vocoder

As in a previous subband WaveNet [26], a block diagram of the proposed subband WaveNet vocoder is described in Fig. 1. In the training stage, fullband speech waveforms $\boldsymbol{x} = [x(1), \cdots, x(T)]$ with a sampling frequency of fs in the training set are decimated by factor M and 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 an overlapped SSB analysis filterbank. 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 common acoustic features $\boldsymbol{h}$. 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 upsampled by M, and each subband waveform with a sampling frequency of fs is obtained by an overlapped SSB synthesis filterbank.

In previous experiments on unconditional subband WaveNet synthesis [26], each estimated sample $\hat{x}_n(t)$ was generated with original past samples $[x_n(1), \cdots, x_n(t-1)]$, and the phase shift between subbands was not a problem. In the conditional subband WaveNet, on the other hand, a problem does exist since each estimated sample $\hat{x}_n(t)$ is generated from already estimated past samples $[\hat{x}_n(1), \cdots, \hat{x}_n(t-1)]$ and acoustic features $\boldsymbol{h}$ with random sampling based on $p_n(\boldsymbol{x}_n|\boldsymbol{h})$. To compensate the phase shift between subbands, a maximum correlation-based approach is introduced. Because the proposed subband WaveNet vocoder introduces an overlapped SSB filterbank, adjacent subbands include a common frequency component. By using the common frequency component, a one-half overlap-and-add frame shift-based linear phase compensation between adjacent subband waveforms is performed sequentially from the low subbands. The analysis window is

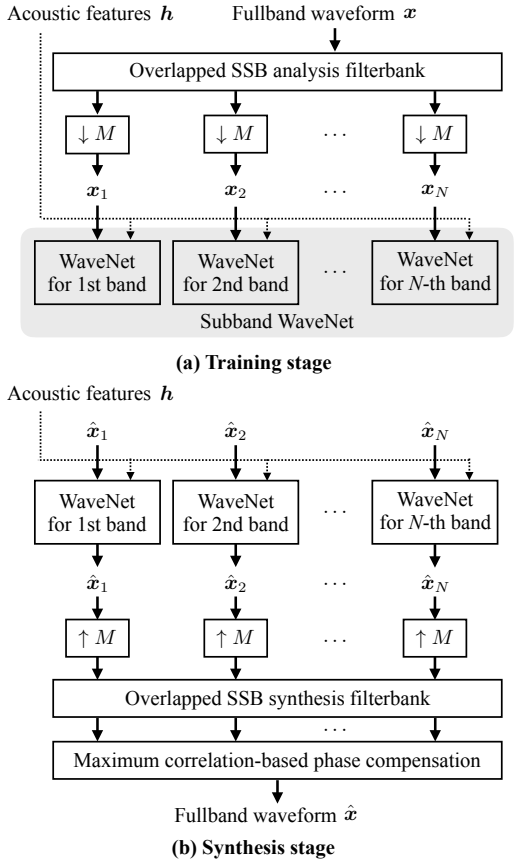


Fig. 1. Block diagram of proposed subband WaveNet vocoder.

a Hann window with length-s samples, and the frame shift length is $s/2$ samples. In the i-th frame, the lower subband waveform is windowed as $\boldsymbol{x}_{\mathrm{low},i}$, and the higher subband waveform is windowed as $\boldsymbol{x}_{\mathrm{high},i}$ with time-shifted $\pm q$ samples that maximize the correlation between $\boldsymbol{x}_{\mathrm{high},i}$ and $\boldsymbol{x}_i$ for considering both the adjacent subband and previous waveforms, where

$$\boldsymbol{x}_i = \boldsymbol{x}_{\mathrm{low},i} + [x(s/2+1)_{\mathrm{high},i-1}, \cdots x(s)_{\mathrm{high},i-1}, \underbrace{0, \cdots 0}_{s/2}]. \tag{2}$$

Higher subband waveform $\boldsymbol{x}_{\mathrm{high},i}$ is then overlap-and-added. All phase-compensated subband waveforms are finally integrated into fullband waveform $\hat{\boldsymbol{x}} = [\hat{x}(1), \cdots, \hat{x}(T)]$.

3. EXPERIMENTS

3.1. Experimental conditions

To evaluate the effectiveness of the proposed speaker-dependent subband WaveNet vocoder with a sampling frequency of 48 kHz, we conducted objective and subjective experiments using a Japanese male speech corpus recorded with a sampling frequency of 48 kHz. In the experiments, 5697 (about 3.7 hours) and 100 utterances were respectively used as the training and test sets [26].

The proposed subband WaveNet vocoder was compared with a fullband WaveNet vocoder that can be calculated with limited model parameters using a consumer GPU as well as conventional

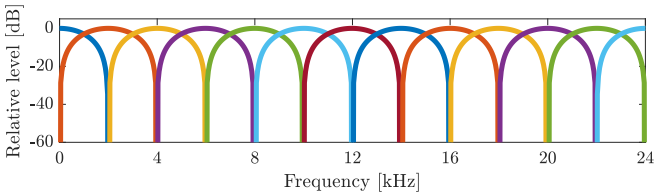


Fig. 2. Frequency response of a square-root Hann window-based overlapped SSB filterbank with decimation factor $M = 6$ and division number $N = 13$ for proposed subband WaveNet vocoder.

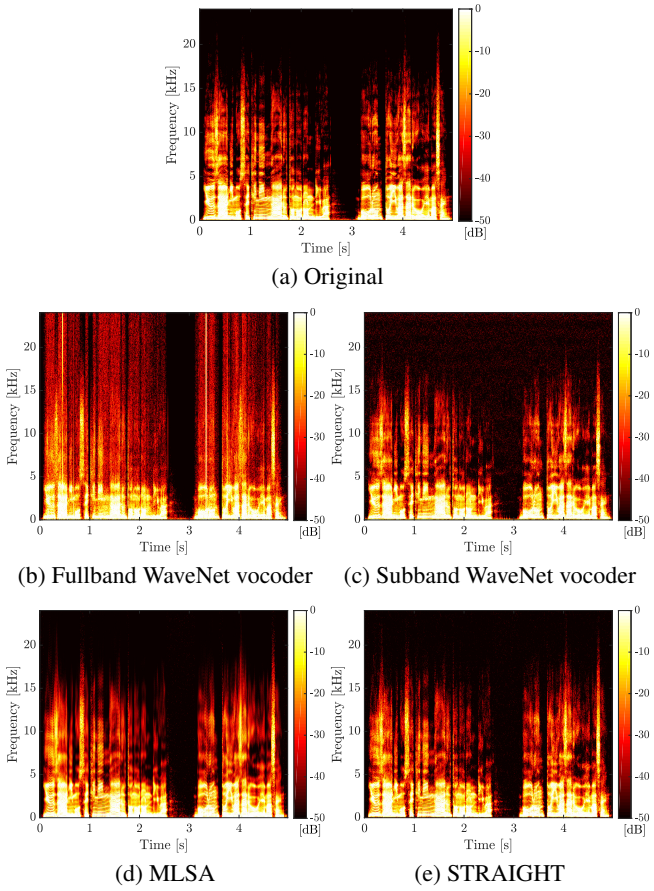
MLSA [5] and STRAIGHT [3] vocoders with a sampling frequency of 48 kHz. Fullband and subband WaveNet vocoders were trained using an Intel Xeon(R) CPU E5-2670 and a single GPU of NVIDIA GeForce GTX 1080. In the experiments, acoustic features were analyzed every 5 ms and the length of an analysis Hann window was 25 ms. Fundamental frequency f_0, analyzed by an NDF algorithm implemented in STRAIGHT [28], was used in all the vocoders. For fullband and subband WaveNet vocoders, the 0-th to 34-th mel-cepstral coefficients (35-dimensions) were analyzed from a simple short-time Fourier transform of windowed speech waveforms with a sampling frequency of 48 kHz and warping coefficient $\alpha = 0.55$ as a default setting in HTS[1]. In addition, the bandlimited acoustic features were analyzed from downsampled waveforms with sampling frequencies of 16 and 8 kHz to explore the bandwidth extension of the proposed subband WaveNet vocoder. The 0-th to 24-th mel-cepstral coefficients (25-dimensions) and the 0-th to 16-th mel-cepstral coefficients (17-dimensions) were respectively analyzed for sampling frequencies of 16 and 8 kHz with $\alpha = 0.42$ and 0.31. In the MLSA vocoder, the 0-th to 49-th mel-cepstral coefficients (50-dimensions) were obtained from the smooth vocal tract spectrum analyzed by STRAIGHT in a previous work [22]. In the STRAIGHT vocoder, the 0-th to 59-th mel-cepstral coefficients (60-dimensions) for the smooth vocal tract spectrum as well as the 0-th to 24-th mel-cepstral coefficients (25-dimensions) for the aperiodicity component were analyzed by STRAIGHT [23, 24].

According to the experimental results of unconditional subband WaveNet [26], a square-root Hann window-based overlapped SSB filterbank was also introduced for the proposed subband WaveNet vocoder. In the experiments, decimation factor M was set to 6 and division number $N = 2M + 1 = 13$. The length of the analysis and synthesis prototype FIR filters was 1536 samples. The sampling frequency of each subband waveform was $(48/6 =)$ 8 kHz, and each subband WaveNet was easily trained by a consumer GPU. The frequency response of the filterbank is plotted in Fig. 2. For phase compensation between subbands, we employed $s = 960$ and $q = 240$ in Eq. (2).

In a fullband WaveNet vocoder with a sampling frequency of 48 kHz, 24 dilated causal convolution layers were introduced as $\{1, 2, 4, \cdots, 2048\} \times 2$, whose receptive field length was $4096 \times 2/48000 = 0.170$ s. The mini-batch size was 20 k samples $(= 0.42$ s). All the dilation and residual channels as well as the number of skip connections were set to 256. These parameters were one of the limit settings using a GPU of NVIDIA GeForce GTX 1080.

In the proposed subband Wavenet vocoder, we employed 27 dilated causal convolution layers as $\{1, 2, 4, \cdots, 256\} \times 3$, whose receptive field length was 0.192 s, for each subband WaveNet. The mini-batch size was 20 k samples $(= 2.5$ s). Both the dilation and residual channels were set to 64. The numbers of skip connections were set to 64 for the 1st and 2nd bands, 16 for the 3rd to 6th bands, and 8 for the 7th to 13th bands.

[1]http://hts.sp.nitech.ac.jp



(a) Original

(b) Fullband WaveNet vocoder (c) Subband WaveNet vocoder

(d) MLSA (e) STRAIGHT

Fig. 3. Spectrograms: (a) test set original speech waveform with a sampling frequency of 48 kHz, (b) synthesized by fullband WaveNet vocoder, (c) synthesized by subband WaveNet vocoder, (d) MLSA vocoder, (e) STRAIGHT vocoder with fullband acoustic features.

In both the fullband and subband WaveNet vocoders, the number of parameter updates was 200 k, and an Adam optimization algorithm updated the neural network parameters with a learning rate of 0.001 as an initial value that was multiplied by 0.5 at all the K parameter updates. In the fullband and subband WaveNet vocoder for the 1st to the 4th bands, $K = 50$ k and for the 5th to 13th bands, $K = 10$ k. In the synthesis, a fast generation algorithm was introduced [29].

3.2. Objective evaluations

Figure 3 shows the spectrograms of a test set original speech waveform and those generated by fullband and subband WaveNet and MLSA and STRAIGHT vocoders with fullband acoustic features. Obviously, the fullband WaveNet cannot correctly generate speech waveforms especially over 5 kHz, since the number of model parameters was insufficient to train WaveNet with a sampling frequency of 48 kHz. Since the speech quality synthesized by the fullband WaveNet vocoder was distinctly lower than the other vocoders, we removed it after the objective and subjective evaluations.

Table 1. Results of objective evaluations of 100 test set utterances.

	SNR [dB]	SD [dB]	MCD [dB]
MLSA	0.70 ± 0.09	$\mathbf{9.70 \pm 0.08}$	$\mathbf{0.51 \pm 0.03}$
STRAIGHT	0.90 ± 0.11	9.85 ± 0.09	0.63 ± 0.03
Subband (A. F. 48 k)	4.60 ± 0.13	11.7 ± 0.11	0.68 ± 0.04
Subband (A. F. 16 k)	$\mathbf{5.50 \pm 0.14}$	12.1 ± 0.10	0.55 ± 0.03
Subband (A. F. 8 k)	5.30 ± 0.13	12.7 ± 0.09	0.61 ± 0.04

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

$$SNR = 10 \log_{10} \left(\frac{\sum_{t=1}^{T} \hat{x}(t)^2}{\sum_{t=1}^{T} (x(t) - \hat{x}(t))^2} \right), \tag{3}$$

$$SD = \frac{1}{A} \sum_{a=1}^{A} \sqrt{\frac{1}{F} \sum_{f=1}^{F} \left(20 \log_{10} \frac{|\hat{X}(f,a)|}{|X(f,a)|} \right)^2}, \tag{4}$$

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 number of frames. Similar to a previous work [18], we introduced a linear phase compensation for each frame to calculate the SNR. As an acoustic feature analysis, the short-time Fourier transform analysis window function was also a Hann window with a frame length of 25 ms and a frameshift of 5 ms. To consider the human auditory perception criterion in the objective evaluation, mel-cepstral distortion (MCD) was also introduced and defined:

$$MCD = \frac{10}{\log 10} \sqrt{2 \sum_{b=1}^{B} (c(b) - \hat{c}(b))^2}, \tag{5}$$

where $c(b)$ and $\hat{c}(b)$ are the b-th mel-cepstral coefficients obtained from $X(f,a)$ and $\hat{X}(f,a)$ with $\alpha = 0.55$ and $B = 34$.

The results of the objective evaluations are shown in Table 1. Just as in previous results [18], subband WaveNet vocoders achieved a higher SNR than MLSA and the STRAIGHT vocoders while the SD and MCD for MLSA were higher than the others since the subband WaveNet vocoders directly generated speech waveforms and their phase components can be reconstructed.

3.3. Subjective evaluations

To subjectively compare the subband Wavenet vocoder with fullband and bandlimited acoustic features with MLSA and STRAIGHT vocoders, mean opinion score (MOS) tests were conducted. In the subjective evaluation, the equivalent Q value was utilized to provide reliability for the evaluation based on P. 830 [30]. Modulated noise reference unit (MNRU) $y(t)$ was first prepared with $Q = 15,\ 20,\ 25,\ 30,\ 35,$ and 40 dB:

$$y(t) = x(t) + 10^{-Q/20} x(t) n(t), \tag{6}$$

where $n(t)$ is Gaussian white noise. Vocoded speech waveforms were then evaluated by MOS tests. Finally, equivalent Q values were obtained based on the above MOS scores of MNRU and vocoded speech. 23 utterances out of the test set were used for each vocoded and MNRU speech as the evaluation set and presented by headphones. As listening subjects, 15 Japanese adult native speakers without hearing loss evaluated $23 \times (5 + 6) = 253$ utterances.

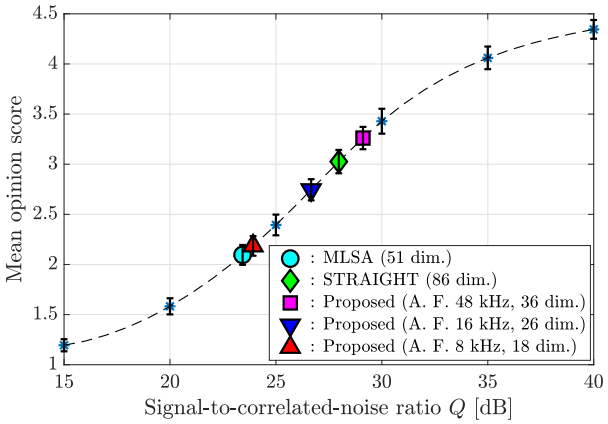


Fig. 4. Results of MOS and equivalent Q value for subjective evaluations with 15 listening subjects.

The MOS results and the equivalent Q values are plotted in Fig. 4. First, the statistical analysis of the result indicates that the proposed subband WaveNet vocoder with 36-dimensional simple fullband acoustic features significantly outperformed other vocoders including STRAIGHT with 86-dimensional features. The t-test result and the equivalent Q value difference between the proposed subband WaveNet vocoder with fullband acoustic features and STRAIGHT were respectively $p = 7.32 \times 10^{-5} \ll 0.05$ and 1.14 dB. Although the subband WaveNet vocoders with bandlimited acoustic features synthesized speech waveforms with a sampling frequency of 48 kHz, their synthesized speech qualities failed to reach those of the subband WaveNet vocoder with fullband features and STRAIGHT. The results indicate that a higher frequency component of acoustic features is required for high-quality synthesis covering the entire human auditory frequency range. Consequently, we validated the availability of the proposed subband architecture in conditional WaveNet training and synthesis.

4. FUTURE WORK

Although there are many common parameters in each subband in the experiments, such parameters as the receptive field length will be set to optimum values in each subband WaveNet for higher-quality synthesis. Experiments using a female speech corpus will also be conducted. Moreover, a method that simultaneously inputs and outputs all or some subband waveforms with a single network will be investigated for phase shift compensation between subbands within a neural network framework instead of the simple maximum correlation-based phase compensation introduced in the experiments.

5. CONCLUSIONS

For a WaveNet vocoder with a sampling frequency of 48 kHz using a consumer GPU, this paper proposed a subband WaveNet vocoder covering the entire human audible frequency range. By introducing multirate signal processing, each subband WaveNet vocoder was successfully trained with a consumer GPU. The results of subjective evaluations using a Japanese male speech corpus showed that the proposed subband WaveNet vocoder with 36-dimensional simple fullband acoustic features significantly outperformed conventional source-filter model-based vocoders including STRAIGHT with 86-dimensional features.

6. REFERENCES

[1] 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.

[2] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, Apr. 2017.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.*, vol. 39, no. 2, pp. 163–166, Mar. 2018.

[9] 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.

[10] 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*, 2016, (unreviewed manuscript).

[11] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017, (unreviewed manuscript).

[12] 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.

[13] 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.

[14] 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.

[15] S. O. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Proc. NIPS*, Dec. 2017, pp. 2966–2974.

[16] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc. ICLR*, Apr. 2018.

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

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

[19] K. Kobayashi, A. Tamamori, T. Hayashi, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," in *Proc. Interspeech*, Aug. 2017, pp. 1138–1142.

[20] 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.

[21] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. ASRU*, Dec. 2017, pp. 712–718.

[22] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Trajectory training considering global variance for speech synthesis based on neural networks," in *Proc. ICASSP*, Mar. 2016, pp. 5600–5604.

[23] 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.

[24] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. SSW 9*, Sept. 2016, pp. 218–223.

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

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

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

[28] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Proc. Interspeech*, Sept. 2005, pp. 537–540.

[29] 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.

[30] ITU-T Recommendation P. 830, *Subjective performance assessment of telephone-band and wideband digital codecs*, 1990.