

Bunched LPCNet : Vocoder for Low-cost Neural Text-To-Speech Systems

Ravichander Vipperla^{1†}, Sangjun Park^{2†}, Kihyun Choo^{2†}, Samin Ishtiaq¹, Kyoungbo Min²
Sourav Bhattacharya¹, Abhinav Mehrotra¹, Alberto Gil C. P. Ramos¹, Nicholas D. Lane^{1,3}

¹Samsung AI Centre, Cambridge, UK

²Samsung Research, Seoul, Republic of Korea

³University of Cambridge, UK

{r.vipperla, sj0.park, khchoo, s.ishtiaq, kyoungbo.min,
sourav.bl, a.mehrotra1, a.gilramos, nic.lane}@samsung.com

Abstract

LPCNet is an efficient vocoder that combines linear prediction and deep neural network modules to keep the computational complexity low. In this work, we present two techniques to further reduce its complexity, aiming for a low-cost LPCNet vocoder-based neural Text-to-Speech (TTS) System. These techniques are: 1) Sample-bunching, which allows LPCNet to generate more than one audio sample per inference; and 2) Bit-bunching, which reduces the computations in the final layer of LPCNet. With the proposed bunching techniques, LPCNet, in conjunction with a Deep Convolutional TTS (DCTTS) acoustic model, shows a 2.19x improvement over the baseline run-time when running on a mobile device, with a less than 0.1 decrease in TTS mean opinion score (MOS).

Index Terms: Neural Text-to-Speech, vocoder, LPCNet, Sample Bunching, Bit Bunching

1. Introduction

LPCNet [1] is a state-of-the-art light-weight vocoder that improves upon WaveRNN [2] in terms of sound quality and inference time. Its design is based on the principles of source-filter model [3] of speech production. The key idea in LPCNet is to separate the burden of vocal tract response prediction using a well-understood, low-cost, linear prediction filter and utilize a smaller, WaveRNN-style neural net's capacity for the prediction of source excitation in order to reconstruct the speech. It is one of the most compute efficient neural vocoders in recent times and has been demonstrated to give excellent performance when used in text-to-speech systems [4, 5] as well as in speech codecs [6]. Due to its small size and low computational complexity, it is well suited as the vocoder component for commercial on-device TTS solutions for mobile and IoT devices. An auto-regressive vocoder in a TTS system is the biggest computational bottleneck as it infers speech samples one at a time conditioned on the previously generated samples. For a wide-band speech signal generation at 16 KHz or 24 KHz, even an efficient LPCNet implementation accounts for 80 to 90% of the total computational cost, the rest accounted for by the NLP and acoustic modeling components that generate the input for the vocoder.

In this paper we address the problem of reducing the computational complexity of LPCNet without sacrificing the synthesized audio quality. Our contributions include two methods: 1) Sample bunching where the architecture of the LPCNet has been modified to generate more than one sample per inference

and 2) Bit bunching where the final, softmax layer is segregated into two bunches to reduce the layer size and computations.

We present a brief overview of LPCNet in Section 2, followed by in-depth description of the sample and bit bunching techniques in Sections 3 and 4 respectively. Evaluation results highlighting the efficacy of the proposed methods are presented in Section 5.

2. LPCNet overview

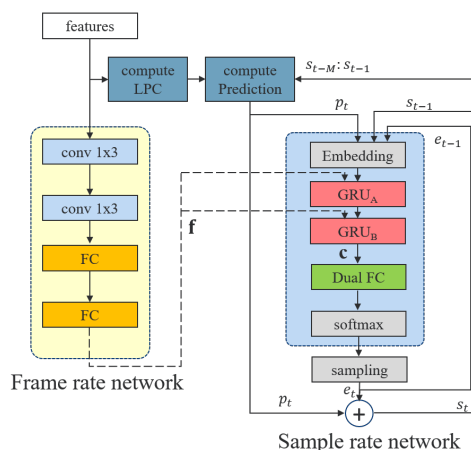


Figure 1: LPCNet overview

LPCNet, as depicted in Figure 1, keeps the computational complexity low by using an all-pole LPC filter (M coefficients) [7] for modeling the vocal tract response and a small neural network for predicting the excitation signals. It comprises a frame rate network (FRN) that runs once per input frame, and a sample rate network (SRN) that runs N (frame size) times per frame generating one audio sample per inference. As a result, most of the computational burden resides in the SRN.

The SRN comprises two gated recurrent units (GRU) and one dual fully-connected (dual FC) layer leading to a softmax layer that models the probability distribution of the excitation signal. The excitation signal e_t is sampled from this distribution and combined with the prediction p_t from the LPC filter to generate the audio sample s_t . The excitation and speech sample from the previous time step along with the prediction for the current time step are fed as input to the SRN via embedded representations.

[†] equal contribution

Weight sparsification [2] applied to the recurrent weight matrix in GRU_A can reduce the complexity without quality degradation. Notably, in GRU_A, while the complexity of the input weight matrix U is larger than that of the recurrent weight matrix W , matrix-vector multiplications of U with the embedding vectors from p , s and e are converted into addition operations using pre-computed lookup tables obtained from multiplication of the embedding tables and the corresponding weights. With these optimizations, LPCNet shows high performance with much lower complexity as compared to other autoregressive (AR) neural vocoders [2, 8].

We note that within the SRN, the two GRU units and the dual FC layer account for about 85% and 15% of the computations respectively. The two proposed techniques, sampling bunching and bit bunching, target reducing computations in these blocks respectively.

3. Sample bunching

The key idea with sample bunching is to get the SRN to generate more than one sample (called a bunch herewith) per inference thereby allowing it to run fewer times resulting in a reduction of computational cost.

The concept of generating multiple samples has been tried in the context of non-autoregressive vocoders [9, 10, 11, 12, 13] via parallel inference on GPUs. Such models do not scale well for CPU processing. In AR vocoders, parallel sample generation is challenging due to the dependence of current inference on past output. SampleRNN [14], an efficient AR model, uses different clock rates for higher and lower layers in the network to increase throughput. In recently proposed Gaussian LPCNet [15], two samples are generated per inference by assuming independence for these samples. In order to maintain the quality of output, the capacity of RNN was increased in that work.

In our approach described below, we maintain the autoregressive nature of LPCNet in conditioning the current output on past outputs and thereby enabling the maintenance of audio quality beyond bunch size of 2. Our approach allows multi-sample generation on any hardware including low-end CPUs.

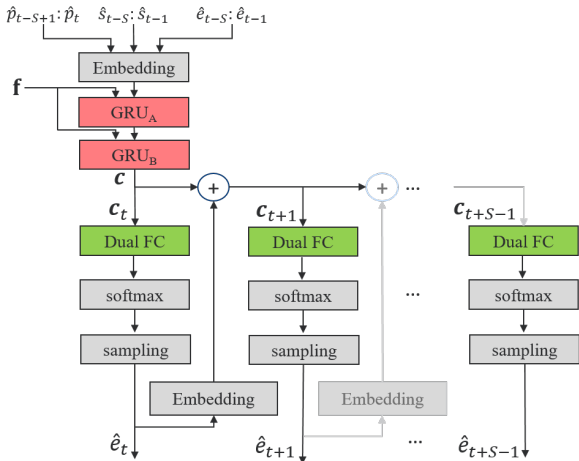


Figure 2: Sample bunching

Our proposition is that the GRUs in the SRN have sufficient model capacity to generate a bunch of samples. As seen in Figure 2, the SRN shares the GRU layers for all the samples in the bunch and has an individual dual FC layer for each exci-

tion prediction in the bunch. The input to the dual FC layer for the first excitation is conditioned only on the output from GRU_B, $\hat{e}_t \sim p(e_t|c)$; while for the rest, it is also conditioned on the previous excitations within the bunch via embedding feeds, $\hat{e}_{t+k} \sim p(e_{t+k}|c, \hat{e}_t, \dots, \hat{e}_{t+k-1})$.

While the number of iterations for the SRN is reduced by the bunch size S , the inputs to GRU_A increase linearly with S resulting in larger input matrix U in GRU_A. However, using the lookup table implementation as described in Section 2, this increase in cost is marginal and the overall gain in computations with sample bunching is proportional to $1/S$.

4. Bit bunching

LPCNet [1] uses a Dual FC layer with a softmax activation layer of size 256 for computing the probability $p(e_t)$. Each softmax output node corresponds to a quantized level of the signal in 8-bits μ -law representation. Inspired by the dual softmax layer in WaveRNN [2], we introduce the idea of bit bunching for this layer to optimize the inference speed further in conjunction with sample bunching.

Modifying the Dual FC layer to split 8 bits into two separate groups, higher bit and lower bit bunches, results in two smaller output layers and thereby gains in computational complexity. Our bit bunching approach also uses a new scaled μ -law quantization to allow higher number of bits for better quality signal.

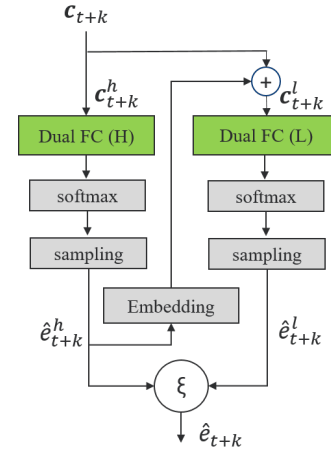


Figure 3: Bit bunching

The idea of bit bunching is depicted in Figure 3. With bit bunching, the input information into the Dual FC layer is not changed from the original LPCNet, i.e., the information for predicting probability $p(e_t)$ remains the same. The higher bit bunch and lower bit bunch outputs map to a coarse prediction and fine correction of the excitation signal respectively. To improve the prediction efficacy of the lower bits, the higher bit predictions are additionally fed in as conditioning input via embedding layer. In our experiments, this additional conditioning was found to be useful in improving the cross-entropy loss for the lower-bit bunch and also assisted in the choice of number of bits to assign for higher-bit bunch. The best cross entropy error was achieved with a split of $\mathbf{B} = (B_h, B_l) = (7, 4)$ where B_h and B_l denote the number of bits in higher bit and lower bit bunches respectively, and this split has been used in all further experimentation. From the higher and lower bit bunch predictions \hat{e}_t^h and \hat{e}_t^l , the excitation signal is calculated

Algorithm 1: Audio sample generation with sample and bit bunching

Input: Samples, excitations and predictions from previous run of SRN $\hat{s}_{t-S} : \hat{s}_{t-1}$, $\hat{e}_{t-S} : \hat{e}_{t-1}$ and $\hat{p}_{t-S+1} : \hat{p}_t$ respectively

Output: $\hat{s}_t : \hat{s}_{t+S-1}$, $\hat{e}_t : \hat{e}_{t+S-1}$ and $\hat{p}_{t+1} : \hat{p}_{t+S}$

- 1 $\mathbf{c} \leftarrow \text{GRU}_B(f, \text{GRU}_A(f, \text{Embedding}(\text{Input})))$
- 2 $\mathbf{c}_t^h = \mathbf{c}$
- 3 **for** $i \leftarrow 0$ **to** $S - 1$ **do**
- 4 $\hat{e}_{t+i}^h \leftarrow \text{sampling from } p(e_{t+i}^h | \mathbf{c}_{t+i}^h)$
- 5 $\mathbf{c}_{t+i}^l \leftarrow \mathbf{c}_{t+i}^h + \text{Embedding}(\hat{e}_{t+i}^h)$
- 6 $\hat{e}_{t+i}^l \leftarrow \text{sampling from } p(e_{t+i}^l | \mathbf{c}_{t+i}^l)$
- 7 $\hat{e}_{t+i} \leftarrow 2^{B_i} \hat{e}_{t+i}^h + \hat{e}_{t+i}^l$
- 8 $\mathbf{c}_{t+i+1}^h \leftarrow \mathbf{c}_{t+i}^h + \text{Embedding}(\hat{e}_{t+i})$
- 9 **end**
- 10 **for** $i \leftarrow 0$ **to** $S - 1$ **do**
- 11 $\hat{s}_{t+i} \leftarrow \hat{e}_{t+i} + \hat{p}_{t+i}$
- 12 $\hat{p}_{t+i+1} \leftarrow \text{compute prediction}(\hat{s}_{t+i-M-1} : \hat{s}_{t+i})$
- 13 **end**

as $\hat{e}_t = \xi(\hat{e}_t^h, \hat{e}_t^l) = 2^{B_i} \hat{e}_t^h + \hat{e}_t^l$. From the generated excitations, the predictions and audio samples are computed with sample and bit bunching according to Algorithm 1.

As in the LPCNet paper, we have employed μ -law quantization algorithm [16] to represent 16-bits PCM values x , where $-32768 \leq x \leq 32767$. It is an efficient method which can represent the waveform with B -bits as shown in (1).

$$y = Q_B(x) = \text{sign}(x) \cdot V_{m2} \cdot \frac{\ln(1 + s_1|x|)}{\ln(V_m)}$$

$$x = Q_B^{-1}(y) = \text{sign}(u) \cdot s_2 \cdot \left(\exp \frac{\ln(V_m)|u|}{V_{m2}} - 1 \right) \quad (1)$$

$$\text{where } V_m = 2^B, V_{m2} = 2^{B-1},$$

$$u = y - V_{m2}, s_1 = \frac{V_m - 1}{2^{15}}, s_2 = \frac{2^{15}}{V_m - 1}$$

$$V_m = w_s 2^B \quad (2)$$

Typically, most systems use it with $B = 8$, but with bit bunching we consider cases with larger B . One downside with $B > 9$ is that the quantization step size is smaller than 1 for values of x close to zero. For instance for $B = 11$, $Q_{11}(0) = 1024$ and $Q_{11}(1) = 1032$. This leads to under-utilization of quantization levels and also discrepancies due to many-to-one mapping from sampled quantization value to the PCM conversion. To address this issue, we added a factor w_s to control the slope of the mapping function as represented in (2), and it was chosen to keep the quantization step always greater than 1. For instance, for $B = 11$, w_s is set to 0.08 as shown in Figure 4.

5. Evaluation

5.1. Experimental environment

For high-fidelity TTS with low complexity, we made the following modifications to the original LPCNet configuration: (1) For 24KHz sampling rate, 10 ms frame size ($N : 240$), 20 Bark cepstral coefficients with 240 shift size and 480 window size are used. (2) We used the RAPT algorithm [17] for pitch tracking. It showed the better performance in the LPCNet and in generation of more natural prosody in the acoustic model. (3)

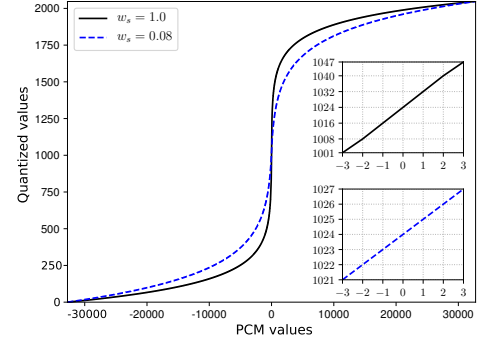


Figure 4: Mapping function of the modified μ -law quantization with $B = 11$

We increased the sparse ratio of the recurrent weight matrix in GRU_A from (0.95, 0.95, 0.8) to (0.99, 0.99, 0.9). This way, the overall complexity was reduced by about 25% with a similar performance, and we employed this modified LPCNet as a baseline system for all comparisons. The learning rate and decay rate were determined by random search method [18] and the other hyper-parameters were set identical to the original LPCNet, i.e., GRU_A with 384 units and GRU_B with 16 units. A total of 8 systems were evaluated with the sample bunching sizes $S \in \{1, 2, 3, 4\}$ and the bit bunching configurations $\mathbf{B} = (B_h, B_l) \in \{(8, 0), (7, 4)\}$. Note that the system with $S = 1$ and $\mathbf{B} = (8, 0)$ corresponds to the baseline system. When $\mathbf{B} = (7, 4)$, the modified μ -law quantization was applied with $w_s = 0.08$.

A phoneme-based DCTS [19] was employed as an acoustic model to evaluate TTS performance. The systems were trained using two datasets, a professional English male speaker (17-hours with 10,000 utterances) and a professional English female speaker (15-hours with 7,612 utterances). 110 utterances were used as a test set, and one percent of the rest were held out as validation set for training.

5.2. Complexity

We measured RTF (Real Time Factor) and CR (Complexity Ratio) on two devices: 1) AWS c5.4xlarge instance - representative of cloud deployment (Intel(R) Xeon(R) Platinum 8124M CPU @ 3.00GHz) and 2) Samsung Galaxy S10+ (Exynos 9820) for on-device deployment. The implementation was optimized using SIMD (Single Instruction Multiple Data) with single thread for each architecture.

The RTFs and CRs summarized in Table 1 highlight the improvements in computational complexity with sample and bit bunching over the baseline : $S = 1$ and $\mathbf{B} = (8, 0)$. The sample bunching algorithm with $S = 4$ achieves a reduction of 39.7% and 43.4% over the baseline on Intel Xeon and Exynos respectively. Bit bunching helps in reducing the RTF further by about 7% and 10% absolute on these devices. With $S = 4$ and $\mathbf{B} = (7, 4)$, we get an overall improvement of 54.2%, i.e., it runs 2.19 times faster than the baseline, on the mobile device.

Since we need one Dual FC layer computation per excitation calculation even with sample bunching, it implies a constant reduction in computations with bit bunching for all values of S . At higher sample bunching values, this accounts for a larger proportion of savings, for e.g, relative gain with bit bunching on Intel Xeon at baseline is 6.6% while at $S = 4$,

Table 1: Real time factor and complexity ratio on the two CPU architectures

S	B	Intel Xeon		Exynos 9820	
		RTF	CR	RTF	CR
1	(8, 0)	0.136	100.0%	0.243	100.0%
	(7, 4)	0.127	93.4%	0.214	88.3%
2	(8, 0)	0.098	72.1%	0.174	71.7%
	(7, 4)	0.089	65.4%	0.149	61.2%
3	(8, 0)	0.087	64.0%	0.147	60.6%
	(7, 4)	0.078	57.4%	0.124	51.2%
4	(8, 0)	0.082	60.3%	0.137	56.6%
	(7, 4)	0.072	52.9%	0.111	45.8%

it results in 12.2% gain.

To verify the efficacy of the bunching approach in complexity reduction, we compare the validation loss of the proposed system and the baseline system with smaller GRU_A units M_A in Figure 5. Sample bunching shows lower complexity for the same validation loss. For example, ($S = 1, M_A = 288$) and ($S = 3, M_A = 384$) with similar loss, 3.3854 and 3.3840, work 25.9% and 56.3% faster than the baseline system ($S = 1, M_A = 384$) respectively. It suggests that the proposed approach is a more efficient method than reducing the units in the RNN layers.

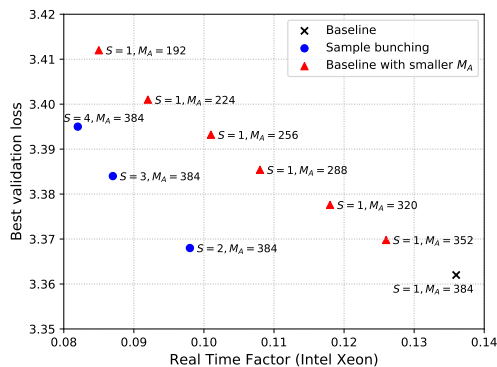


Figure 5: Validation loss versus RTF (baseline systems with varying M_A and sample bunching systems)

5.3. Quality

For quality evaluation, we conducted a MOS (Mean Opinion Score) test and a DMOS (differential MOS) test using the Amazon Mechanical Turk platform on 11 systems including one original speech as a high anchor, and two degraded ones generated by Griffin-Lim [20] and 5 bits μ -law quantization as low anchors, with 100 people and 100 unseen test utterances. Each listener rated 110 utterances, 10 utterances each for the 11 systems.

For the evaluation, we choose Degradation Category Rating (DCR) manners for *Anal-Synth* and Absolute Category Rating (ACR) manners for *LPCNet+DCTTS* test defined in ITU-T P.800 [21], for precise analysis of the vocoder performance. In the test, the subjects are requested to rate the amount of degradation compared with given reference in case of *Anal-Synth* and

to rate absolute quality in the case of *LPCNet+DCTTS*. There are 10 categories in our test [0.5, 5] with a step size of 0.5 and a higher score corresponds to better quality. To remove non-discriminative ratings, we applied a screening rule where results were discarded if the original item scored lower than 4.0. The results of *Anal-Synth* using the extracted features from original speech and *LPCNet+DCTTS* using the predicted features from the DCTTS are summarized in Tables 2 and 3 respectively¹.

Table 2: DMOS scores with 95% confidence intervals

S	B	<i>Anal-Synth</i> (DMOS)	
		male	female
Original		4.62 ± 0.03	4.66 ± 0.03
1	(8, 0)	4.35 ± 0.06	4.40 ± 0.06
	(7, 4)	4.32 ± 0.06	4.39 ± 0.06
2	(8, 0)	4.30 ± 0.06	4.30 ± 0.06
	(7, 4)	4.25 ± 0.06	4.29 ± 0.06
3	(8, 0)	4.26 ± 0.06	4.25 ± 0.07
	(7, 4)	4.18 ± 0.06	4.20 ± 0.06
4	(8, 0)	4.22 ± 0.06	4.09 ± 0.07
	(7, 4)	4.16 ± 0.06	4.00 ± 0.07

Table 3: MOS scores with 95% confidence intervals

S	B	<i>LPCNet+DCTTS</i> (MOS)	
		male	female
Original		4.41 ± 0.03	4.44 ± 0.03
1	(8, 0)	4.08 ± 0.05	4.09 ± 0.05
	(7, 4)	4.02 ± 0.05	4.10 ± 0.05
2	(8, 0)	4.05 ± 0.05	4.06 ± 0.05
	(7, 4)	4.04 ± 0.05	4.06 ± 0.05
3	(8, 0)	4.01 ± 0.06	4.09 ± 0.05
	(7, 4)	4.00 ± 0.05	4.06 ± 0.05
4	(8, 0)	4.01 ± 0.06	4.06 ± 0.05
	(7, 4)	3.99 ± 0.05	4.03 ± 0.06

In the *Anal-Synth* case, it is confirmed that lower the complexity, lower is the DMOS. Especially, DMOS decreases rapidly when $S = 4$, which can be attributed to the insufficient capacity of GRU_B. On the other hand, in the TTS case, MOS degradation is insignificant, and the proposed method with $S = 4$ and $B = (7, 4)$, shows a less than 0.1 MOS drop compared to the baseline. In fact, DCR test is more sensitive than ACR and the difference according to the test methodology shows up in the DMOS and MOS score of original utterance.

6. Conclusion

In this work, we have proposed sample and bit bunching techniques to reduce the computational complexity of the LPCNet vocoder. We have demonstrated that bunched-LPCNet can provide more than 2X speed-up over baseline with negligible loss in quality. It is thus a strong proposition for use within a commercial on-device TTS system for lower-spec mobile, IoT and embedded devices.

¹Audio samples available at <https://bunchedlpcnet.github.io/>

7. References

- [1] J.-M. Valin and J. Skoglund, "LPCNet: Improving Neural Speech Synthesis through Linear Prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5891–5895.
- [2] 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 *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80, 2018, pp. 2410–2419.
- [3] T. Chiba and M. Kajiyama, *The Vowel, its Nature and Structure*. Tokyo-Kaiseikan Publishing Company, Ltd, Tokyo, 1941.
- [4] Z. Kongs, S. Shechtman, A. Sorin, C. Rabinovitz, and R. Hoory, "High quality, lightweight and adaptable TTS using LPCNet," in *INTERSPEECH*, 2019.
- [5] M.-J. Hwang, E. Song, R. Yamamoto, F. Soong, and H.-G. Kang, "Improving LPCNET-Based Text-to-Speech with Linear Prediction-Structured Mixture Density Network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7219–7223.
- [6] J.-M. Valin and J. Skoglund, "A Real-Time Wideband Neural Vocoder at 1.6 kb/s Using LPCNet," in *INTERSPEECH*, 09 2019, pp. 3406–3410.
- [7] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
- [8] A. v. d. 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.
- [9] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. 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 *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018, pp. 3918–3926.
- [10] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.
- [11] K. Peng, W. Ping, Z. Song, and K. Zhao, "Parallel neural text-to-speech," *arXiv preprint arXiv:1905.08459*, 2019.
- [12] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.
- [13] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech," in *International Conference on Learning Representations (ICLR)*, 2018.
- [14] M. Soroush, K. Kundan, G. Ishaan, K. Rithesh, J. Shubham, S. Jose, C. Aaron, and B. Yoshua, "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," in *ICLR*, 2017.
- [15] V. Popov, M. Kudinov, and T. Sadekova, "Gaussian LPCNet for Multisample Speech Synthesis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6204–6208.
- [16] *ITU-T Recommendation G. 711.*, Pulse Code Modulation (PCM) of voice frequencies, 1988.
- [17] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp. 497–518, 1995.
- [18] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of machine learning research*, vol. 13, no. Feb, pp. 281–305, 2012.
- [19] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4784–4788.
- [20] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.
- [21] *ITU-T Recommendation P. 800*, Methods for subjective determination of transmission quality, 1996.