10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Deep neural network based real-time speech vocoder with periodic and aperiodic inputs

Keiichiro Oura[1,2], *Kazuhiro Nakamura*[2], *Kei Hashimoto*[1,2], *Yoshihiko Nankaku*[1], *and Keiichi Tokuda*[1,2]

[1]Department of Computer Science, Nagoya Institute of Technology, Japan
[2]Department of Research and Development, Techno-Speech, Inc., Japan

uratec@nitech.ac.jp

Abstract

In this paper, we propose a framework for speech synthesis taking both periodic and aperiodic inputs. Recently, a method of modeling speech waveforms directly, called WaveNet [1], was proposed. WaveNet is able to model speech waveforms accurately and is able to generate natural speech directly, so it is being used, particularly as a speech vocoder [2], in various research [3, 4, 5]. However, it has an autoregressive structure that generates speech sample from the sequence of past speech samples, so parallel computation cannot be used for synthesis, and consequently real-time synthesis is not possible. It also uses pitch information as an auxiliary feature, so it is unable to generate waveforms with a pitch outside of the range in the training data [6], and even if a pitch within the range of the training data is specified, a waveform with a different pitch could be generated. To address these issues, we propose a method that uses periodic and aperiodic input signals to generate the speech sample sequence at once. With the proposed method, speech can be generated faster than real-time, and speech waveforms with pitch outside the range of the training data can be generated. We also conducted a subjective evaluation of the naturalness of the speech, which indicated better synthesized speech quality than WaveNet.

Index Terms: DNN, GAN, signal processing, speech synthesis, singing synthesis

1. Introduction

Speech is one of the most familiar communication channels for humans, so it has long been a subject of research. Recently, speech related technologies based on digital signal processing, such as speech coding, speech recognition, and speech synthesis, have been incorporated into smartphones, household electronics and other devices, and become practical in people's daily lives. The digital signal processing we refer to here assumes that discrete-time signals converted from analog speech signals are linear-time invariant systems, and is based on theory such as the Fourier transform and the z-transform. This signal processing is based on a generative model of speech that consist of models for a sound source, the glottis, the vocal tract and emission, and is the most basic approach used in speech-related research. However, past speech related research has been limited by frameworks handling transformations and processing in this way, and excessive constraints on model structure have limited performance.

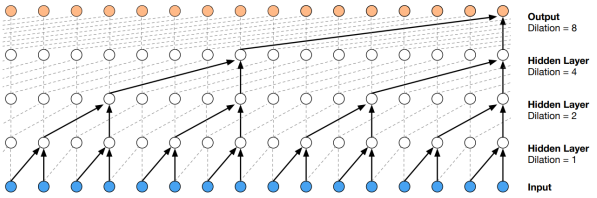

Figure 1: *WaveNet*

Then, in 2016, WaveNet [1] was proposed, using a waveform-generating model with an autoregressive structure (Figure 1). This model used the sequence of several thousand speech samples previously output by the model itself to predict the following speech sample, and was able to generate very high-quality speech. WaveNet can be used as a speech vocoder [2] by using auxiliary features such as mel cepstrum and log F0, and has been used often in recent research on speech synthesis [3, 4, 5]. WaveNet has an autoregressive structure, so parallel processing cannot be used for synthesis, and real-time synthesis is not possible. Therefore, fast generation techniques with autoregressive structure such as WaveRNN [7], LPCNET [8], etc. have been proposed. On the other hand, non autoregressive methods such as Parallel WaveNet [9] and ClariNet [10] have been proposed to resolve this issue. Parallel WaveNet, which uses Wavenet as the teacher model to train a student model that does not have autoregressive structure, is a high-speed generative model that can synthesize speech with the same quality as WaveNet. The student model has the same structure as the teacher model, but when generating speech sample output, rather than instead of using past speech samples generated by the model itself as input as the teacher model does, it uses a noise sequence. Training is done using distillation, in which the output probability distribution of the student model, which uses forward propagation, is trained to approximate the output probability distribution of the teacher model, which has autoregressive structure (Figure 2). Since Parallel WaveNet does not have autoregressive structure, multiple speech samples can be output at once, greatly reducing the time for synthesis relative to WaveNet.

Speech vocoders using autoregressive structure based method [1, 7, 8] and normalizing-flow based method [9, 10, 11]

10.21437/SSW.2019-3

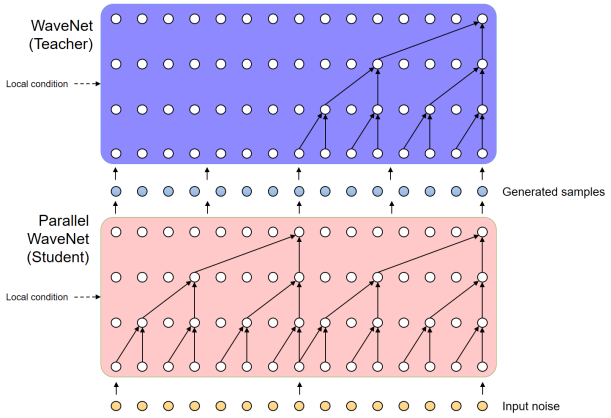

Figure 2: *Probability density distillation of Parallel WaveNet*

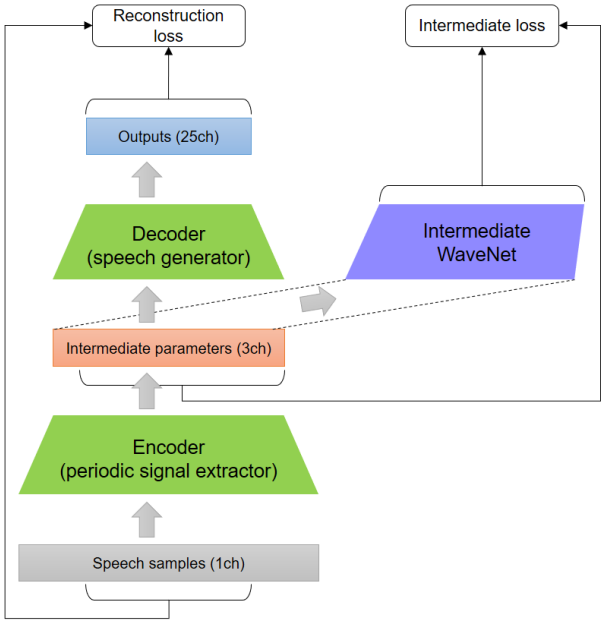

Figure 3: *Overview of the proposed model*

can synthesize speech with very high quality, but they also have several issues. Since these models do not have explicit input of a periodic signal, they cannot generate waveforms with pitch outside of the range in the training data [6], and further, they occasionally generate waveforms of a different pitch, even if a pitch within the range of the training data is specified. The error of intonation and pitch accent can decrease the quality of text-to-speech synthesis. Because accurate reproduction of pitch in singing synthesis has a strong effect on quality, this is an issue that needs to be resolved. Parallel WaveNet can perform real-time speech synthesis, but several types of loss besides probability density distillation loss must be used (power loss, perceptual loss, and contrastive loss), so tuning is difficult during training. As such, in this paper, we propose a framework for speech vocoders based on neural networks generating speech using periodic and aperiodic signals as inputs.

2. Proposed model

2.1. Overview

An overview of the proposed model is shown in Figure 3. Its structure is similar to that of Deep Auto-Encoders (DAE), which have been actively researched recently, with an Encoder (periodic signal extractor), which converts the speech waveform to intermediate parameters, in series with a Decoder (speech generator), which converts the intermediate parameters to speech. Note that in this paper we use three channels sequence, which consists of sample-level sine, cosine and voiced/unvoiced-flag sequences, as the intermediate parameters. "Intermediate WaveNet," which represents the intermediate parameters, is trained beforehand by using the sample-level intermediate parameters generated from the frame-level F0 parameters of the training data. The Encoder/Decoder are trained simultaneously to minimize the sum of Intermediate loss and Reconstruction loss. Intermediate loss, which is the mean square error between input and 1-sample-shifted output of "Intermediate WaveNet," evaluates how well the Encoder (periodic signal extractor) generated the intermediate parameters. Reconstruction loss evaluates how well the input speech waveform was reproduced. An output sequence with the same periodicity as the intermediate parameters can be expected during synthesis by inputting arbitrary intermediate parameters to the Decoder (speech generator).

The speech waveform is assumed to be the sum of periodic and aperiodic waveforms in this framework, so the Decoder (speech generator) output consists of 25 channels: a single channel of the periodic waveform, and 24 channels of the aperiodic power in each frequency band. Reproduction of the speech waveform is shown on the right in Figure 4. First, 24 channels of aperiodic power for each frequency band, which represent the speech aperiodic waveform, are split from the 25 channel signals output from the Decoder (speech generator), and are multiplied by Gaussian noise partitioned into each frequency band. The speech waveform can be generated by summing the speech aperiodic waveforms of all frequency band and the speech periodic waveform.

The related approaches which use periodic signals as input have been proposed [14, 15]. In the training process of these approaches, short-time Fourier transform (STFT) include framing and windowing is used. Since STFT is not used in the proposed model training, STFT parameters such as FFT size, frame shift, window type and so on are not required to be considered.

2.2. Reconstruction loss

Two types of loss were used to evaluate reproduction of the speech waveform.

Gauss loss The aperiodic waveform is obtained by taking the difference between the natural speech signal and the single-channel periodic waveform from the 25 Decoder (speech generator) output channels. Under the zero-mean Gaussian distributions which use the other 24

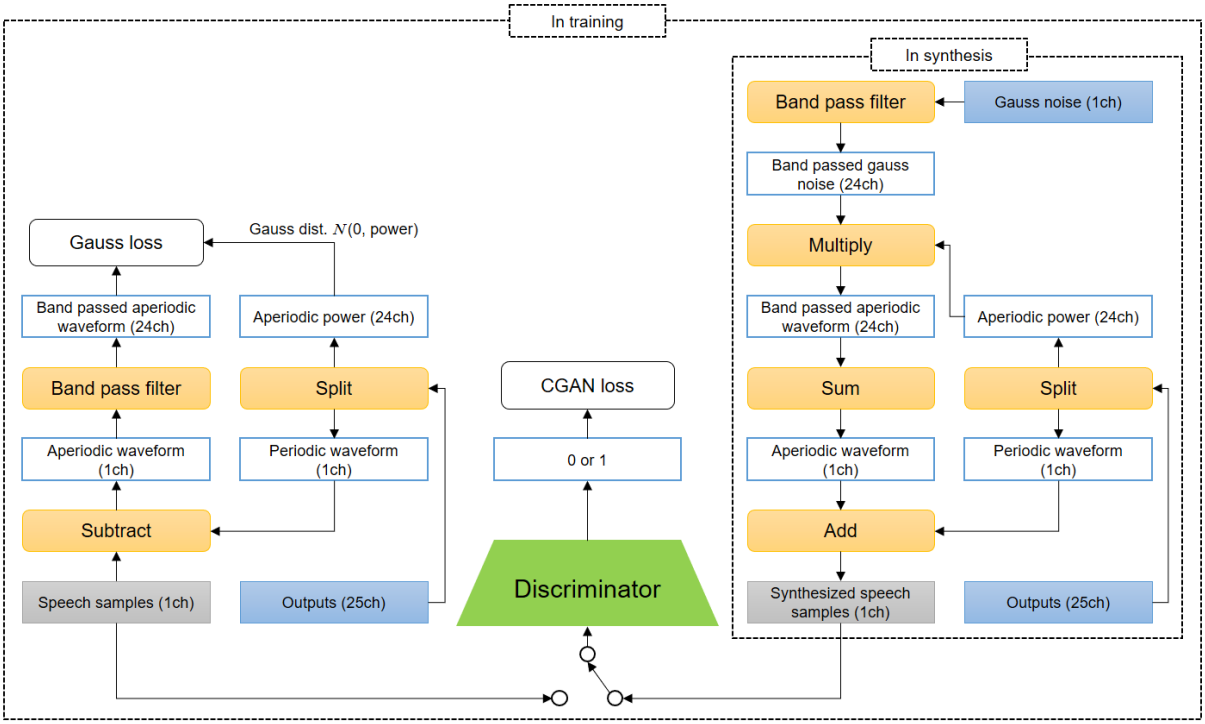

Figure 4: *Reconstruction loss calculation*

channels as standard deviation, the log likelihood per band is calculated sample-by-sample for the band-passed aperiodic waveform.

CGAN loss A generative adversarial network (GAN) [12, 13] framework is introduced. In this manner, a Discriminator estimates the probability that speech samples came from the training data rather than Decoder (speech generator). The Encoder/Decoder are trained to maximize the probability of the Discriminator making a mistake. Note that updates alternate between the Discriminator and Encoder/Decoder.

2.3. Model structure

Detailed structure of WaveNet [2], "Intermediate WaveNet," Encoder/Decoder, and Discriminator are shown in Figures 5, 6, 7, and 8, respectively. "Intermediate WaveNet" to evaluate the output of the Encoder (periodic signal extractor) in the proposed method have three channels for output instead of the 256 channels for the discrete probability distribution in WaveNet. WaveNet and Parallel WaveNet use past speech sample outputs from the model and a noise sequence as inputs respectively, while the proposed model explicitly uses a periodic signal (three channels: sine, cosine, and whether voiced or not) as inputs. Since the proposed model does not have autoregressive structure, there is no need to restrict input to past data, and filters have left-right symmetry on all dilated convolution layers, as shown in Figure 7. To output the current speech sample, both past and future inputs are used. The Encoder/Decoder merges the outputs from the highest dilated convolution layer with skip layer information from all layers. With Parallel WaveNet, a periodic signal must be generated from noise, so more Inverse-Autoregressive Flow (IAF) layers than WaveNet dilated convolution layers are needed. However, with the proposed model, a periodic signal is input explicitly, so it has the same number of dilated convolution layers as WaveNet. During synthesis, WaveNet selects a μ-law class by sampling based on the output discrete probability distribution, but for the proposed model, the speech waveform is generated by using the 25 channels of continuous values (Figure 4, right).

3. Experiments

To show the effectiveness of the proposed method, we compared modeling performance using a database of 70 Japanese children songs. We used 60 songs (approx. 65 min.) for training, and ten songs (approx. 6.4 min.) for testing. The speech format is mono-channel with 48 kHz sampling frequency and 16 bit quantization. The structure of WaveNet, "Intermediate WaveNet," Encoder/Decoder, and Discriminator are shown in Figures 5, 6, 7, and 8, respectively. The auxiliary features used for them included 50 dimensions of mel cepstrum coefficients derived using WORLD [16], 50 dimensions of aperiodic measures, the log F0, and voiced/unvoiced data, totaling 102 dimensions.

To assist training, we performed initial training of the Encoder/Decoder. REAPER [17] was used to extract glottal closure instants, these were then used to generate sine/cosine waves, which in-turn, were used to initialize both the Encoder

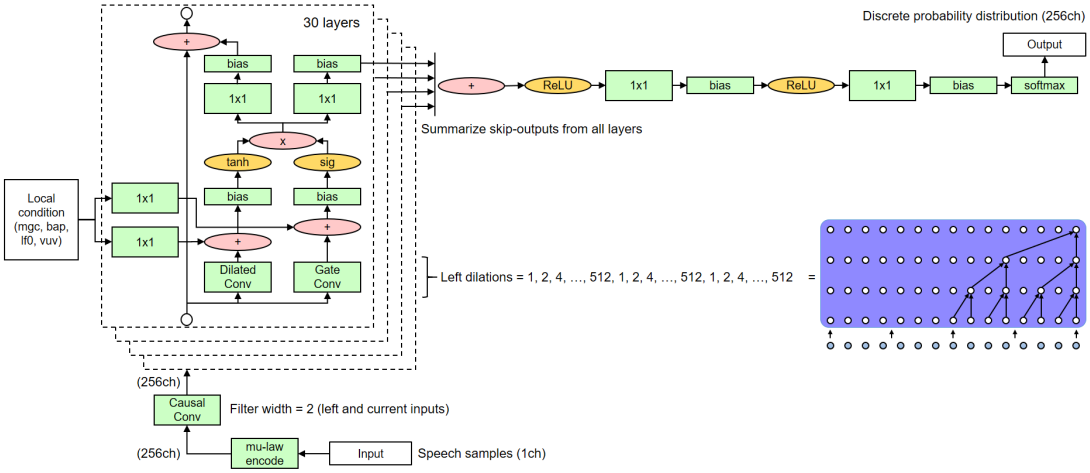

Figure 5: *WaveNet structure*

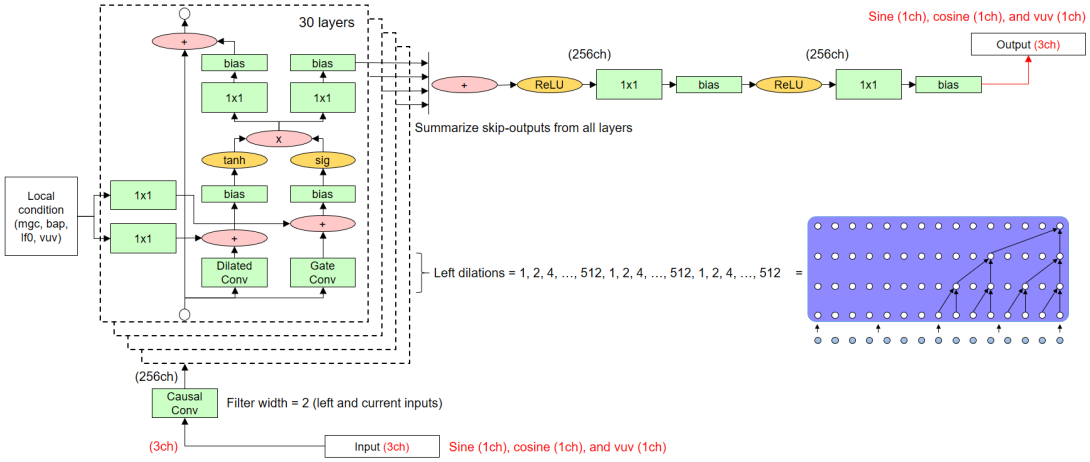

Figure 6: *Intermediate WaveNet structure of the proposed framework*

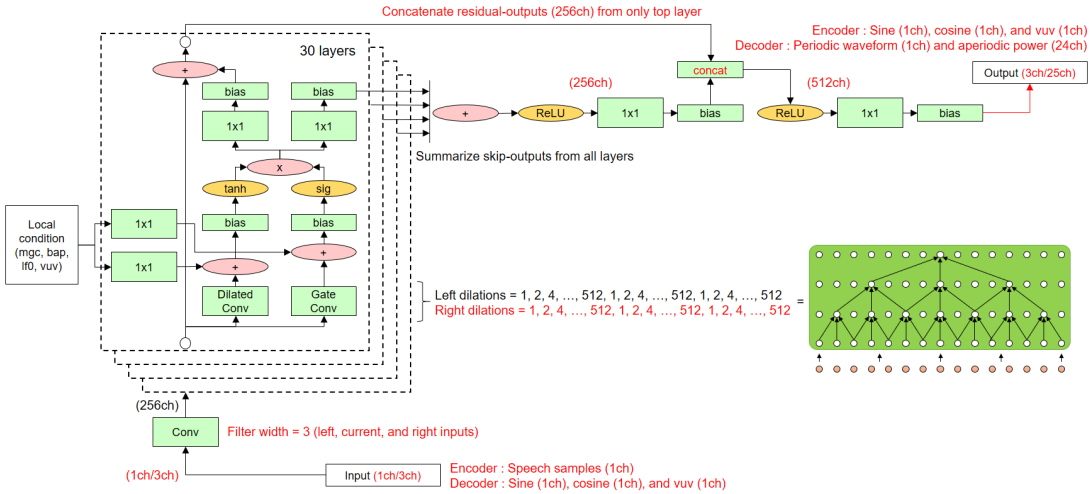

Figure 7: *Encoder/Decoder structure of the proposed framework*

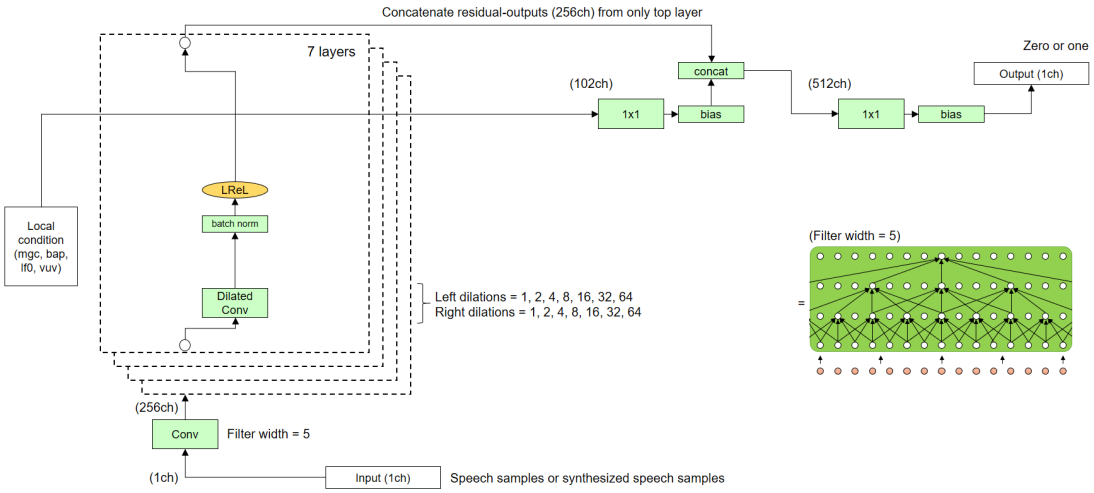

Figure 8: *Discriminator structure of the proposed framework*

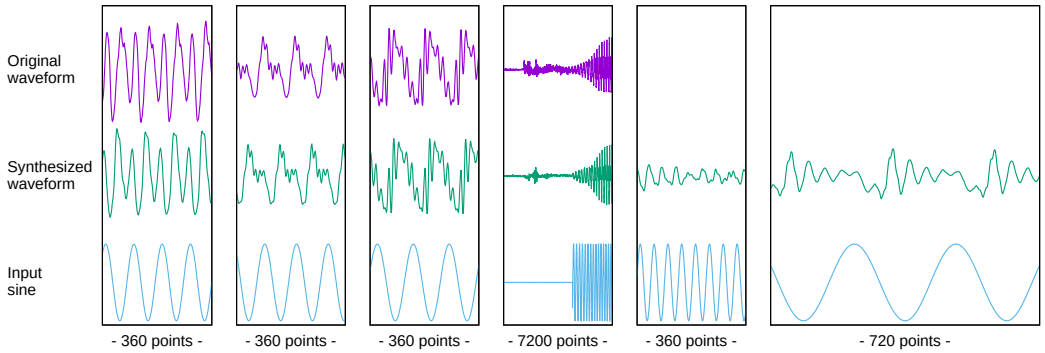

Figure 9: *Original waveform (upper), synthesized waveform (middle), and input sine (lower). /a/, /o/, /i/, and /ko/ in the test data are shown from the left. /e/ and /a/ exceeded under/over the pitch range are shown from the right.*

(speech generator) and Decoder (periodic signal extractor) separately. The Adam optimizer is used with learning rate of 0.001. We measured the time required for synthesis using the proposed method, and showed that using an NVIDIA GTX 1080, speech was synthesized at approximately five-times faster than real-time.

Waveforms generated using the proposed method are shown in Figure 9. The upper waveform is the natural waveform, the middle is synthesized waveform, and the lower is the periodic sine wave. The four figures on the left show that the synthesized waveform had the same pitch of the input periodic signal and had similar contour to the natural waveform in the test data. The two figures on the right show results of inputting periodic signals outside of the range of the training data, with test data one octave higher and lower respectively. The synthesized waveform was less natural, but the pitch was maintained, as can be seen in the figures.

We conducted a subjective evaluation with 16 subjects, selecting data randomly from among ten phrases per method and having the subjects listen to them and evaluate them for naturalness on a five-level scale. With the proposed method, we used two methods: Gauss loss only and the sum of Gauss loss and CGAN loss as Reconstruction loss.

The evaluation results are shown in Figure 10. The figure shows that the proposed method achieves higher naturalness than WaveNet. It also shows that adding CGAN loss to Gauss loss does not improve naturalness for the proposed method. There is some possibility that Discriminator take notice of whether waveform are quantized instead of the discrimination of natural waveform and synthesized waveform. Although the both inputs of the Discriminator are 32 bit floating point number, natural waveform input is converted from the corpus of 16 bit signed integer.

4. Conclusions

This paper has proposed a framework for a real-time speech vocoder based on neural networks with periodic and aperiodic inputs. By using periodic and aperiodic signals as input, the proposed method generates the speech sample sequence at once. The proposed method is able to generate speech waveforms

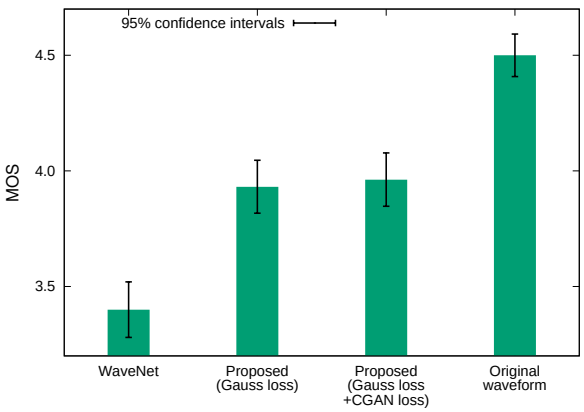

Figure 10: *Mean opinion scores for the naturalness*

faster than real-time, and can generate waveforms with pitch outside of the range of the training data. In subjective evaluations, we confirmed that the method produces better quality speech than WaveNet. Future work include test using a larger database.

5. Acknowledgements

Casio Science Promotion Foundation, Kayamori Foundation of Informational Science Advancement, JSPS Grant JP18K11163, JSPS Grant JP19H04136.

6. References

[1] A. van den Oord, *et. al.* "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[2] A. Tamamori, *et. al.* "Speaker-dependent WaveNet vocoder," *INTERSPEECH* 2017, pp. 1118–1122, 2017.

[3] J. Shen, *et. al.* "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017.

[4] J. Niwa, *et. al.* "Statistical voice conversion based on WaveNet," *ICASSP* 2018, pp. 5289–5293, IEEE, 2018.

[5] K. Sawada, *et. al.* "The nitech text-to-speech system for the blizzard challenge 2018," *Blizzard Challenge 2018 Workshop*, 2018.

[6] Y. Chiao Wu, *et. al.* "Quansi-Periodic WaveNet Vocoder: A Pitch Dependent Dilated Convolution Model for Parametric Speech Generation," *CoRR*, vol. abs/1907.00797, 2019.

[7] N. Kalchbrenner, *et. al.* "Efficient Neural Audio Synthesis," *CoRR*, vol. abs/1802.08435, 2018.

[8] J.-Marc Valin, *et. al.* "LPCNET: Improving neural speech synthesis through linear prediction," *ICASSP* 2019, pp. 5891–5895, IEEE, 2019.

[9] A. van den Oord, *et. al.* "Parallel WaveNet: Fast high-fidelity speech synthesis," *CoRR*, vol. abs/1711.10433, 2017.

[10] W. Ping, *et. al.* "ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech," *CoRR*, vol. abs/1807.07281, 2018.

[11] R. Prenger, *et. al.* "WaveGlow: A Flow-based Generative Network for Speech Synthesis," *CoRR*, vol. abs/1811.00002, 2018.

[12] I. Goodfellow, *et. al.* "Generative adversarial nets," in *NIPS*, 2014.

[13] M. Mirza, *et. al.* "Conditional Generative Adversarial Nets," CoRR, vol. abs/1411.1784, 2014.

[14] X. Wang, *et. al.* "Neural source-filter-based waveform model for statistical parametric speech synthesis," *ICASSP* 2019, pp. 5916–5920, IEEE, 2019.

[15] O. Watts, *et. al.* "Speech waveform reconstruction using convolutional neural networks with noise and periodic inputs," *ICASSP* 2019, pp. 7045–7049, IEEE, 2019.

[16] M. Morise, *et. al.* "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. E99. D, no. 7, pp. 1877–1884, 2016.

[17] "REAPER: Robust epoch and pitch estimator," https://github.com/google/REAPER.