Triple M: A Practical Text-to-speech Synthesis System With Multi-guidance Attention And Multi-band Multi-time LPCNet

Shilun Lin, Fenglong Xie, Li Meng, Xinhui Li, Li Lu

Tencent, Beijing, China

{cirolin, fenglongxie, leemeng, hiccupli, adolphlu}@tencent.com

Abstract

In this work, a robust and efficient text-to-speech (TTS) synthesis system named Triple M is proposed for large-scale online application. The key components of Triple M are: 1) A sequence-to-sequence model adopts a novel multi-guidance attention to transfer complementary advantages from guiding attention mechanisms to the basic attention mechanism without in-domain performance loss and online service modification. Compared with single attention mechanism, multi-guidance attention not only brings better naturalness to long sentence synthesis, but also reduces the word error rate by 26.8%. 2) A new efficient multi-band multi-time vocoder framework, which reduces the computational complexity from 2.8 to 1.0 GFLOP and speeds up LPCNet by 2.75x on a single CPU.

Index Terms: Speech synthesis, sequence-to-sequence model, attention, transfer learning, vocoder, LPCNet

1. Introduction

In the past few years, speech synthesis has attracted a lot of attention due to advances in deep learning. Sequence-to-sequence neural network [1] with attention mechanism is one of the most popular text-to-feature models [2, 3]. Attention mechanism is applied to align the input and output sequences. Therefore, the training is no longer fragmented. Content-based attention [4] applied in Tacotron [2] does not exploit the monotonicity and locality between text and acoustic features. Tacotron2 [3] attempts to enhance the robustness of attention by introducing location-related information. However, there are still some problems in this hybrid attention mechanism such as lack of monotonic restriction and weak robustness of long sentence synthesis. In order to alleviate these problems, alternative attention mechanisms such as forward attention [5], GMM-based attention [6], stepwise monotonic attention [7] and dynamic convolution attention [8] are proposed. These attention mechanisms have their own advantages and implementations. It would be nice to transfer those advantages to the basic attention mechanism and keep its original implementation. Knowledge transfer or transfer learning [9] between attention mechanisms would be desirable. In this work, we propose a novel multi-guidance attention mechanism which enables the basic location-sensitive attention mechanism to learn different strengths from multiple guidance sources. Forward attention and GMM-based attention are selected as the guiding attention mechanisms after careful analysis. Thanks to the multi-guidance attention mechanism, the convergence speed and the robustness of long sentence synthesis are significantly improved without affecting perceptual quality. More importantly, experiments show that the advantages of guidance attention mechanisms can be transferred to the basic attention through a reasonable combination.

For a large-scale online custom speech synthesis system, it is very important to reduce the cost of the neural vocoder. The flow based neural vocoders, such as Parallel WaveNet [10] and Clarinet [11], reduce the computational overhead of WaveNet [12] by learning from the teacher WaveNet and realize real-time synthesis on the GPU. Recently, autoregressive models with simple structures like WaveRNN [13] and LPCNet [14] have been proposed for real-time synthesis on CPU by introducing sparse GRU. In this work, we propose a multi-band multi-time LPCNet that can simultaneously calculate the excitation of different subbands at adjacent moments in each forward process. The redundant computation is reduced by making full use of the correlation in frequency domain and time domain.

2. Proposed Method

In this section, we describe main components of triple M as shown in Figure 1. First, a text-to-feature module with novel multi-guidance attention will be introduced. The basic location-sensitive attention mechanism learns different strengths, such as fast convergence, stable feature generation and robust long sentence synthesis, from multiple guidance sources during training. In this way, new attention properties which make the system more robust can be integrated without modifying the inference framework of the online system. Next, a novel multi-band multi-time LPCNet which can significantly improve the efficiency of the system will be described.

2.1. Multi-guidance attention

The elaboration of multi-guidance attention will be divided into three parts. Firstly, a basic text-to-feature module which learns from the guidance attention and serves as the final inference module will be described. Then, we will explain why forward and GMM-based attention are chosen as guidance mechanisms. Finally, the details of guidance training will be given.

2.1.1. Basic setup

The basic text-to-feature module used in our work is based on Tacotron [2]. The improved hybrid location-sensitive attention proposed in Tacotron2 [3] is applied as the basic attention. CBHG (convolutional bank, highway network and bidirectional gated recurrent unit) encoder is used to transform Chinese Pinyin sequences with tone and prosody information $\{x_i\}_{i=1}^L$ into hidden text representations $\{h_i\}_{i=1}^L$ that are more suitable for attention mechanism (Eq.1). An attention RNN uses the last state, context vector and decoding result of the previous time step as input, and outputs the current state s_t for computing the attention score (Eq.2). The hybrid location-sensitive attention module takes the current state, hidden representations and location-related information as input to get the attention score a_t. And then the context vector c_t at the current moment is calculated (Eq.3). Finally, last decoder RNN state, current attention RNN state and the context vector are fed to the decoder
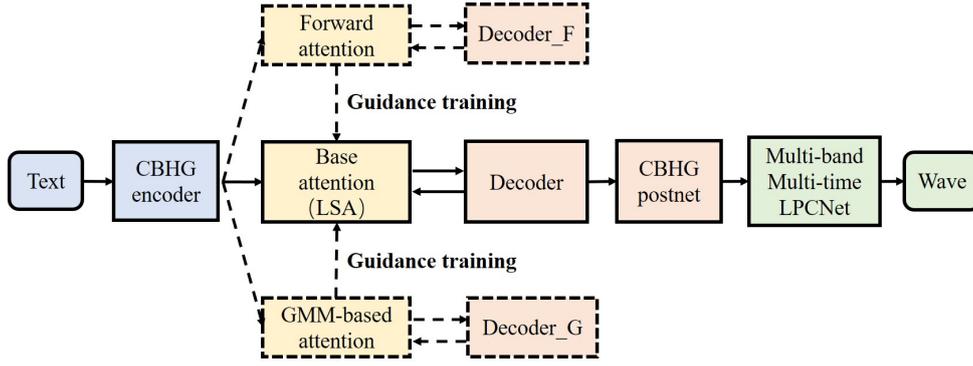
Figure 1: *System architecture of our proposed Triple M.*

RNN. Then the current state of the decoder d_t is obtained and passed to an affine function to obtain the final decoding result o_t (Eq.4).

$$\{h_i\}_{i=1}^{L} = CBHGEncoder(\{x_i\}_{i=1}^{L}) \tag{1}$$

$$s_t = AttentionRNN(s_{t-1}, c_{t-1}, o_{t-1}) \tag{2}$$

$$a_t = LSA(s_t, h_i, l_t) \qquad c_t = \sum_i^L a_{t,i} h_i \tag{3}$$

$$d_t = DecoderRNN(d_{t-1}, c_t, s_t) \quad o_t = Affine(d_t) \tag{4}$$

2.1.2. Guidance sources selection

The multi-guidance attention mechanism is introduced into above basic structure and the supervision attention mechanisms are carefully selected. Two complementary attention mechanisms, forward attention [5] and GMM-based attention [6] are selected as guidance sources after comparing different attention mechanisms. As a kind of monotonic attention mechanisms, forward attention only considers the alignment paths that satisfy the monotonic condition at each decoding time step to ensure the monotonicity of the final alignment path. This method has been verified to accelerate the convergence and improve the stability of feature generation. However, it is observed that the training process often fails to learn valid alignments. Using a fixed diagonal mask to regulate the attention alignment matrix in training is helpful to solve this problem [15]. The relatively stable basic location-sensitive attention here can play the role of providing diagonal mask. It helps to avoid the collapse of forward attention during training. In turn, the well trained forward attention can give the basic attention an explicit monotonic restriction which is helpful to its guidance learning. For large-scale online applications, we need to improve the robustness of long-form speech synthesis [8] in addition to ensuring the stability of in-domain sentence synthesis. Different from forward attention, GMM-based attention is a purely location-related attention mechanism. It can bring benefits to the basic location-sensitive attention that are different from those of the monotonic attention mechanism. Using GMM-based attention as a guiding mechanism can give the basic location-sensitive attention the ability to synthesize long sentences without modifying the forward computation. This effectively prevents the system from crashing due to long-form input.

2.1.3. Guidance training

The guidance attention mechanisms affect the learning of basic attention mechanism through the loss function during training.

The training loss function consists of three parts (Eq.5). The first part contains l_1 distances between all decoder outputs and the real acoustic feature r. o, o_f, o_g represent the output of basic decoder, forward attention decoder and GMM-based attention decoder respectively. Then there is the l_1 distance between the output of CBHG postnet p and the real acoustic feature r. The last part L_ga (Eq.6) includes l_1 distances between the basic alignment score a and all guidance alignment scores. a_f and a_g represent alignment score of forward attention and GMM-based attention. λ is used to control the intensity of guidance learning.

$$L = \ell_1(o, r) + \ell_1(o_f, r) + \ell_1(o_g, r) + \ell_1(p, r) + L\_ga \tag{5}$$

$$L\_ga = \lambda(\ell_1(a, a_f) + \ell_1(a, a_g)) \tag{6}$$

All guidance-related modules are trained together with the basic text-to-feature module. When the guidance attention mechanisms become reliable, the basic attention mechanism begins to learn from them. Only the basic text-to-feature module is retained in inference.

2.2. Multi-band multi-time LPCNet

2.2.1. LPCNet

As a relatively lightweight neural vocoder, LPCNet [14] models the vocal tract response through a low-cost linear prediction filter and adopts a smaller network to obtain the excitation. However, the computational overhead of LPCNet still has room for improvement. The original LPCNet is composed of a frame rare network (FRN) and a sample rate network (SRN). The former extracts high-level representations from conditional acoustic features. The latter including two GRU layers and a dual FC layer needs to loop T times per frame to obtain T corresponding samples. Therefore, the SRN occupies a major part of the overall calculation.

2.2.2. Multi-band multi-time processing

As analyzed above, the computational complexity of LPCNet is mainly concentrated on SRN. Decreasing the number of SRN forward steps can effectively reduce the computational overhead. Therefore, a multi-band multi-time processing framework is proposed in this work to further explore the acceleration potential of LPCNet by making full use of the correlation in both frequency domain and time domain. The pipeline is shown in Figure 2.

Multi-band processing is initially used for parallel acceleration of vocoder [16, 17]. The speed can be effectively in-

creased by deploying multiple models to calculate different sub-band signals in parallel, but the computational overhead cannot be truly reduced. Durian proposed in [18] introduces multi-band processing into a single WaveRNN model, which reduces the total computational complexity from 9.8 to 3.6 GFLOPS. We apply the multi-band strategy to LPCNet for the purpose of further exploiting the latent capacity of SRN. A kind of Cosine-Modulated Filter Bank (CMFB) named Pseudo Quadratue Mirror Filter Bank (Pseudo-QMF) [19] is applied to multi-band processing. After processing, the original signal is divided into N subbands. Down-sampling each subband N times will not cause the loss of original information. Intuitively, the number of forward steps will be reduced by N times through using SRN to predict N down-sampled subband signals simultaneously. Since Pseudo-QMF is a low-cost filter bank, the cost of reconstructing the original signal from the subband signals is much less than that saved by reducing the number of SRN forward steps. Multi-band strategy improves the efficiency of LPCNet from the frequency domain. The multi-time strategy takes two adjacent sampling points in the subband signal into consideration. SRN predicts adjacent points in N subbands at the same time, which can reduce the number of SRN forward steps by 2N times. Bunched LPCNet [20] predicts a bunch of original samples in an autoregressive way. Within a bunch, the first excitation is only conditioned on the output from GRU-B. While for the rest excitation, it depends on all the previous excitation in the same bunch. Different from Bunched LPCNet, our multi-band multi-time LPCNet predicts the adjacent samples of each subband signal at the same time. Due to the small time span, we remove the autoregressive module for simplicity.
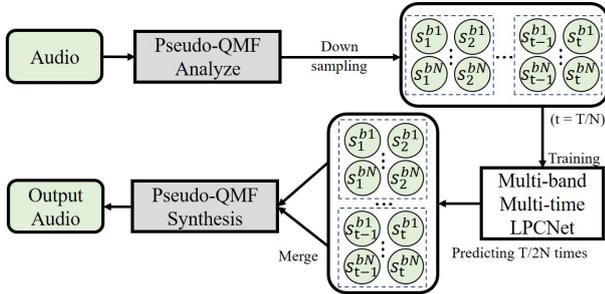

Figure 2: *The flowchart of multi-band multi-time processing.*

2.2.3. Audio generation framework

Except for the dual FC layer, the rest of the SRN layers are shared in each forward process as shown in Figure 3. Take four subbands as an example: The excitation ($e_{t-1}^{b1:b4}$ and $e_{t-2}^{b1:b4}$), audio samples ($s_{t-1}^{b1:b4}$ and $s_{t-2}^{b1:b4}$) from the last adjacent time and obtained predictions (at the last time $p_{t-1}^{b1:b4}$ and the current time $p_t^{b1:b4}$) are used as input to the first GRU layer (GRU-A). The output of the second GRU layer (GRU-B) is sent to 8 independent dual FC layers to predict the excitation of subbands at adjacent time ($e_t^{b1:b4}$ and $e_{t+1}^{b1:b4}$). After that, the audio samples at the current adjacent time ($s_t^{b1:b4}$ and $s_{t+1}^{b1:b4}$) can be acquired recursively (Eq.7,8,9). Finally, the LPC queues of all subbands are updated to prepare for next round. In summary, the input matrix of GRU-A layer will become larger while the table lookup operation makes this overhead negligible. Although 7 more dual FC layers are introduced, the multi-band multi-time strategy reduces the number of required cycles by 8 times which reduces

the computational complexity from 2.8 to 1.0 GFLOPS.

$$s_t^{b1:b4} = e_t^{b1:b4} + p_t^{b1:b4} \tag{7}$$

$$p_{t+1}^{b1:b4} = LPCPrediction(s_{t-16+1}^{b1:b4} : s_{t-1+1}^{b1:b4}) \tag{8}$$

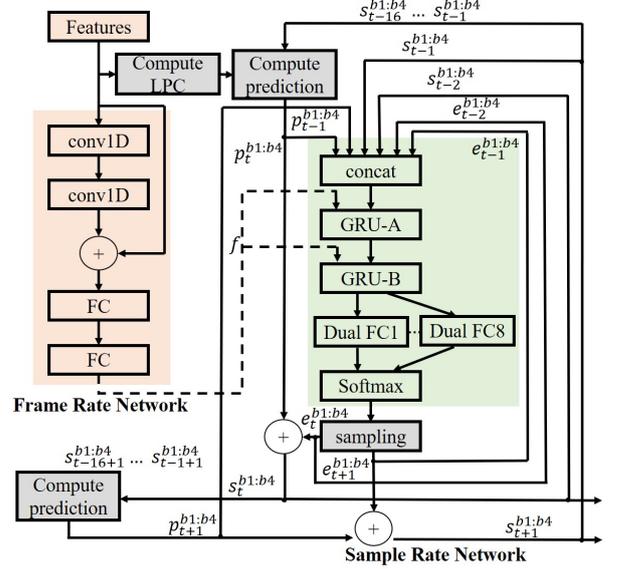$$s_{t+1}^{b1:b4} = e_{t+1}^{b1:b4} + p_{t+1}^{b1:b4} \tag{9}$$

Figure 3: *The architecture of multi-band multi-time LPCNet.*

3. Experiments

3.1. Experimental setup

A Mandarin corpus recorded by a Chinese male speaker was used in our experiments. All recordings were sampled at 16kHz with 16-bit quantization. About 16 hours recordings with an average length of 90 characters were used for training. 100 regular sentences and 50 paragraphs selected from WeChat official account with an average length of 1000 characters were tested. Consistent with original LPCNet, 18 Bark cepstral coefficients and 2 pitch parameters were extracted as conditions of the vocoder. At the same time, they were also used as the prediction targets of the text-to-feature module.

The text-to-feature module was a sequence-to-sequence model with a CBHG encoder and a CBHG postnet. The main component of decoder was a two-layer 512-dimensional unidirectional LSTM. All attention modules were composed of a unidirectional GRU layer and corresponding attention mechanism. λ was used to control the intensity of guidance learning, which was set to 10 in our experiment. All modules were trained together at the beginning. When the guidance attention modules aligned the text and acoustic features stably, the basic attention module began to learn from them.

In the original LPCNet, SRN consisted of a 90% sparse 384-dimensional GRU layer, a normal 16-dimensional GRU layer and a 256-dimensional dual FC layer. As to multi-band multi-time LPCNet, dimensions of all GRU layers kept unchanged. Different from [14], condition feature was fed to GRU-B in addition to being input to GRU-A. In order to reach a reasonable compromise between the quality and the computational cost, the hyper-parameters of subband and time span were set to 4 and 2 respectively.

Audio samples related to this work were available on the accompanying web page[1].

3.2. Evaluations

The experiment found that the basic attention module can align the text and acoustic features stably at the second epoch by learning from forward attention and GMM-based attention. It took much less epochs than 10 epochs required without guidance. This prevented the model falling into bad alignment in early stage of training and provided a stable basis for the subsequent training. The well-learned basic attention provided an approximate diagonal mask which gave a constraint to the forward attention. Thus it was not easy to collapse during training.

Table 1: *Long sentence failure rate and word error rate.*

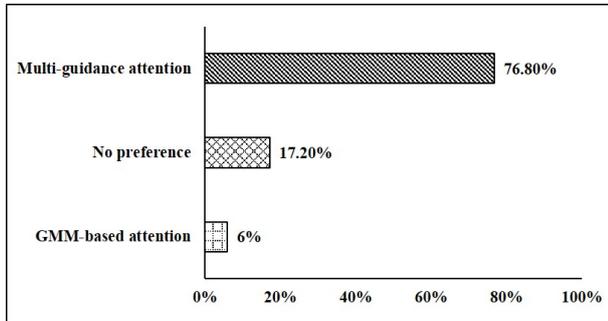| Model	Failure rate	Word error rate
Baseline	>60%	NA
Single-guidance attention	>40%	NA
GMM-based attention	10%	4.1%
Multi-guidance attention	**2%**	**3.0%**

Figure 4: *Multi-guidance attention vs. GMM-based attention, the p-value of t-test between the two systems p<0.01.*

In the long sentence synthesis experiment, 50 paragraphs selected from WeChat official account covering the fields of politics, sports, entertainment, literature, cooking and so on were tested. The experimental results are recorded in Table 1. The failure rates were more than 40% when forward attention and GMM-based attention were severally used as guidance mechanism. Compared with the baseline model, there was no satisfactory improvement in robustness of long sentences synthesis. The failure was mainly identified by whether the synthesized audio ended early, repeated the same clip or contained meaningless clip which seriously affected the understanding of the content. Applying two attention mechanisms simultaneously to guide the learning of the basic attention reduced the failure rate to 2%. These results showed that only using forward attention as a supervisory signal cannot provide long sentence synthesis capabilities. But without the constraint of forward attention, basic attention cannot learn long sentence synthesis ability from GMM-based attention well. Only using GMM-based attention can achieve a 10% failure rate with 3.3% in-domain MOS degradation. In the A/B preference test, 10 listeners were asked to listen to the paired sentences synthesized by two different systems in random order and then gave their preference. The result in Figure 4 shows that the audio synthesized by the system with multi-guidance attention was much better.

The word error rate (WER) measured by an ASR system described in [21] was 4.1% when only GMM-based attention was used. Using multi attention mechanisms to guide the learning reduced the WER to 3.0%. This might be related to the ability of forward attention to generate features stably.

In order to verify the influence of the multi-guidance attention mechanism on perceptual quality, Mean Opinion Score (MOS) tests were implemented. For all MOS tests in this work, three groups of native Chinese speakers (5 in each group) were invited to listen and score 125 audio each time. 100 test utterances synthesized by the corresponding model were mixed with 25 original recordings, and the listener did not know which category each audio belonged to. Scores ranged from 1 to 5, with 5 representing "completely natural speech". The final MOS was obtained by averaging the scores of the three groups. Table 2 shows the results of the MOS test. Experiments showed that the basic attention can learn the advantages of the guiding attention while maintaining the in-domain naturalness. In this way, new properties can be assigned to the text-to-speech system without modifying the online service. In addition, multi-guidance attention also provided a new way to improve the performance of attention mechanism through transfer learning.

Table 2: *MOS with 95% confidence intervals.*

Model	MOS
Baseline	4.52 ± 0.08
GMM-based attention	4.42 ± 0.08
Multi-guidance	$\mathbf{4.57 \pm 0.05}$
Ground truth	4.65 ± 0.04

According to [14], the complexity of the SRN was around 2.8 GFLOPS. The multi-band multi-time strategy reduced it to around 1.0 GFLOPS (Eq.10).

$$C = \frac{(3dG_A^2 + 3G_B(G_A + G_B) + 2G_BQN_BN_T)2F_S}{N_BN_T} \tag{10}$$

Comparison experiment of MOS and real-time factor (RTF) was implemented to demonstrate that the multi-band multi-time strategy can accelerate synthesis without significantly degrading the perceptual quality. The results are presented in Table 3. It can be found that the multi-band multi-time strategy sped up LPCNet by 2.75x while the MOS has only dropped by 2.6%.

Table 3: *MOS with 95% confidence intervals and RTF.*

Model	MOS	RTF
LPCNet	4.57 ± 0.05	0.303
Multi-band & Multi-time	$\mathbf{4.45 \pm 0.07}$	**0.110**

4. Conclusions

Robustness and efficiency are essential for a large-scale online system. In this paper, we proposed Triple M, a new practical text-to-speech system. It integrates the advantages of various attention mechanisms by a novel multi-guidance strategy, which enables the system to deal with ultra-long sentence robustly while guaranteeing in-domain performance. Compared with single attention, multi-guidance attention reduces the WER of long sentence synthesis by 26.8% without MOS degradation. In addition, exploiting the acceleration potential of the original LPCNet through our multi-band multi-time strategy reduces the computational complexity by approximately 64%.

5. References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014, pp. 3104–3112.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[4] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[5] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Forward attention in sequence-to-sequence acoustic modeling for speech synthesis," in *Proc. ICASSP*, 2018, pp. 4789–4793.

[6] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[7] M. He, Y. Deng, and L. He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural tts," in *Proc. Interspeech*, 2019, pp. 1293–1297.

[8] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," in *Proc. ICASSP*, 2020, pp. 6194–6198.

[9] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[10] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *Proc. ICML*, 2018, pp. 3918–3926.

[11] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR*, 2018.

[12] 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," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[13] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. ICML*, 2018, pp. 2410–2419.

[14] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, 2019, pp. 5891–5895.

[15] D. Lim, W. Jang, O. Gyeonghwan, H. Park, B. Kim, and J. Yoon, "Jdi-t: Jointly trained duration informed transformer for text-to-speech without explicit alignment," in *Proc. Interspeech*, 2020, pp. 4004–4008.

[16] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Improving fft-net vocoder with noise shaping and subband approaches," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 304–311.

[17] 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*, 2018, pp. 5654–5658.

[18] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for speech synthesis," in *Proc. Interspeech*, 2020, pp. 2027–2031.

[19] F. Cruz-Roldán, P. Amo-López, S. Maldonado-Bascón, and S. S. Lawson, "An efficient and simple method for designing prototype filters for cosine-modulated pseudo-qmf banks," *IEEE signal processing letters*, vol. 9, no. 1, pp. 29–31, 2002.

[20] R. Vipperla, S. Park, K. Choo, S. Ishtiaq, K. Min, S. Bhattacharya, A. Mehrotra, A. G. C. Ramos, and N. D. Lane, "Bunched lpcnet: Vocoder for low-cost neural text-to-speech systems," in *Proc. Interspeech*, 2020, pp. 3565–3569.

[21] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang *et al.*, "Transformer-based acoustic modeling for hybrid speech recognition," in *Proc. ICASSP*, 2020, pp. 6874–6878.