INVESTIGATIONS OF REAL-TIME GAUSSIAN FFTNET AND PARALLEL WAVENET NEURAL VOCODERS WITH SIMPLE ACOUSTIC FEATURES

Takuma Okamoto[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

This paper examines four approaches to improving real-time neural vocoders with simple acoustic features (SAF) constructed from fundamental frequency and mel-cepstra rather than mel-spectrograms. The investigations are as follows: 1) the effectiveness of single Gaussian (SG) autoregressive (AR) WaveNet and FFTNet vocoders with SAF, 2) the possibility of SG parallel WaveNet vocoder training and synthesis with SAF, 3) the impact of noise shaping on SG AR neural vocoders, and 4) the efficacy of bandwidth extension to synthesize speech waveforms at a sampling frequency of 24 kHz by SG AR neural vocoders from SAF for that of 16 kHz. The results of experiments indicate that SG AR WaveNet and real-time SG AR FFTNet vocoders with noise shaping using SAF can realize sufficient synthesis quality with bandwidth extension effect. Moreover, a real-time SG parallel WaveNet vocoder can also be trained using SAF.

Index Terms— Speech synthesis, neural vocoder, FFTNet, parallel WaveNet, Gaussian inverse autoregressive flow, noise shaping

1. INTRODUCTION

Neural network-based raw audio autoregressive (AR) generative models such as WaveNet [1], SampleRNN [2], FFTNet [3], and WaveRNN [4] have been recently investigated and outperform the conventional source-filter vocoders typically employed in conventional statistical parametric speech synthesis (SPSS) [5] and voice conversion (VC) [6]. Such raw audio generative models can realize end-to-end text to speech (TTS), converting text to raw speech waveforms. The speech quality of English synthesized by Tacotron 2 can match that of natural speech at a sampling frequency fs of 24 kHz [7].

In contrast to TTS, a WaveNet vocoder that directly synthesizes raw speech waveforms from acoustic features [8] has been proposed to drive conventional source-filter vocoders within a neural network-based raw audio generative model framework. Neural vocoders based on WaveNet [8] and SampleRNN [9] have been applied to SPSS [10, 11] and VC [12–14], and have been shown to outperform conventional source-filter vocoders.

Compared with conventional source-filter vocoders, however, the synthesis speed of AR neural vocoders remains problematic, because the sequential synthesis of each sample requires a huge amount of calculation time [1,2]. To overcome this problem, FFTNet [3] has a simpler structure based on a 1×1 convolutional network and rectified liner unit layers. As a result, FFTNet can synthesize raw audio in real time. In addition, noise shaping [15, 16] and subband approaches [17, 18] have been applied to 256-way categorical FFTNet to improve the synthesized speech quality [19]. However, its performance cannot reach that of 256-way categorical WaveNet with noise shaping [19].

Parallel WaveNet [20] introduces teacher–student knowledge distillation based on the inverse autoregressive flow (IAF) [21] and WaveRNN [4] introduces a single-layer recurrent neural network with sparse and subscale modifications. These methods enable real-time synthesis with 16-bit raw audio prediction. However, the training in parallel WaveNet is unstable because of the intractable Kullback–Leibler (KL) divergence between the student logistic and teacher mixture of logistics (MoL) distributions [22].

To solve this problem, single Gaussian (SG) parallel WaveNet was proposed. This method is based on the Gaussian IAF [21]; the non-AR student WaveNet is trained using a SG AR teacher WaveNet. An SG-based AR WaveNet outperforms a MoL-based AR WaveNet and the student network is better trained than in a MoL-based parallel WaveNet. The entire end-to-end TTS based on SG parallel WaveNet is called ClariNet [22].

The SG modeling is only applied to AR and parallel WaveNet vocoders with 80-band mel-spectrograms for end-to-end TTS [22]. In conventional SPSS and VC, however, source-filter vocoders are still widely employed to synthesize speech waveforms from estimated and converted simple acoustic features (SAF) mainly constructed from the fundamental frequency and mel-cepstra [23, 24] rather than mel-spectrograms [7, 14, 22]. Therefore, it is important to investigate the effectiveness of SG modeling in AR WaveNet and FFTNet with SAF, as employed in previous neural vocoders [8–10, 12, 13, 15, 16, 18, 19]. In addition, it is also important to determine how parallel WaveNet can be successfully trained from SAF because this remains uninvestigated. Furthermore, the impact of noise shaping for these SG-based neural vocoders should also be investigated at $fs = 24$ kHz because all previous investigations used 256-way categorical WaveNet and FFTNet neural vocoders at $fs = 16$ kHz [15, 16, 19].

In Tacotron 2 [7], high quality speech waveforms at $fs = 24$ kHz are synthesized from 80-band mel-spectrograms with a frequency band of 125–7,600 Hz. The result indicates that SAF for lower sampling frequencies might also be able to synthesize high quality speech waveforms. Therefore, it is also important to investigate whether speech waveforms at $fs = 24$ kHz can be synthesized by neural vocoders from SAF for $fs = 16$ kHz to explore the possibility of bandwidth extension. If this is possible, existing SPSS and VC systems operating at $fs = 16$ kHz could be able to synthesize speech waveforms at $fs = 24$ kHz.

Consequently, to improve real-time neural vocoders with SAF, this paper investigates the following four topics: 1) the effectiveness of the SG AR WaveNet and FFTNet vocoders with SAF, 2) the possibility of SG parallel WaveNet training and synthesis with SAF, 3) the impact of noise shaping on SG AR neural vocoders, and 4) the efficacy of bandwidth extension to synthesize speech waveforms at $fs = 24$ kHz by SG AR neural vocoders from SAF at $fs = 16$ kHz.

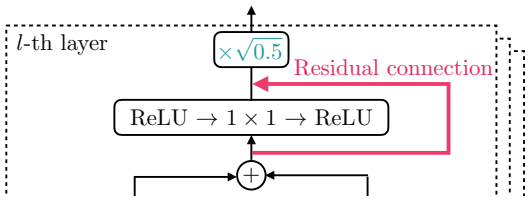

Fig. 1. Proposed FFTNet with residual connections and additional multiplications.

2. INVESTIGATIONS OF SINGLE GAUSSIAN NEURAL VOCODERS WITH SIMPLE ACOUSTIC FEATURES

2.1. SG AR WaveNet and FFTNet neural vocoders with SAF

Given acoustic features $\boldsymbol{h}$, AR WaveNet [8] and FFTNet [3] neural vocoders model conditional probability distribution $p(\boldsymbol{x}|\boldsymbol{h})$ of raw audio waveform $\boldsymbol{x} = [x_1, \cdots, x_T]$ as

$$p(\boldsymbol{x}|\boldsymbol{h}) = \prod_{t=1}^{T} p(x_t|x_{<t}, \boldsymbol{\theta}; \boldsymbol{h}), \tag{1}$$

where $\boldsymbol{\theta}$ are the parameters of the model. In AR WaveNet, Eq. (1) is modeled by a stack of dilated causal convolution layers, allowing the efficient input of very long audio samples with relatively few layers. However, the network model size of WaveNet vocoders is still too large to synthesize speech waveforms in real time. In contrast, to significantly reduce the network model size, FFTNet uses simple 1×1 convolution layers instead of the dilated causal convolution layers, and can therefore synthesize speech waveforms in real time with a fast generation algorithm [25].

Although the vanilla WaveNet and FFTNet models introduce a categorical distribution of the next sample x_t based on μ-law companding algorithm defined in G.711 [26], the MoL distribution has been employed in parallel WaveNet [20] and several WaveNet vocoders [11, 13] including Tacotron 2 [7] for predicting 16 bit law audio to improve synthesized speech quality. In these neural vocoders, 10-component MoL is used and the output consists of 30 channels. In parallel WaveNet, the MoL distribution is also used for teacher–student knowledge distillation-based training using the IAF [21]. However, a parallel WaveNet employs a Monte Carlo method to approximate the intractable KL divergence between the student logistic and teacher MoL distributions. Therefore, a double-loop sampling is required for the student input and estimation of the intractable KL divergence. [22].

To solve this problem in MoL-based parallel WaveNet, an SG output distribution is provided. The conditional distribution of $p(x_t|x_{<t}, \boldsymbol{\theta}; \boldsymbol{h})$ in Eq. (1) is defined as:

$$p(x_t|x_{<t}, \boldsymbol{\theta}; \boldsymbol{h}) = \mathcal{N}(\mu(x_{<t}; \boldsymbol{\theta}), \sigma(x_{<t}; \boldsymbol{\theta})), \tag{2}$$

where $\mu(x_{<t}; \boldsymbol{\theta})$ and $\sigma(x_{<t}; \boldsymbol{\theta})$ are the mean and standard deviation predicted by the AR WaveNet, respectively. Network parameters $\boldsymbol{\theta}$ are trained using maximum likelihood estimation [22]. Compared with MoL distribution, the SG AR WaveNet can synthesize speech waveforms with higher quality when using mel-spectrogram input although only two output channels are used. In addition, the SG modeling can efficiently train parallel WaveNet because the KL divergence between the two Gaussian distributions of the teacher and student models can be analytically calculated without sampling [22].

The SG modeling has only been applied to AR and parallel WaveNet vocoders with mel-spectrogram input for end-to-end TTS [22]. Therefore, this paper investigates the effectiveness of the SG modeling for these neural vocoders with SAF constructed from the fundamental frequency and mel-cepstra for directly applying the conventional SPSS and VC frameworks.

SG modeling can also be directly applied to an FFTNet vocoder with SAF. Compared with the categorical FFTNet, which has 256 input and output channels, there are one input and two output channels in the SG WaveNet. Real-time synthesis is also realized by SG FFTNet vocoder. In a previous work, the residual connections in all layers were found to improve the FFTNet model while keeping the network model size [19]. In this paper, additional multiplications with $\sqrt{0.5}$ are introduced after the residual connections in all layers to improve the stability of the network, as shown in Fig. 1 and the effectiveness of MoL and SG FFTNet vocoders with SAF are also investigated.

2.2. SG parallel WaveNet vocoder with SAF

The SG parallel WaveNet is based on the Gaussian IAF [21]. In SG parallel WaveNet, white noise $\boldsymbol{z}^{(0)}$ is first generated from the isotropic Gaussian distribution $\mathcal{N}(0, \boldsymbol{I})$ as the input of the non-AR student WaveNet where $\boldsymbol{I}$ is a unit matrix. Using n stacked Gaussian IAFs, where each flow is parameterized by a WaveNet, student input $\boldsymbol{z}^{(0)}$ is repeatedly transformed as $\boldsymbol{z}^{(0)} \rightarrow \boldsymbol{z}^{(1)} \rightarrow \cdots \rightarrow \boldsymbol{z}^{(n)}$. For acoustic feature input, the upsampling layer trained in the teacher WaveNet is commonly used. Then, the student WaveNet outputs a synthesized waveform $\boldsymbol{x}_q$ and student Gaussian distribution $q(x_t|x_{<t}, \boldsymbol{\theta}; \boldsymbol{h})$ with mean μ_q and standard deviation σ_q where $\boldsymbol{x}_q = \boldsymbol{z}^{(n)} = \boldsymbol{z}^{(0)} \odot \boldsymbol{\sigma}_q + \boldsymbol{\mu}_q$. To match student Gaussian distribution $q(x_t|x_{<t}, \boldsymbol{\theta}; \boldsymbol{h})$ to teacher Gaussian distribution $p(x_t|x_{<t}, \boldsymbol{\theta}; \boldsymbol{h})$ provided by the SG AR teacher WaveNet, the KL divergence between the two distributions is minimized. In contrast to the intractable KL divergence in MoL-based parallel WaveNet [20], the KL divergence between two Gaussian distributions can be analytically calculated. In SG parallel WaveNet, the regularized KL divergence loss is introduced as

$$\begin{aligned}\mathrm{KL}^{\mathrm{reg}}(q||p) =& \lambda|\log\sigma_p - \log\sigma_q|^2 \\ &+ \log\frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2 - \sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_p^2},\end{aligned} \tag{3}$$

where $\lambda = 4$ is used in [22].

However, only minimizing the regularized KL divergence loss leads to whisper voices [20, 22]. To avoid the problem, the spectrogram frame loss between synthesized and ground truth waveforms $\boldsymbol{x}_q$ and $\boldsymbol{x}$ is additionally considered and total loss $\mathcal{L}$ is calculated as

$$\mathcal{L} = \mathrm{KL}^{\mathrm{reg}}(q||p) + \frac{1}{B}\left\| |\mathrm{STFT}(\boldsymbol{x}_q)| - |\mathrm{STFT}(\boldsymbol{x})| \right\|_2^2 \tag{4}$$

in SG parallel WaveNet, where $|\mathrm{STFT}(\boldsymbol{x})|$ is the magnitude of the short-term Fourier transform (STFT), $B = 1025$ is the number of frequency bins, and the STFT size is 2048 with a 12.5 ms frame-shift 50 ms Hann window [22]. Using SG parallel WaveNet instead of SG AR WaveNet enables speech waveforms to be directly generated by the student SG WaveNet with white noise $\boldsymbol{z}^{(0)}$ and acoustic features $\boldsymbol{h}$ in real time.

As SG AR WaveNet vocoder, SG parallel WaveNet vocoder has also only been investigated with mel-spectrogram input [22]. Therefore, the possibility of SG parallel WaveNet vocoder training and synthesis with SAF instead of mel-spectrograms is also investigated in this paper.

2.3. Impact of noise shaping on SG neural vocoders with SAF

The speech signals synthesized by neural vocoders often suffer from noise caused by prediction errors, and these noise signals tend to cause large spectral distortions in high-frequency bands. Thus, the noise signals degrade the synthesized speech quality [16]. To reduce the adverse effects of the noise signals generated by neural vocoders, predictive pulse code modulation (PPCM) [27]-based time-invariant noise shaping, which is a perceptual weighting technique, has been applied to categorical WaveNet and FFTNet vocoders [15, 16, 19]. This noise shaping should improve the synthesized speech quality of SG AR WaveNet and FFTNet as well as SG parallel WaveNet vocoders. Therefore, the impact of noise shaping on SG neural vocoders with SAF is also investigated in this paper.

2.4. Efficacy of bandwith extention on SG neural vocoders with SAF

As described in Sec. 1, high quality speech waveforms at $fs = 24$ kHz are synthesized by a MoL-based WaveNet vocoder using 80-band mel-spectrograms with a frequency band of 125–7600 Hz in Tacotron 2 [7]. Therefore, SG WaveNet and FFTNet neural vocoders are trained to synthesize speech waveforms at $fs = 24$ kHz by using SAF for $fs = 16$ kHz to investigate the bandwidth extension effect.

3. EXPERIMENTS

3.1. Experimental conditions

To evaluate the impact of the SG modeling and noise shaping methods for AR WaveNet and FFTNet as well as parallel WaveNet neural vocoders with SAF at $fs = 24$ and 16 kHz, a series of objective and subjective experiments were conducted using a Japanese male speech corpus recorded at $fs = 48$ kHz and downsampled to 24 kHz, as used in [17–19]. In the experiments, 5,697 utterances (about 3.7 h) were used as the training set and 20 utterances were used as the test set.

In the experiments, acoustic features $\boldsymbol{h}$ were analyzed every 5 ms over a Hann window of length 25 ms. Fundamental frequency f_{o}, analyzed by an NDF algorithm [28], was used in all the vocoders with SAF, as in [18, 19]. For the neural vocoders with SAF, the 0–th to 34–th mel-cepstral coefficients (35 dimensions) were analyzed from a simple STFT of windowed speech waveforms at $fs = 24$ kHz with warping coefficient $\alpha = 0.46$. In addition, the 0–th to 24–th mel-cepstral coefficients (25 dimensions) were also analyzed from speech waveforms at $fs = 16$ kHz with warping coefficient $\alpha = 0.42$ to investigate the possibility of bandwidth extension. In the neural vocoders with SAF for $fs = 24$ kHz, $(1 + 1 + 35 =)$ 37-dimensional vectors constructed from continuous logarithmic f_{o}, voice/unvoice one-hot vector, and mel-cepstrum coefficients (normalized to have a zero-mean and unit-variance) were used as acoustic features $\boldsymbol{h}$. For the acoustic features for $fs = 16$ kHz, $(1 + 1 + 25 =)$ 27-dimensional vectors were employed.

The SG WaveNet with mel-spectrogram input is used as a reference. In this method, 80-dimensional log-mel-spectrograms were analyzed every 12.5 ms over a Hann window of length 85.3 ms with a frequency band 125 to 7,600 Hz and normalized to the range of [0, 1], as in [7]. Similar in [19], transposed convolution [1] was applied for upsampling the acoustic features in all the neural vocoders and the upsampling layer was also trained with neural vocoder models.

To directly compare the results with those of the original SG WaveNet vocoders [22], the same network parameters as those of the original SG WaveNet vocoders were employed. Both the residual and skip channels of AR and parallel WaveNet vocoders were set to 128. Twenty layers (10 dilations $\times$ 2 cycles) with a kernel size of two were used for the dilated causal convolution layers, giving a receptive field of 2,047 samples for the AR teacher WaveNet vocoder. In the parallel WaveNet vocoder, 60 layers, where each Gaussian IAF is parameterized by a 10-layer WaveNet with a kernel size of three, were used, as in [22].

In the FFTNet vocoders, $L = 11$ layers were introduced and the receptive field was $2^{11} = 2,048$ samples, as used in the original FFTNet [3]. The channel number of each FFTNet layer was 256 [3]. In FFTNet, 10-component MoL models were also investigated.

The AR WaveNet, AR FFTNet, and parallel WaveNet vocoders required 3,000,000, 3,000,000, and 1,500,000 parameter updates, respectively. In addition, an Adam optimization algorithm [29] updated the neural network parameters with learning rates of 0.0002, 0.001, and 0.0002, respectively. As in MoL-based parallel WaveNet and Tacotron 2, exponential moving averaging [30] with a decay rate of 0.999 was used for the parameters. The minibatch sizes of all neural vocoders were 2 $\times$ 5,000 samples. They were trained using a single GPU of an NVIDIA Tesla P100.

Similar to previous approaches [15, 16, 19], a mel-generalized cepstrum [23]-based noise shaping filtering implemented by the mel-log spectrum approximation (MLSA) filter [24] was introduced. A parameter to control noise energy in the formant regions was set to 0.5 for noise shaping according to the results of the WaveNet and FFTNet vocoder investigations [15, 16, 19]. Noise shaping was not applied to the SG AR WaveNet with mel-spectrograms. The results of preliminary experiments showed that the speech qualities synthesized by MoL and SG FFTNet without residual connections and noise shaping were obviously lower than those of the others. Hence, they were omitted after the objective and subjective evaluations. As a result, nine models were evaluated, as described in Table. 1 and Fig. 2. Although SG parallel WaveNet with noise shaping was also investigated, it was difficult to train to the same quality as obtained without noise shaping because the loss score cannot be reduced. Therefore, determining a method to effectively train SG parallel WaveNet with noise shaping is a future task.

3.2. Objective evaluations

To objectively evaluate the synthesized test set speech waveforms, the signal-to-noise ratio (SNR) and spectral distortion (SD) between original waveform $\boldsymbol{x}$ and synthesized $\hat{\boldsymbol{x}}$ were computed. As in previous studies [8, 16, 18], a linear phase compensation for each frame was introduced to calculate the SNR. For acoustic feature analysis, the STFT analysis window function was also a Hann window with a frame length of 25 ms and frameshift of 5 ms. To consider the human auditory perception criterion in the objective evaluation, the mel-cepstral distortion (MCD) was also computed with weighting factor $\alpha = 0.46$. The results of the objective evaluations are presented in Table 1.

3.3. Subjective evaluations

To subjectively evaluate the speech waveforms synthesized by nine models, mean opinion score (MOS) tests [31] were conducted. All 20 utterances of the test set were used as the evaluation set. These were presented through headphones to 13 Japanese adult native speakers without hearing loss (20 utterances $\times$ 9 conditions including the original test set waveforms = 180 utterances). The MOS results are plotted in Fig. 2.

Table 1. Results of objective evaluations of 20 test set utterances. "NS" and "Parallel WN" denote noise shaping and parallel WaveNet, respectively.

Method	Network	Type	Input features	Real-time	NS	SNR [dB]	SD [dB]	MCD [dB]
(a):WN-SG-MSPC	WaveNet	SG	Mel-spectrogram			3.30 ± 0.39	9.34 ± 0.20	3.71 ± 0.12
(b):WN-SG-SAF	WaveNet	SG	SAF 24 kHz			$\mathbf{5.40 \pm 0.44}$	8.02 ± 0.08	2.55 ± 0.07
(c):WN-SG-SAF-NS	WaveNet	SG	SAF 24 kHz		✓	3.90 ± 0.73	$\mathbf{7.57 \pm 0.08}$	$\mathbf{2.20 \pm 0.04}$
(d):WN-SG-SAF16k-NS	WaveNet	SG	SAF 16 kHz		✓	3.70 ± 0.69	8.26 ± 0.07	2.89 ± 0.05
(e):PWN-SG-SAF	Parallel WN	SG	SAF 24 kHz	✓		5.20 ± 0.41	8.09 ± 0.06	2.73 ± 0.07
(f):FN-MoL-SAF-NS	FFTNet	MoL	SAF 24 kHz	✓	✓	3.20 ± 0.66	7.96 ± 0.08	2.69 ± 0.05
(g):FN-SG-SAF-NS	FFTNet	SG	SAF 24 kHz	✓	✓	2.90 ± 0.66	8.01 ± 0.08	2.80 ± 0.05
(h):FN-SG-SAF16k-NS	FFTNet	SG	SAF 16 kHz	✓	✓	3.10 ± 0.66	8.53 ± 0.07	3.36 ± 0.05

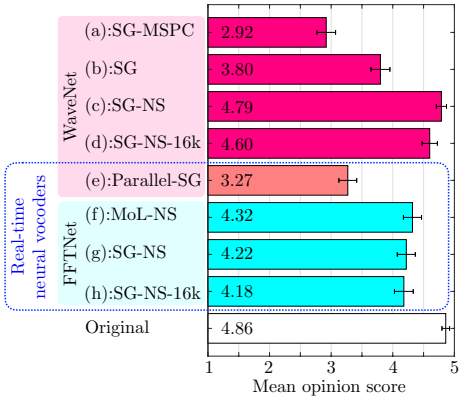

Fig. 2. Results of MOS test with 13 listening subjects. "WN", "FN" and "SAF" are omitted, as defined in Table. 1.

4. DISCUSSIONS

4.1. Effectiveness of SG AR WaveNet vocoder with SAF

From the results of (a) and (b), the synthesis quality of the SG AR WaveNet vocoder with mel-spectrograms was insufficient and lower than that with SAF. This might be due to the lack of training data (only 3.7 hours) since original SG AR WaveNet vocoders with mel-spectrograms have been trained using about 20 hours of training data [22]. The other reason might be due to the lack of network parameters since Tacotron 2 [7] and a WaveNet-based VC with mel-spectrograms [14] have used larger numbers of network parameters than those of these investigations and original SG AR WaveNet [22]. These results indicate that SG AR WaveNet with SAF might be easily trained compared with that with mel-spectrograms when using not so large amount of training data since SAF can might more simply represent speech information than mel-spectrograms although the fundamental frequency estimation is required. However, the synthesis quality of the SG AR WaveNet vocoder without noise shaping using SAF was not so high although highest SNR can be achieved. To confirm this hypothesis and to improve the synthesis quality of the SG AR WaveNet vocoder with SAF, further investigations with a larger amount of training data are required as future work.

4.2. Possibility of SG parallel WaveNet vocoder with SAF

From the results of (e), SG parallel WaveNet can be successfully trained from SAF with reasonable SNR although synthesized waveforms still include spectral distortions and the synthesis quality was not so high. The reason might also be due to the lack of training data since at least 9 hours of training data has been used in the original parallel WaveNet [20]. To improve the synthesis quality of the SG parallel WaveNet vocoder with SAF, a huge amount of training data (more than 10 hours) might also be required.

4.3. Impact of noise shaping on SG AR neural vocoders

From the results of (b), (c), (f) and (g), noise shaping can significantly improve the SG WaveNet, MoL and SG FFTNet neural vocoders to sufficient synthesis quality over MOS values of 4.0 compared with the SG WaveNet without noise shaping when using not so large amount of training data. The results especially for FFTNet, are important since FFTNet can synthesize speech waveforms in real-time and categorical FFTNet vocoders with noise shaping cannot reach sufficient synthesis quality in a previous investigation [19].

4.4. Efficacy of bandwidth extension on SG AR WaveNet and FFTNet vocoders with SAF

From the results of (d) and (h), the SG WaveNet and FFTNet vocoders with SAF for $fs = 16$ kHz can still realize sufficient synthesis quality over MOS values of 4.0 and the bandwidth extension effect was successfully validated in SG neural vocoders although this effect was not confirmed in a previous subband approach [18].

5. FUTURE WORK

To improve the synthesis quality of SG AR FFTNet and SG parallel WaveNet neural vocoders with SAF, further investigations with a larger amount of training data should be required and they should be compared with other real-time neural vocoders [4, 32, 33] as future work. Furthermore, an efficient learning method for training SG parallel WaveNet with noise shaping should also be investigated.

6. CONCLUSIONS

This paper investigated methods for improving real-time neural vocoders with SAF. The results of these investigations indicated that real-time MoL and SG AR FFTNet vocoders with noise shaping using SAF can realize sufficient synthesis quality with bandwidth extension effect. Moreover, a real-time SG parallel WaveNet vocoder can also be trained from SAF. These investigations are important for improving the existing SPSS and VC systems based on source-filter vocoders with SAF.

7. 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," *arXiv:1609.03499*, Sept. 2016, (unreviewed manuscript).

[2] 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.

[3] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "FFTNet: A real-time speaker-dependent neural vocoder," in *Proc. ICASSP*, Apr. 2018, pp. 2251–2255.

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

[5] 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.

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

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

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

[9] Y. Ai, H.-C. Wu, and Z.-H. Ling, "SampleRNN-based neural vocoder for statistical parametric speech synthesis," in *Proc. ICASSP*, Apr. 2018, pp. 5659–5663.

[10] 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.

[11] L. Juvela, V. Tsiaras, B. Bollepalli, M. Airaksinen, J. Yamagishi, and P. Alku, "Speaker-independent raw waveform model for glottal excitation," in *Proc. Interspeech*, Sept. 2018, pp. 2012–2016.

[12] 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.

[13] J. Niwa, T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Statistical voice conversion based on WaveNet," in *Proc. ICASSP*, Apr. 2018, pp. 5289–5293.

[14] K. Chen, B. Chen, J. Lai, and K. Yu, "High-quality voice conversion using spectrogram-based WaveNet vocoder," in *Proc. Interspeech*, Sept. 2018, pp. 1993–1997.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

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

[21] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *Proc. NIPS*, Dec. 2016, pp. 4743–4751.

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

[23] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis — A unified approach to speech spectral estimation," in *Proc. ICSLP*, Sept. 1994, pp. 1043–1046.

[24] 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.

[25] 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.

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

[27] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 27, no. 3, pp. 247–254, June 1979.

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

[30] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Control Optim.*, vol. 30, no. 4, pp. 838–855, July 1992.

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

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

[33] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. ICASSP*, May 2019.