INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

TacoLPCNet: Fast and Stable TTS by Conditioning LPCNet on Mel Spectrogram Predictions

Cheng Gong[1], *Longbiao Wang*[1,*], *Ju Zhang*[2,*], *Shaotong Guo*[1], *Yuguang Wang*[2], *Jianwu Dang*[1,3]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Huiyan Technology (Tianjin) Co., Ltd, Tianjin, China
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan

{gongchengcheng, longbiao_wang, shaotong_guo}@tju.edu.cn, juzhang@huiyan-tech.com

Abstract

The combination of the recently proposed LPCNet vocoder and a seq-to-seq acoustic model, i.e., Tacotron, has successfully achieved lightweight speech synthesis systems. However, the quality of synthesized speech is often unstable because the precision of the pitch parameters predicted by acoustic models is insufficient, especially for some tonal languages like Chinese and Japanese. In this paper, we propose an end-to-end speech synthesis system, TacoLPCNet, by conditioning LPCNet on Mel spectrogram predictions. First, we extend LPCNet for the Mel spectrogram instead of using explicit pitch information and pitch-related network. Furthermore, we optimize the system by model pruning, multi-frame inference, and increasing frame length, to enable it to meet the conditions required for real-time applications. The objective and subjective evaluation results for various languages show that the proposed system is more stable for tonal languages within the proposed optimization strategies. The experimental results also verify that our model improves synthesis runtime by 3.12 times than that of the baseline on a standard CPU while maintaining naturalness.

Index Terms: Text-to-speech synthesis, real-time, LPCNet, optimization, Mel spectrogram

1. Introduction

Text to speech (TTS) has attracted significant attention in recent years owing to advances in deep learning. Significant developments have been made in neural end-to-end TTS synthesis models for generating high-quality speech using a simplified pipeline. Deep neural network-based systems have become increasingly popular for TTS, such as Tacotron [1], Tacotron 2 [2], FastSpeech [3], and the fully end-to-end ClariNet [4]. Usually, these models first generate a Mel spectrogram autoregressively from text input and then synthesize speech from the Mel spectrogram using a vocoder such as Griffin-Lim [5], WaveNet [6], or WaveGlow [7].

A drawback of this approach is that the resulting systems use large neural network (NN) models. This may lead to a computationally heavy and slow synthesis process, even on a powerful GPU. An efficient neural vocoder called LPCNet was recently introduced [8]. The LPCNet inference runs faster than real time on a single CPU, while producing a high-quality speech output. In [9], a considerable quality improvement was achieved by modifying a TTS system that produced the WORLD [10] vocoder parameters to predict the parameters for LPCNet. In [11], LPCNet was used to build a high-quality streaming speech synthesis system with low sentence-length-independent latency. In [12], several parallelization techniques applicable to LPCNet were investigated, and a TTS system suitable for low-to-mid range mobile devices was proposed.

Although several solutions [9, 11, 12, 13, 14] have been claimed to be capable of faster-than-real-time on-device speech generation and used relatively complex optimization schemes, input acoustic features have not been extensively discussed in the literature, to the best of our knowledge. Tacotron2 and LPCNet are usually integrated by replacing the output Mel spectrogram of the original Tacotron2 with the native features of LPCNet, that is, a 20-dimensional vector consisting of 18 Bark-scale cepstral coefficients (BFCCs) and two pitch parameters (period and correlation). However, it is yet unclear if the 20-dimensional input acoustic features are sufficiently robust within the combined optimization methods.

Unlike conventional approaches [1, 2] to end-to-end speech synthesis, which typically consider the spectral features in which the fundamental frequency (F_0) is implicit as targets, LPCNet uses explicit pitch prams as features. Suprasegmental features, such as pitch, are important cues in speech signals. Achieving high-quality end-to-end speech synthesis in pitch accent or tonal languages (such as Japanese and Chinese) will be significantly difficult, largely because of the character diversity and pitch accents [15, 16]. In [15], it was identified that choosing vocoder parameters such as F_0 is beneficial to pipeline systems; however, this is in contradiction to the case of Tacotron. This is because Tacotron generates alignment errors owing to the prediction of longer F_0 sequences. Even the small error in pitch parameters prediction from acoustic model will cause significant audio quality decrease in tonal languages.

In this paper, we propose a novel end-to-end TTS model TacoLPCNet by replacing the explicit pitch parameters with implicit spectral features. The contributions of this paper are highlighted as follows:

- We extend the original LPCNet for Mel spectrogram and analyze two different acoustic features (explicit and implicit pitch parameters) for LPCNet input.
- For real-time speech synthesis on devices running without a GPU, we describe further optimizations of the synthesis system by model pruning, multi-frame inference, and increasing frame length.
- To the best of our knowledge, this is the first study to analyze the adaptability between two different features and various languages under an on-device LPCNet-based synthesis system.

* Corresponding Author

http://dx.doi.org/10.21437/Interspeech.2021-852

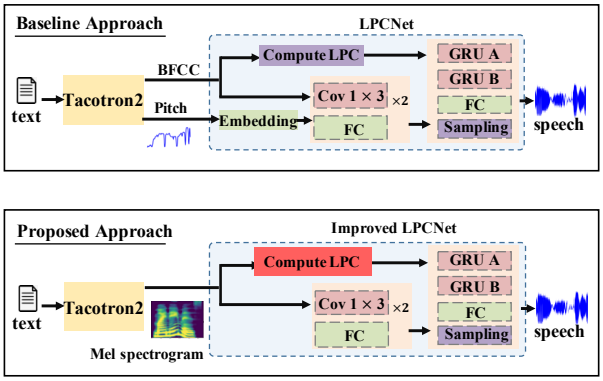

Figure 1: *An overview of the proposed model architecture.*

We conducted experiments on different language datasets, including tonal and non-tonal languages. We demonstrate that the proposed system can generate 16 kHz speech faster than real-time on a general-purpose CPU.

The paper is organized as follows: in Section 2, we describe the design of the improved LPCNet vocoder for Mel spectrograms; in Section 3, various optimization techniques are discussed; in Section 4, objective and subjective evaluation results are presented along with efficiency measurement; finally, the conclusion is provided in Section 5.

2. Improved LPCNet vocoder

When using LPCNet as the vocoder of Tacotron2, most previous researchers directly replaced the Mel spectrogram with 20-dim LPCNet features for Tacotron2, similar to the baseline in Figure 1. Only small changes are necessary to the output of Tacotron2. We propose another integration method that keeps the acoustic feature prediction intact so that the 80-channel Mel spectrogram outputs from Tacotron2 are directly fed to the LPCNet model

2.1. Integration through spectrogram features

In the original LPCNet work, the input of the synthesis is limited to just 20-dim features: 18-dim BFCC and two pitch parameters (period and correlation). To learn better pitch-related information, the above two pitch parameters are first passed into an embedding layer. Then, the learned pitch embedding and the 18-dim BFCC features are simultaneously fed into the frame-rate network to generate frame-level context vectors.

When integrating through the Mel spectrogram, the pitch and voice features are not explicitly presented to the neural vocoder but are instead contained implicitly in the input. Without two pitch parameters, a pitch embedding layer is no longer required. Only the 80-dim Mel spectrogram is fed into the frame-rate network to generate frame-level context vectors, as proposed approach shown in Figure 1.

2.2. Linear prediction

Knowing that the vocal tract response can be represented reasonably well by a simple all-pole linear filter, LPCNet use the linear filter to model the vocal tract rather than the NN model. The speech signal is reconstructed by passing the generated excitation signal through the linear prediction synthesis filter, as follows:

$$x_t = e_t + p_t \tag{1}$$

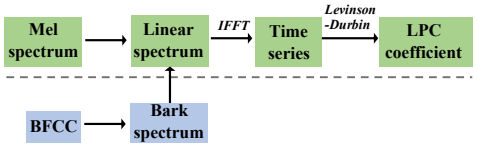

Figure 2: *LPC coefficient estimation process from features.*

$$p_t = \sum_{k=1}^{m} a_k s_{t-k} \tag{2}$$

where x_t, e_t, and p_t denote the n^th sample of speech, excitation, and intermediate prediction terms, respectively, and a_k denotes the i^th LPC coefficient with the order m.

In TacoLPCNet, we developed an algorithm for a_k from the Mel spectrogram, similar to the LPCNet, as shown in Figure 2. The prediction LPC coefficients, a_k, were computed by first converting the Mel spectrum into a linear spectrum. The linear spectrum was then converted to an auto-correlation time series using an inverse FFT. From the time series, the Levinson-Durbin[17] algorithm was used to compute the LPC coefficient. Although the LPC analysis computed using this approach is not as accurate as that computed on the input signal (owing to the low resolution of the Mel spectrum), the effect on the output is minimal because the network can learn to compensate.

3. Optimization strategies

Tacotron2 and LPCNet have a similar RNN autoregressive structure and predict a frame Mel spectrogram or a sample instantly. Thus, the time of speech synthesis is composed of two parts: the Mel spectrogram and sample generation. The final time, T_s, required for TacoLPCNet to generate S_l samples is as follows:

$$\begin{aligned} S_l &= F_l * N_f \\ T_s &= T_t + T_l \end{aligned} \tag{3}$$

where F_l and N_f denote that the length of each frame is l and the number of frames is N, and T_t and T_l represent the time consumed by Tacotron2 and LPCNet, respectively. Subsequently, T_t can be defined as follows:

$$T_t = (C_t * N_f)/M_f \tag{4}$$

where C_t is the Tacotron2 per frame execution time, M_f is the number of frames generated at once decoder step. Similarly, T_l can be defined as follows:

$$T_l = C_l * F_l * N_f = C_l * S_l \tag{5}$$

where C_l denotes the time required for LPCNet to generate one sample; therefore, the time consumed by LPCNet is only related to C_l and the length of speech S_l. To reduce its huge computational complexity and deploy real-time services, we made the following optimization attempts.

Model pruning. The first bottleneck was created by the high computational cost of the Tacotron2 decoder comprising a two-layer LSTM with 1024 units. We chose to decrease the size of the LSTM using three different level factors. As in [8], the complexity of the LPCNet model can be mainly attributed to two GRU, as well as a dual fully connected layer. In TacoLPCNet, we also decreased the mainly GRUA size of the LPCNet. In other words, we reduce the execution times C_t and C_l by straight-forward model pruning.

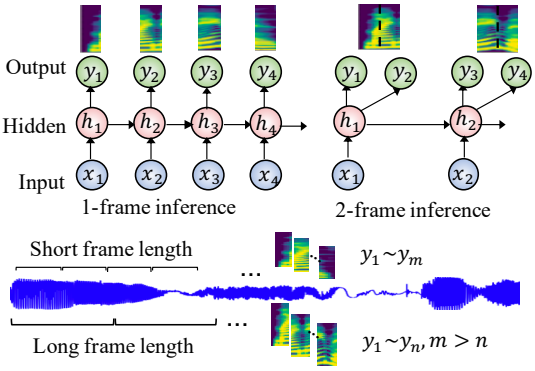

Figure 3: *Computation graph of multi-frame inference and increasing frame length.*

Multi-frame inference and increasing frame length. Instead of reducing C_t, we can also reduce the time T_t consumed by Tacotron2 by decreasing N_f or increasing M_f. To synthesize an entire piece of audio with length of S_l, N_f is only related to the per-frame length F_l. The longer the per-frame length F_l is generated at each step, the fewer steps the model decodes. This means that we can decrease N_f in Eq. 4 by increasing the frame length F_l in Eq. 3.

As long as Tacotron operated on speech frames, Tacotron2 was also capable of predicting several speech frames in one step [1, 2]. Instead of predicting one acoustic frame, multiple acoustic frames are jointly predicted at the same time. That is M_f in Eq. 4 is increased. Figure 3 illustrates the concept of multi-frame inference and the increasing frame length. Moreover, these two methods, multi-frame inference and increasing frame length, can be used independently or simultaneously. Note that multi-frame inference and increasing frame length do not affect the time consumed by LPCNet because the length of the total speech S_l remains constant in Eq. 5.

4. Experiments

4.1. Data

We conducted experiments on four different language corpora, including English, Chinese, Japanese, and Korean. We used the LJSpeech [18] dataset for English, which contains 13,100 paired text/audio examples. We used the BiaoBei [19] dataset for Chinese. High-quality speech corpora, including Japanese and Korean, recorded by native speakers in a professional studio, were also used in our experiments, and these two private datasets were provided by Huiyan-tech [1]. We used a sampling rate of 16 KHz for all audios. The details of the corpora are listed in Table 1. One hundred utterances that were not included in the training data were used for evaluation. A subset of the samples used in the listening test is available at `https://gongchenghhu.github.io/TacoLPCNet-demo/`.

4.2. Model

We trained the following two models:

- **Baseline:** The baseline model trained with 20-dim LPCNet features and original LPCNet.
- **TacoLPCNet:** The proposed model was trained with Mel spectrograms and improved LPCNet.

[1] https://huiyan-tech.com/

Table 1: *Details of the corpus used in our experiments.*

	English	Chinese	Japanese	Korean
Duration (h)	23.4	11.3	4.21	3.46
#Sent	13,100	10,000	2500	2469

Table 2: *Mean opinion score (MOS) evaluations with 95% confidence intervals computed from the t-distribution for various systems. Note that the number of units in the LSTM of Tacotron2 decoder is fixed at 512 in this group test, and LPCNet is evaluated using the main GRUA of size equal to 384.*

Corpus	Methods		Ground-truth
	Baseline	TacoLPCNet	
English	3.878 ± 0.099	4.072 ± 0.121	4.543 ± 0.102
Chinese	4.038 ± 0.108	4.260 ± 0.083	4.633 ± 0.072
Japanese	3.857 ± 0.128	4.119 ± 0.113	4.562 ± 0.117
Korean	3.552 ± 0.200	4.225 ± 0.119	4.547 ± 0.139

For both systems, we trained Tacotron2 with 16 GB RAM and Titan RTX GPUs, using a batch size of 36 for 200k steps and trained LPCNet using a batch size of 64 for 120 epochs.

4.3. Evaluation of different features in various languages.

Subjective evaluation. In terms of audio quality, we used the mean opinion score (MOS) as the metric, as in [2]. In all tests, 30 native listeners were asked to rate the performance of 50 randomly selected synthesized utterances from the test set. Table 2 summarizes the MOS test results, the trends of which can be analyzed as follows: (1) TacoLPCNet outperformed the baseline model. This result confirms that spectral features are better than spectral envelopes for generating speech using the LPCNet vocoder. (2) Although there is no explicit pitch period parameters, the Mel spectrogram features already contain sufficient fundamental frequency information. (3) The pitch embedding structure in the LPCNet vocoder is not necessary.

Comparing the synthesized voice, the most significant defect of the sound synthesized by the baseline model is the existence of obvious tremolo in Chinese, Korean, and Japanese. In terms of English, the performance of the two models is similar. It is because English is not a tonal or pitch accent language, and the speech generated by the baseline model has few tremolos. The lack of character diversity and pitch accents makes it easy to generate speech without tremolo.

Objective evaluation. We also used the root mean square error of F_0 (F_0RMSE) and mel-cepstrum distortion (MCD) as the objectivity metrics , as in [20]. The FastDTW [21] algorithm was adopted to align the predicted acoustic feature sequences with the natural ones. The objective results are presented in Table 3. From Table 3, we can observe that TacoLPCNet outperforms baseline in the objective evaluation as well. In both systems, Korean and Chinese have a high F_0 error. This can be explained by the fact that Korean and Chinese had higher and more variational F_0 contours, which were more difficult to model. Although a small F_0 error was obtained for Japanese, the subjective test for Japanese did not provide satisfying results. This means that the original LPCNet is particularly sensitive to the pitch parameters. Although the baseline has a pitch embedding layer, even the small error in pitch parameter prediction will cause significant shaking in the final synthesized sound. Figure 4 shows an example where we use the baseline

Table 3: *Objective evaluation results of the baseline model and proposed TacoLPCNet.*

Corpus	Metric	Methods	
		Baseline	TacoLPCNet
English	F_0RMSE(Hz)	32.69	31.18
	MCD(dB)	10.82	10.75
Chinese	F_0RMSE(Hz)	42.13	36.81
	MCD(dB)	5.76	5.38
Japanese	F_0RMSE(Hz)	23.71	19.07
	MCD(dB)	7.52	7.48
Korean	F_0RMSE(Hz)	44.26	37.13
	MCD(dB)	6.71	6.15

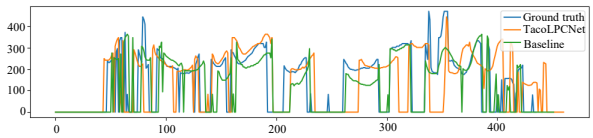

Figure 4: *Groud truth, TacoLPCNet and baseline pitch contours.*

model and TacoLPCNet to generate one Chinese speech. From the Figure 4, we can found that TacoLPCNet's pitch contour is closer to the source than that of the baseline model.

4.4. Evaluation of optimization methods

To evaluate the efficiency of the optimization methods, we tested the all systems on an Intel(R) Xeon(R) Silver 4110 CPU. We calculated the real time factor (RTF), which is defined as the time necessary to synthesize a piece of audio divided by the duration of the audio. For convenience, we define three different optimization levels H(high), M(medium), and L(low), as shown in Table 4. For example, MH_LM has a configuration of 512 Tacotron2 decoder LSTM units, 384 LPCNet GRUA units, 1-frame inference, and a frame length of 20 ms. Due to space limitations, we only show the optimization results for Korean.

Which system is more robust within the optimization?: From Figure 5, we can observe that TacoLPCNet achieves a better performance than the baseline. Although the HH_LL model with baseline has achieved a good result in audio quality, the inference speed is the worst with the low-level optimization methods. Once this baseline model is optimized, such as increasing the frame length (MH_LM) and 2-frame inference (MH_ML), the performance of the baseline drops sharply. When using the same optimization methods, TacoLPCNet can achieve a far better audio quality. Therefore, we consider that TacoLPCNet is more robust with optimization, and the improved LPCNet with Mel spectrogram could maintain a stable performance with a high inference speed.

The effect of the different level optimization methods: *Model pruning.* As shown in Figure 5, comparing HH_LL with MH_LL in baseline, the reduction of units in the LSTM of Tacotron2 decoder could speed up feature prediction (low RTF), which also results in degrading the sound quality to a certain extent. For LPCNet, the reduction of units in the LPCNet GRUA cloud increases the speed of waveform generation and only causes a slight loss of audio quality.

Multi-frame inference and increasing frame length. It is well known that speech signals remain stable for a short time [22]. Comparing MH_LL with MH_LM in the baseline

Table 4: *Different optimization levels.*

Corpus	Optimization level		
	High (H)	Medium (M)	Low (L)
T2 LSTM	1024	512	384
LPCNet GRUA	384	256	192
Frames per step	3	2	1
Frame length (ms)	30	20	10

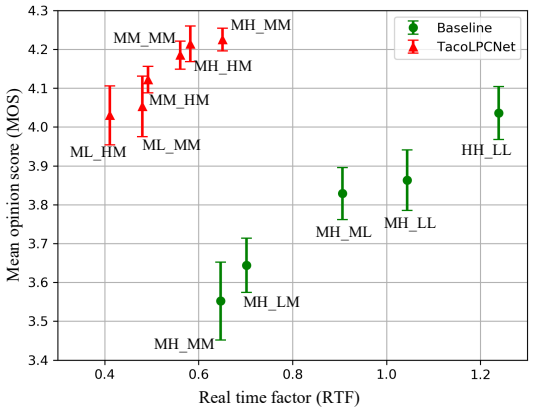

Figure 5: *A comparison of between baseline and TacoLPCNet within various level optimization.*

model, although the speed has been dramatically improved, the increase in the frame length causes a drop in audio quality. Therefore, we increased the frame length from 10 ms to 20 ms and maintained it at 20 ms. It is still a challenge to predict acoustic features in the long term. For multi-frame inference, when generating multiple frames (such as two or three) simultaneously, the speed is improved with a slight drop in audio quality. Comparing MH_LM and MH_ML in the baseline model, the acceleration effect of multi-frame inference is not as good as increasing the frame length.

5. Conclusions

In this article, we present a new TTS system TacoLPCNet by extending LPCNet for Mel spectrograms. Furthermore, we optimized this synthesis system to enable it to meet the conditions required for real-time applications. The optimization strategies included model pruning, multi-frame inference, and increasing frame length. TacoLPCNet addresses the challenging goal of producing stable speech while operating at a faster rate than real time without GPU support. Objective and subjective evaluation results on different language corpus show that TacoLPCNet is more robust for tonal languages at a faster runtime and maintains naturalness. The results also demonstrate that TacoLPCNet can generate high-quality audio at a 3.12 times faster runtime than baseline owing to our optimizations.

6. Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2018YFB1305200, the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330 and the National Natural Science Foundation of China under Grant 61771333.

7. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, Z. Y. Jaitly, Y. Xiao, Z. Chen, S. Bengio, Q. Le *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[2] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, 2019, pp. 3165–3174.

[4] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *International Conference on Learning Representations*, 2018.

[5] 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.

[6] 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.

[7] 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.

[8] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[9] Z. Kons, S. Shechtman, A. Sorin, C. Rabinovitz, and R. Hoory, "High quality, lightweight and adaptable tts using lpcnet," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 176–180, 2019.

[10] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[11] N. Ellinas, G. Vamvoukakis, K. Markopoulos, A. Chalamandaris, G. Maniati, P. Kakoulidis, S. Raptis, J. S. Sung, H. Park, and P. Tsiakoulis, "High quality streaming speech synthesis with low, sentence-length-independent latency," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2022–2026, 2020.

[12] V. Popov, S. Kamenev, M. Kudinov, S. Repyevsky, T. Sadekova, V. K. Bushaev, and D. Parkhomenko, "Fast and lightweight on-device tts with tacotron2 and lpcnet," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 220–224, 2020.

[13] 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)*. IEEE, 2020, pp. 6204–6208.

[14] 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," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020.

[15] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6905–6909.

[16] Y. Zhao, H. Li, C.-I. Lai, J. Williams, E. Cooper, and J. Yamagishi, "Improved prosody from learned f0 codebook representations for vq-vae speech waveform reconstruction," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 4417–4421, 2020.

[17] P. Castiglioni, "Levinson-durbin algorithm," *Encyclopedia of Biostatistics*, vol. 4, 2005.

[18] K. Ito, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[19] B. D.-B. Sciene and T. Ltd, "Chinese standard mandarin speech copus," https://www.data-baker.com/open_source.html, 2018.

[20] P.-f. Wu, Z.-h. Ling, L.-j. Liu, Y. Jiang, H.-c. Wu, and L.-r. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *Asia-Pacific Signal and Information Processing Association (APSIPA)*, 2019.

[21] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[22] M. Portnoff, "Short-time fourier analysis of sampled speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 364–373, 1981.