MULTI-SAMPLE SUBBAND WAVERNN VIA MULTIVARIATE GAUSSIAN

Hiroki Kanagawa and Yusuke Ijima

NTT Corporation

ABSTRACT

This paper proposes a high-speed neural vocoder for CPU implementation. Two approaches for speeding up autoregressive neural vocoders have been proposed, 1) simultaneous multiple sample generation and 2) subband signal-based vocoder; so far they have been employed independently. Our neural vocoder is extremely fast as it generates multiple samples of subband signals simultaneously. Although there is an association between each subband signal, the conventional subband-based vocoder can degrade quality because each subband signal is generated from an independent probability distribution. To overcome this problem, we also introduce waveform generation that takes account of the association of each subband by employing multivariate Gaussian. Experiments show that 1) our proposed method is 1.81 times as fast as the conventional subband WaveRNN on a single-threaded CPU; 2) it outperformed the conventional method in a subjective evaluation in terms of naturalness, and achieved a mean opinion score (MOS) of 4.08 on text-to-speech task.

***Index Terms*—** neural vocoder, subband signal, multi-sample generation, multivariate Gaussian, speech synthesis

1. INTRODUCTION

The quality of text-to-speech (TTS) has been greatly enhanced by the recent adoption of neural vocoders. WaveNet is the first successful model; it predicts the probability distribution of speech samples by a large auto-regressive model with dilated causal convolution [1]. In order to overcome WaveNet's drawback of slow inference speed, several techniques have been proposed, including Parallel WaveNet, Clarinet, WaveGlow, Parallel WaveGAN [2, 3, 4, 5]. Although these models are faster, parallel processors such as GPUs are required for fast inferencing. In order to overcome the GPU requirement, alternative models for CPU implementation have been proposed. WaveRNN achieves real-time operation on a CPU by replacing WaveNet's huge dilated causal convolution with a simple RNN [6]. LPCNet, ExcitNet, and LP-WaveNet have also introduced signal processing knowledge and can realize even smaller models than WaveRNN [7, 8, 9]. Recently, for further speed enhancement, a new method has been proposed that predicts a shortened subband signal by Pseudo quadrature mirror filter (PQMF) [10], instead of handling fullband speech signals. Subband WaveRNN and FeatherWave have applied PQMF to WaveRNN and LPCNet, and have more than doubled the speed by reducing the sequence length to be predicted [11, 12].

As another approach to reduce the sequence length to be predicted, multiple samples can be generated in a single forward propagation step. For example, an extension of LPCNet for multiple sample generation has been proposed [13]. This method achieves 1.5 times speed-up by not only generating multiple excitation signals at the same time but also replacing the DNN target with a Gaussian distribution of raw waveforms instead of bit-value categorical distribution. [14] also models the joint-probability of multiple samples by Laplace distribution, and succeeded in simultaneous generation more than two samples on WaveNet. The combination of these speed-up methods can be used to generate multiple subband signals simultaneously, but this has yet to be explored. It is also unclear how many samples can be generated simultaneously while still maintaining quality.

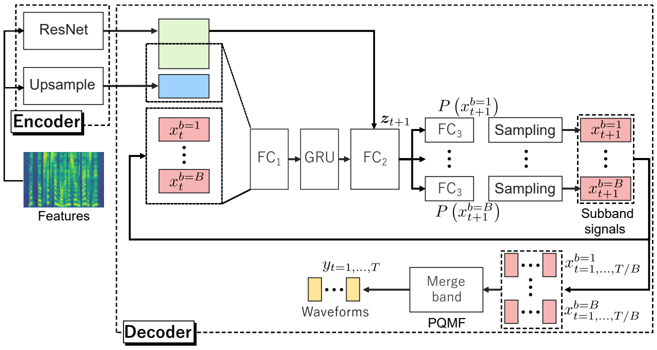


Fig. 1. Overview of subband WaveRNN.

In order to investigate these issues, we propose a fast neural vocoder that can generate multiple subband signals simultaneously. The number of samples to be generated simultaneously and the resulting inference speeds are investigated. In addition, the conventional subband WaveRNN generate each subband signal from independent distributions, thus the quality may be degraded due to the lack of their association. So this work also propose the subband signals' joint-modeling via a multivariate Gaussian to take into account these associations. Our proposed two-sample simultaneous generation vocoder via multivariate Gaussian is 1.81 times faster than the conventional subband WaveRNN on a single-threaded CPU, and its real-time factor (RTF) was less than 0.1. A subjective evaluation for naturalness, including ground truth, showed that the proposed vocoder achieved a mean opinion score (MOS) of 4.08 for acoustic features predicted by TTS. These results confirm the effectiveness of using multivariate Gaussian and that no degradation in quality is observed even if up to two samples are generated simultaneously.

2. SUBBAND WAVERNN

Subband WaveRNN [11] utilizes PQMF to convert a fullband signal with T sample length into a subband signal divided into B bands; this reduces the sequence length to be predicted to T/B. Figure 1 overviews subband WaveRNN. The model mainly consists of an encoder and a decoder, which operate on a frame-by-frame and sample-by-sample basis, respectively. The encoder upsamples the acoustic features and maps them to corresponding frames and samples. The decoder predicts the next time's subband signal by taking the encoder's output and the previous time's subband signal. Unlike the

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

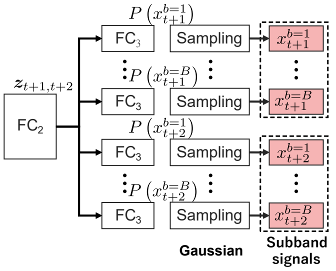

$P\left(x_{t+1}^{b=1}\right)$
$z_{t+1,t+2}$
FC$_2$
FC$_3$
Sampling
$x_{t+1}^{b=1}$
$P\left(x_{t+1}^{b=B}\right)$
$x_{t+1}^{b=B}$
$P\left(x_{t+2}^{b=1}\right)$
$x_{t+2}^{b=1}$
$P\left(x_{t+2}^{b=B}\right)$
$x_{t+2}^{b=B}$
Gaussian
Subband signals

Fig. 2. The decoder of proposed multi-sample subband WaveRNN. This model can generate two subband signals in a single forward propagation step.

original WaveRNN, the decoder's prediction target is the subband signal $x_t^b \forall_t \in [1, T/B], \forall_b \in [1, B]$, where t, b are the time's index and the band's one, respectively. These subband signals are upsampled by feeding them to PQMF to indirectly obtain the speech waveform $y_t \forall_t \in [1, T]$. Since we use a Gaussian distribution as subband signal probability distribution $P\left(x_t^b\right)$, the loss function is defined by negative log likelihood (NLL) during training as:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T/B} \sum_{b=1}^{B} \ln \mathcal{N}\left(x_t^b; \mu^b\left(\boldsymbol{z}_t|\theta\right), \sigma^b\left(\boldsymbol{z}_t|\theta\right)\right), \quad (1)$$

where θ, $\boldsymbol{z}_t$, μ^b, and σ^b are the DNN model parameters, FC2's output, and mean and standard deviation of $P\left(x_t^b\right)$, respectively.

3. PROPOSED MULTI-SAMPLE SUBBAND VOCODER VIA MULTIVARIATE GAUSSIAN

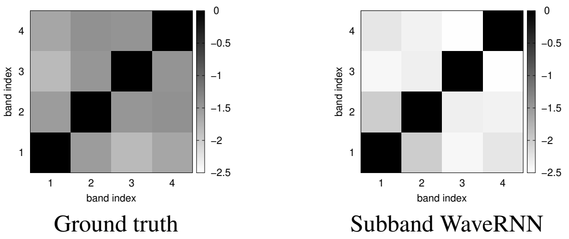
3.1. Multi-sample subband WaveRNN

Figure 2 shows the difference between the proposed decoder and the conventional subband WaveRNN (Fig. 1). Unlike the conventional subband WaveRNN, the feed-forward (FC3) corresponding to each band signal for time $t+1$ and $t+2$ is added after FC2 to generate multiple subband signals in a single forward propagation step. Multiple subband signals are generated by sampling from the probability distribution $P\left(x_\tau^b\right) \forall_\tau \in [t+1, t+2], \forall_b \in [1, B]$ obtained from FC3.

3.2. Multivariate Gaussian Prediction

The subband signals are considered to have associations because there are overlaps in the bands of the PQMF. However, it is difficult to analyze them with simple linear correlations. To analyze the associations between band signals, we use the maximum information coefficient (MIC) [15] instead of linear correlation in Fig. 3 to show the non-linear associations between subband signals. These MICs were obtained from the ground truth and synthesized speech by Subband WaveRNN, respectively. Since the conventional subband WaveRNN utilizes independent distributions for each subband signal modeling, generated signals would not be taken the association among subbands into account. Consequently, generated speech signals by the conventional method lost the association among subbands, nature of ground truth signal. Therefore, jointly modeling these association may lead to the quality improvement of subband WaveRNN.

To address this issue, our proposed vocoder predicts all subband signals simultaneously using the multivariate Gaussian as a probability distribution of subband signals. Figure 4 presents the difference between the proposed vocoder and the conventional one (Fig. 1). Let $\boldsymbol{x}_t \in \mathbb{R}^B$ denote all subband signals at time t, and modify FC3 to predict the joint probability $P\left(\boldsymbol{x}_\tau\right) \forall_\tau \in [t+1, t+2]$. At training time, the loss function is given by:



Ground truth

Subband WaveRNN

Fig. 3. The heatmap of maximum information coefficient (MIC) [15] between subband signals. These MICs were obtained from the ground truth and synthesized speech by subband WaveRNN. To make it easier to see the differences between systems, MICs are taken the logarithmic. Note that subband WaveRNN had off-diagonal components with smaller value than ground truth.

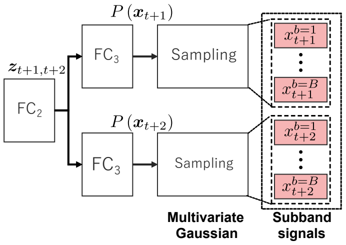


Fig. 4. The decoder of proposed multi-sample subband WaveRNN via multivariate Gaussians. This vocoder aims to model the association between subband signals.

$$\mathcal{L}'(\theta) = -\sum_{t=1}^{T/B} \ln \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}\left(\boldsymbol{z}_t, \theta\right), \boldsymbol{\Sigma}\left(\boldsymbol{z}_t, \theta\right)\right), \quad (2)$$

where $\boldsymbol{\mu} \in \mathbb{R}^B$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{B \times B}$ are the mean vector and covariance matrix of the multivariate Gaussian distribution, respectively.

3.3. Computation complexity

The neural vocoder's computation complexity is mainly created by the decoder's DNN module. Its basic tasks are the operations of addition and multiplication, and floating-point operations (FLOPs) of our vocoder proposed in section 3.1 is given by:

$$\begin{aligned} C = \big[d\left\{\left(D_{\text{in}1} + D_{\text{in}2}\right) N_\alpha + 3N_\alpha^2 + \left(D_{\text{in}2} + N_\alpha\right) N_\beta\right\} \\ + N_\beta D_{\text{out}} BM\big] \times \frac{f_s}{BM}, \end{aligned} \quad (3)$$

where N_α and N_β are unit sizes of GRU and FC2, respectively. $D_{\text{in}1}$, $D_{\text{in}2}$, and D_{out} are the Upsample's dimension, the half dimension of ResNet, and FC3's output dimension, respectively. d, f_s, and M are the density of the sparse DNN, the sampling frequency, and the number of simultaneous generated subband signals, respectively. When $M = 1$, its complexity is equal to that of a conventional subband WaveRNN. In particular, when a Gaussian distribution is used as $P\left(x_t^b\right)$, $D_{\text{out}} = 2$ with mean and logarithmic standard deviation.

When FC3 predicts the mean vector and the lower triangular covariance of a multivariate Gaussian distribution, the computational complexity of the vocoder proposed in section 3.2 can likewise be

Table 1. The compared methods, their probability distributions and the number of multiple samples to be generated.

Method	M	Distribution
Conventional (M=1, w/o MV)	1	Gaussian
Proposed (M=1, w/ MV)	1	Multivariate Gaussian
Proposed (M=2, w/o MV)	2	Gaussian
Proposed (M=2, w/ MV)	2	Multivariate Gaussian
Proposed (M=4, w/ MV)	4	Multivariate Gaussian

reformulated as follows:

$$C' = \Big[d\left\{\left(D_{\text{in}1} + D_{\text{in}2}\right) N_\alpha + 3N_\alpha^2 + \left(D_{\text{in}2} + N_\alpha\right) N_\beta\right\} + N_\beta \left(B + \frac{B}{2}\left(B+1\right)\right) M\Big] \times \frac{f_s}{BM}. \qquad (4)$$

Using $d = 0.4$, $N_\alpha = 256$, $N_\beta = 128$, $B = 4$, $f_s = 22050$, $D_{\text{in}1} = 80$, $D_{\text{in}2} = 64$, and $M = 1$, Eqs (3) and (4) are 0.607 and 0.609 GFLOPs, respectively. We can see that the computational complexity does not significantly increase (only 0.3%) even if multivariate Gaussian is employed.

4. EXPERIMENTS

4.1. Setup

We used speech data uttered by a Japanese professional female speaker. The sampling frequency was 22.05 kHz. Ninety utterances were extracted as evaluation data (5.4 minutes), and the others were used for training and validation (9.6 hours).

Eighty-dimensional logarithmic mel-spectrograms were used as the conditioning feature of the neural vocoder. The analysis frame shift was 5 ms [1]. Since the subband signal's amplitudes significantly differ in each band, the amplitude was flattened by calculating the subband signals after pre-emphasis of the speech waveform at training time. At synthesis time, subband signals were converted into speech waveforms by PQMF, and de-emphasis was applied. Both pre-emphasis and de-emphasis used the coefficient of 0.97.

The number of training steps was 5000k. For fast vocoding, we performed pruning [16] in the same manner as WaveRNN using:

$$d_s = d\left[1 - \left\{1 - \left(s - s_0\right)/S\right\}^3\right], \qquad (5)$$

where $s_0 = 2000$k, $S = 2500$k and s is the training step's index. In order to utilize vector operations by block sparsification [17], the pruning block sizes were set to 4×1 for FC1 and 16×1 for GRU and FC2. The parameters for the computational complexity are the same ones described in section 3.3. The ResNet of the encoder has ten residual blocks, which consist of 1D-convolution with 128 units, batch normalization, and activation. ReLU was used for all activations of the vocoder. To guarantee spectral reproducibility, we also calculated the STFT loss for the speech and subband signals as studied in [18]. These losses were added to Eqs. (1) and (2) without scaling. The vocoder's optimization was performed using RAdam [19], with $\alpha = 1.0 \times 10^{-4}$, $\beta = (0.9, 0.999)$, and $\varepsilon = 1.0 \times 10^{-8}$.

Table 1 shows the compared methods, their probability distributions used for sampling, and the number of samples generated simultaneously. "M" and "w/ MV" in the table denote the number of simultaneous generated samples and the use of a multivariate Gaussian, respectively. When M was greater than 2, the variance often failed to be predicted and clicking sounds were observed. To avoid this problem, we 1) eliminate variance outliers and 2) clip sampled results to $\mu \pm 3\sigma$ in a similar way as [13].

Table 2. Average RTFs obtained from all evaluation data. "Speed enhancement" denotes the improvement over conventional subband WaveRNN.

Method	RTF	Speed enhancement
Conventional (M=1, w/o MV)	0.170	-
Proposed (M=1, w/ MV)	0.171	1.00x
Proposed (M=2, w/o MV)	0.094	1.81x
Proposed (M=2, w/ MV)	0.094	1.81x
Proposed (M=4, w/ MV)	0.052	3.27x

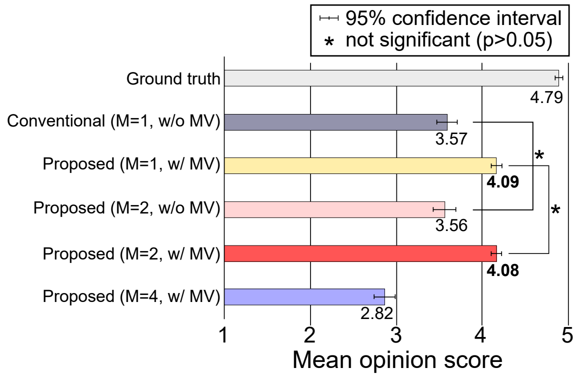


Fig. 5. Mean opinion scores of naturalness of synthetic speech. Acoustic features are predicted by Tacotron2 [20].

4.2. Speed comparison

The real-time factors (RTFs) were calculated to measure the inference speeds of the conventional subband WaveRNN and our proposed vocoder. The RTF definition is given by:

$$\text{RTF} = T_{\text{inference}}/T_{\text{data}}, \qquad (6)$$

where T_{data} and $T_{\text{inference}}$ are speech length and single-thread inference time measured on an Intel Core i7-8750H CPU 2.20 GHz, respectively. Table 2 shows method, averaged RTFs from all evaluation data, and RTF improvements.

A comparison with the conventional methods, Conventional (M=1, w/o MV) and Proposed (M=1, w/ MV), shows that the use of multivariate Gaussian distribution has no impact on the RTF. This is a reasonable result since the computational complexity of the decoder does not increase significantly as discussed in section 3.3. We also confirmed that M values of 2 and 4 improve the speed by 1.81x and 3.27x, respectively, from $M = 1$. Furthermore, these systems achieved extremely fast vocoding with RTF below 0.1.

4.3. Subjective evaluation

To investigate our proposed vocoders' performance, we conducted the subjective evaluation using acoustic features predicted by TTS. In order to predict mel-spectrograms from evaluation texts, we used NVIDIA's implementation[2] of Tacotron2 [20]. The input sequence

[1] Although we also investigated the commonly used frame shift of 12.5 ms in our preliminary experiments, we chose to set it to 5 ms because it reproduced better the pitch of synthetic speech.

[2] https://github.com/NVIDIA/tacotron2.git

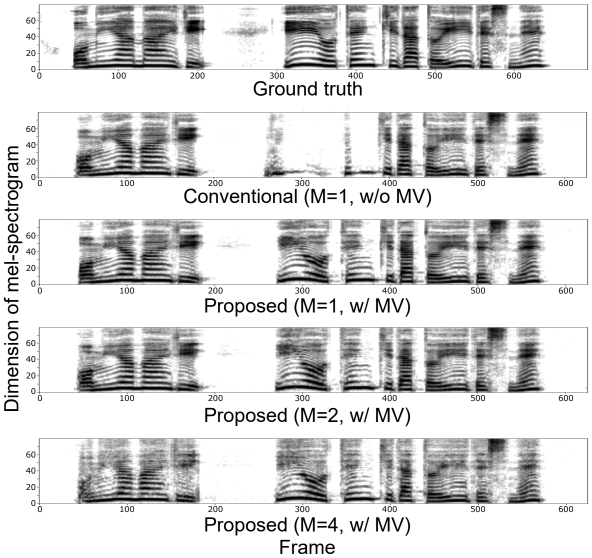


Fig. 6. The comparison of mel-spectrograms of synthetic speech. From top to bottom: Ground truth, Conventional (M=1, w/o MV), Proposed (M=1, w/ MV), (M=2, w/ MV), and (M=4, w/ MV). Conventional (M=1, w/o MV) struggled in some cases with silence, whereas Proposed (M=1, w/ MV) and (M=2, w/ MV) were still able to reproduce spectrograms close to Ground truth. Proposed (M=4, w/ MV) was hard to achieve in the mid and high-frequencies due to sample discontinuities.

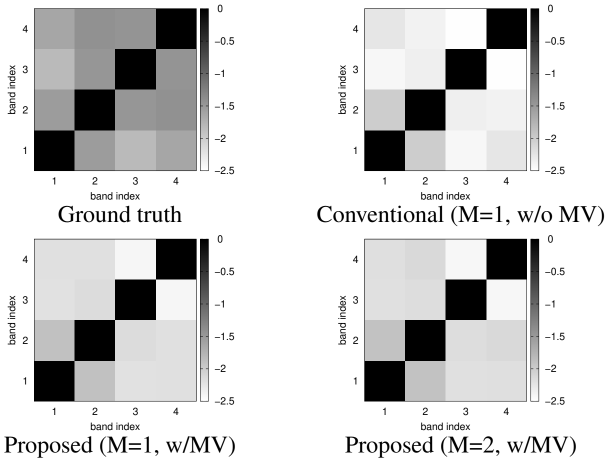


Fig. 7. The comparison of logarithmic MICs obtained from subband signals. From top left to bottom right are those of Ground truth, Conventional (M=1, w/o MV), Proposed (M=1, w/ MV), and (M=2, w/ MV). Proposed (M=1, w/ MV) and (M=2, w/ MV) had larger off-diagonal components than Conventional (M=1, w/o MV), indicating that the multivariate Gaussian was effective for the association modeling between subband signals.

consisted of 58 phonemes including start, end, pause, and accent position symbols. Guided attention loss [21] was also incorporated to stabilize the training, and the model was trained in 500k steps.

We subjectively evaluated the naturalness of synthetic speech by using a mean opinion score (MOS) on a five-point scale ranging from 5: very natural to 1: very unnatural. Thirty of all evaluation utterances were randomly chosen for each method. Fifteen listeners participated in the test, each of them giving scores for synthetic speech. Figure 5 shows the preference test results including ground truth. Proposed(M=1, w/ MV) outperformed Conventional(M=1, w/o MV) and achieved MOS 4.08$\pm$0.06, where $\pm$ denotes 95% confidence interval. This result confirms the effectiveness of multivariate Gaussian distribution in simultaneous subband signal sampling. There was no significant differene between Conventional (M=1, w/o MV) and Proposed (M=2, w/o MV), or between Proposed (M=1, w/ MV) and (M=2, w/ MV). This shows that the simultaneous generation of two samples does not degrade the naturalness of synthetic speech. On the other hand, Proposed (M=4, w/ MV) drastically degraded quality. We will discuss in the next section the reasons behind this subjective evaluation result through spectrogram observations and subband associations analyses.

4.4. Discussion

Figure 6 shows a comparison of the mel-spectrograms of ground truth and synthesized speech. Conventional (M=1, w/o MV) sometimes suffered from silence, as shown around the 300th frame in Fig. 6. This was due to the improbable subband signals prediction ignoring their associations, and PQMF was unable to reconstruct it into the speech waveform. On the other hand, Proposed (M=1, w/ MV) could reproduce mel-spectrograms relatively close to those of the ground truth, and Proposed (M=2, w/ MV) also worked as well as Proposed(M=1, w/ MV). Proposed(M=4, w/ MV) was hard to reproduce the mid- and high-frequency components, which might have contributed to its lower quality. Although this reason may be led by predicting too many samples without considering the association between each sample, we believe that it can be improved by joint-modeling between samples as proposed in [14].

We also investigated the contribution of the multivariate Gaussian to the MIC of subband signals. Figure 7 shows the MICs of the subbands obtained from the evaluation sets of Ground truth, Conventional (M=1, w/o MV), Proposed (M=1, w/ MV), and (M=2, w/ MV). Conventional (M=1, w/o MV) had off-diagonal components with smaller value than Ground truth. Whereas the off-diagonal component of Proposed (M=1, w/ MV) had a larger value, indicating that the association between bands could be modeled by a multivariate Gaussian distribution. In addition, since the results of Proposed (M=2, w/ MV) were similar to those of Proposed (M=1, w/ MV), we can confirm multivariate Gaussian is still effective in combination with multiple sample generation. These results revealed that Proposed (M=1, w/ MV) and (M=2, w/ MV) could improve the representation of the association between subband signals, resulting in improved synthesized speech quality.

5. CONCLUSION

In this work, we proposed the subband WaveRNN-based neural vocoder incorporating multi-sample generation for faster inference on a CPU. In addition, focusing on the fact that the association between subband signals is lost in the conventional subband WaveRNN, we also proposed their joint-modeling via multivariate Gaussian to improve synthesized subband signals' quality. Speed test shows that the proposed method with two-sample simultaneous generation performed 1.81 times faster than the conventional method, demonstrating the effectiveness of generating multiple-sample subband signals. The subjective evaluation also confirmed that 1) there was no quality degradation in the two-sample simultaneous generation, and 2) the subband signals' joint-modeling via multivariate Gaussian improved the quality compared to the conventional method.

6. REFERENCES

[1] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[2] Aäron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Van Den Driessche, Edward Lockhart, Luis C. Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," *Proc. ICML*, vol. 9, 2018.

[3] Wei Ping, Kainan Peng, and Jitong Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *Proc. ICLR*, 2019.

[4] Bryan Catanzaro Ryan Prenger, Rafael Valle, "WaveGlow: A flow-based generative network for speech synthesis," *Proc. ICASSP*, pp. 3617–3621, 2019.

[5] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," *Proc. ICASSP*, pp. 6199–6203, 2020.

[6] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron Van Den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient Neural Audio Synthesis," *Proc. PMLR*, pp. 2410–2419, 2018.

[7] Jean-Marc Valin and Jan Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," *Proc. ICASSP*, pp. 5891–5895, 2019.

[8] Eunwoo Song, Kyungguen Byun, and Hong-Goo Kang, "ExcitNet vocoder: A neural excitation model for parametric speech synthesis systems," *Proc. EUSIPCO*, pp. 1–5, 2019.

[9] Min-Jae Hwang, Frank Soong, Eunwoo Song, Xi Wang, Hyeonjoo Kang, and Hong-Goo Kang, "LP-WaveNet: Linear prediction-based wavenet speech synthesis," *Proc. APSIPA*, pp. 810–814, 2020.

[10] Truong Q. Nguyen, "Near-perfect-reconstruction pseudo-QMF banks," *IEEE Trans. Speech and Audio Processing*, vol. 42, no. 1, pp. 65–76, 1994.

[11] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, Dan Su, and Dong Yu, "DurIAN: Duration informed attention network for speech synthesis," *Proc. INTERSPEECH*, pp. 2027–2031, 2020.

[12] Qiao Tian, Zewang Zhang, Heng Lu, Ling Hui Chen, and Shan Liu, "FeatherWave: An efficient high-fidelity neural vocoder with multi-band linear prediction," *Proc. INTERSPEECH*, pp. 195–199, 2020.

[13] Vadim Popov, Mikhail Kudinov, and Tasnima Sadekova, "Gaussian LPCNet for multisample speech synthesis," *Proc. ICASSP*, pp. 6204–6208, 2020.

[14] Patrick Lumban Tobing, Yi-Chiao Wum, Tomoki Hayashi, Kazuhiro Kobayashi, and Tomoki Toda, "Efficient shallow wavenet vocoder using multiple samples output based on laplacian distribution and linear prediction," *Proc. ICASSP*, pp. 7204–7208, 2020.

[15] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[16] Michael Zhu and Suyog Gupta, "To prune, or not to prune: exploring the efficacy of pruning for model compression," *Proc. ICLR*, 2018.

[17] Sharan Narang, Eric Undersander, and Gregory Diamos, "Block-sparse recurrent neural networks," *arXiv preprint arXiv:1711.02782*, 2017.

[18] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," *Proc. SLT*, pp. 492–498, 2021.

[19] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han, "On the variance of the adaptive learning rate and beyond," *Proc. ICLR*, 2020.

[20] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *Proc. ICASSP*, pp. 4779–4783, 2018.

[21] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," *Proc. ICASSP*, 2018.