INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

DurIAN: Duration Informed Attention Network For Speech Synthesis

Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, Dan Su, Dong Yu

Tencent AI Lab

{czyu,bearlu,ninahu,raymondyu,cweng,kunxu,deyituo,shiyinkang,dansu,dyu}@tencent.com

Abstract

In this paper, we present a robust and effective speech synthesis system that generates highly natural speech. The key component of proposed system is Duration Informed Attention Network (DurIAN), an autoregressive model in which the alignments between the input text and the output acoustic features are inferred from a duration model. This is different from the attention mechanism used in existing end-to-end speech synthesis systems that accounts for various unavoidable artifacts. To improve the audio generation efficiency of neural vocoders, we also propose a multi-band audio generation framework exploiting the sparseness characteristics of neural network. With proposed multi-band processing framework, the total computational complexity of WaveRNN model can be effectively reduced from 9.8 to 3.6 GFLOPS without any performance loss. Finally, we show that proposed DurIAN system could generate highly natural speech that is on par with current state of the art end-to-end systems, while being robust and stable at the same time.

Index Terms: speech synthesis, end-to-end, neural vocoder

1. Introduction

Traditional speech synthesis approaches, including concatenative methods [1, 2] and statistic parametric systems [3, 4, 5], are all based on acoustic feature analysis and synthesis. These approaches are still predominantly used in industrial applications due to their advantages in robustness and efficiency. However, these approaches suffer from the inferior naturalness of generated speech. End-to-end approaches [6, 7, 8, 9, 10, 11] have gained much attention recently due to the naturalness of their synthesized results and simplified training pipelines. However, existing end-to-end systems are lack of robustness as they produce unpredictable artifacts where random words in the source text are repeated or skipped in generated speech [7, 11]. The lack of robustness in end-to-end systems has significantly limited it's applications in real production environments.

To combine the strength of traditional parametric systems and current end-to-end models, we propose a new speech synthesis framework, duration informed attention network (DurIAN), that generates highly natural and robust speech [1] at the same time. DurIAN is a hybrid of traditional parametric and recent end-to-end systems where the end-to-end attention mechanisms in recent end-to-end model are replaced with an alignment model that is similar to the one used in parametric systems [2]. While end-to-end speech synthesis systems have advanced the traditional parametric ones from various perspectives, the end-to-end attention mechanism has been the root cause of instability in generated speech. Therefore, the motivation of proposed DurIAN model is to preserve most advancements in existing end-to-end systems while discarding end-to-end attention mechanism causing various unstable artifacts. We show that the proposed DurIAN model could generate highly natural speech that is on par with current end-to-end systems, while at the same time the generated speech is much more robust and stable.

Another limitation of end-to-end systems is the generalization of speech synthesis performance on out-of-domain text, especially for language such as Mandarin where word and prosody boundaries plays more important role. While it can be argued that both word and prosody segmentations can be jointly learned in end-to-end systems, it normally does not generalize well on out-of-domain context due to relatively limited training data for speech synthesis. Therefore, it is still more desirable to have a separate model for learning word and prosody boundaries from larger dataset [13, 14, 15] in tasks such as Mandarin speech synthesis. However, how to incorporate predicted prosodic boundaries in hybrid model such as proposed DurIAN or FastSpeech [12] has yet been investigated. In this study, we propose a skip encoder structure to effectively utilize predicted hierarchical prosodic boundaires for improved generalization of DurIAN model in Mandarin speech synthesis tasks.

Finally, the computational complexity of neural vocoders [16, 17, 18, 19] is another challenge when applying recent end-to-end models in realtime speech syntheis applications. While neural vocoders are capable of achieving highly natural speech that significantly surpasses the conventional vocoders, their computational complexity is also significantly higher causing slow inference speed, larger latency, and more expensive deployment cost. As most neural vocoders are designed to predict audio signals sample by sample, even one second of speech/audio requires tens of thousands of inference steps. In this work, we propose a novel multi-band audio generation framework that effectively exploits the sparseness characteristics of neural network to reduce the total computational complexity of WaveRNN model [18] from 9.8 to 3.6 GFLOPS, and improves the speech generation speed more than two times on a single CPU core.

2. DurIAN

In this section, we describe the main components in the DurIAN speech synthesis system. As DurIAN is a text driven system, it takes a sequence of symbols converted from text and outputs mel-spectrogram. The architecture of DurIAN is illustrated in Figure. 1.

[1] Sound and video demo can be found at https://tencent-ailab.github.io/durian/

[2] At the time of preparing this paper, we became aware of a preprint paper [12] where a similar idea was proposed to address the robustness issues related to the end-to-end systems. Our work is independently developed with many design choices are completely different.

http://dx.doi.org/10.21437/Interspeech.2020-2968

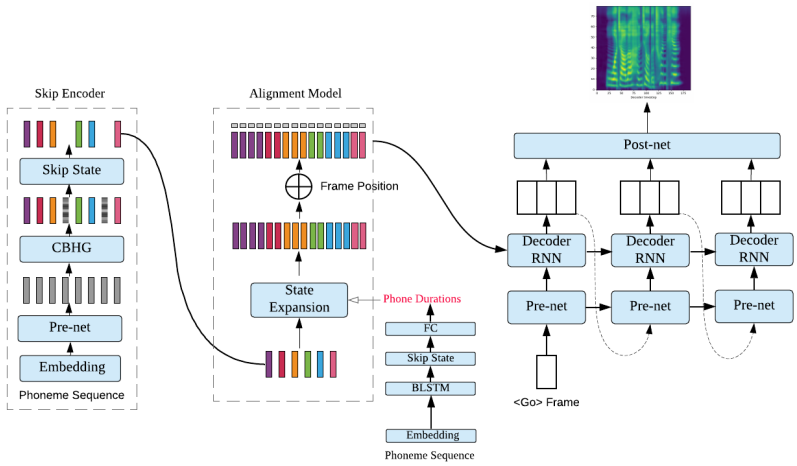

Figure 1: *Model architecture of DurIAN. The model takes a sequence of symbols, including phonemes and prosodic boundaries between them, and outputs the corresponding mel-spectrogram.*

2.1. Skip Encoder

The main objective of the skip encoder is to encode the representation of phoneme sequences as well as hierarchical prosody boundaries in the hidden states. The prosodic boundaries is an important component for improved generalization of speech synthesis system on out-of-domain text in Mandarin speech synthesis tasks.

To generate the input to the skip encoder, the source text is first converted to a sequence of phonemes. To encode different levels of prosody structures, we insert special symbols representing different levels of prosody boundaries between input phonemes. Figure. 2 illustrates an example how these special symbols representing prosodic boundaries are inserted.

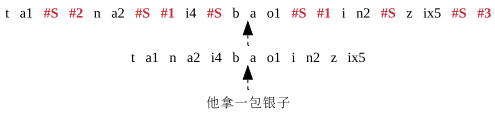

Figure 2: *An illustration of how prosodic boundaries are inserted between input phonemes. The symbol #S represents the boundary of syllables, #1 represents the boundary of prosodic words, #2 represents the boundary of prosodic phrase, and #3 represents the boundary of intonational phrase.*

The main network components in the skip encoder is inherited from the encoder in the Tacotron 1 [6] system. Each phoneme and inserted prosodic symbol in the input phoneme sequence is first converted to a continuous vector in the embedding space. The embedded representation of the phoneme sequence is then passed through the *pre-net* [6] that contains two fully connected layers followed by the CBHG [6] module. Dropout with probability of 0.5 is applied on the *pre-net* during training. The output from the CBHG module is a sequence of hidden states containing sequential representation of the input text. Since prosodic boundaries are not physically aligned with any target acoustic features, the hidden states associated with prosodic boundaries are excluded from the output of the CBHG model. An alternative approach for encoding prosodic boundaries is to convert the phoneme sequence into manually composed linguistic features where prosodic structures are encoded. However, our early experiments show that using skip encoder could generate speech that is more natural than using linguistic features.

2.2. Alignment Model

One important task in speech synthesis is uncovering the hidden alignment between the phoneme sequence and the target feature/spectrum sequence. End-to-end systems rely on attention based mechanism to discover such alignment. However, existing end-to-end attention mechanism frequently generates unpredictable artifacts where some words are skipped or repeated in the generated speech. Since production speech synthesis systems have very low tolerance on such instability, end-to-end speech synthesis systems have not been widely deployed in practical applications. In DurIAN, we replace the attention mechanism with an alignment model [20, 21], in which the alignment between the phoneme sequence and the target acoustic sequence is inferred from a phoneme duration prediction model as illustrated in Figure. 1. The duration of each phoneme is measured by the number of aligned acoustic frames. During training, the alignment between the acoustic frame sequence and the input phoneme sequence can be obtained through forced alignment widely used in speech recognition. The alignment is then used for hidden state expansion, which simply replicates hidden states according to phoneme duration. During synthesis, a separate duration model is exploited to predict the duration of each phoneme. This duration model is trained to minimize the mean squared error between the predicted phoneme duration and the duration obtained through forced alignment, given the whole sentence. After state expansion, the relative position of every frame inside each phone is encoded as a value between 0 and 1, and appended to the encoder state. The expanded encoder states are analogous to the attention context estimated in the end-to-end system, except that in DurIAN they are inferred from the predicted phone duration.

The duration model used in DurIAN is similar to the ones used in the conventional statistical synthesis models. It consists of three 512-unit bidirectional LSTM layers. Similar to that in the skip encoder, the states associated with the prosodic boundaries are also skipped before the final fully connected layer.

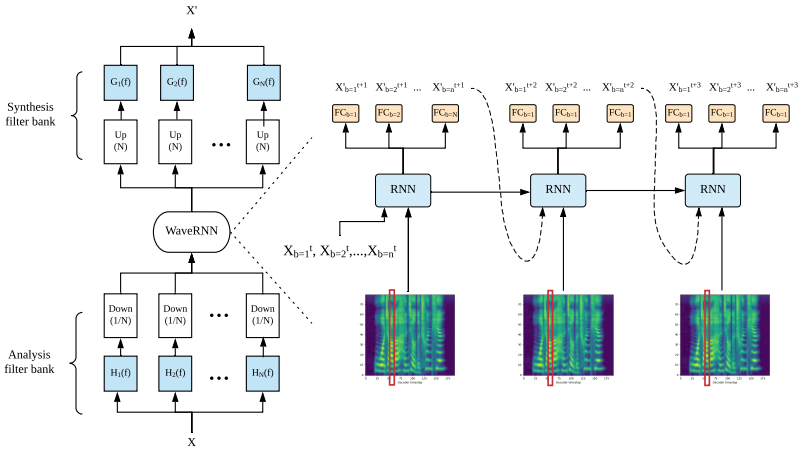

Figure 3: *Model architecture of Multi-band WaveRNN.*

2.3. Decoder

The decoder used in DurIAN is similar to the one used in Tacotron 1 [6]. The only difference is that the attention context concatenated with the *pre-net* output is replaced with the encoder states derived from the alignment model in DurIAN. As in Tacotron, the decoder network can output single frame or multiple non-overlapped frames at each time step. When the target is multiple non-overlapped frames, a restricted attention is applied to the encoder states aligned with the target frames, and then concatenated with the output of the *pre-net* at each time step. The attention mechanism used in DurIAN is different from that used in the end-to-end systems. In DurIAN, the attention context is computed from a small number of encoder frames that are aligned with the target frames. As long as the number of frames per decoder time step is not extraordinary large, it will not cause the similar artifacts observed in the end-to-end systems. The content-based tanh attention [22] is used in our system and dropout with a probability of 0.5 is applied to the *pre-net* in the decoder network during both training and inference.

3. Multi-band WaveRNN

3.1. WaveRNN

The WaveRNN model we use follows the structure in [18]. A single-layer recurrent network and a dual-softmax layers is used to generate 16-bit audio. To accelerate WaveRNN inference speed, we performed 8-bit quantization (quantize aware training) on the hidden-layer weights of Gated recurrent units (GRU) as well as other four fully connected layers following it. The quantization significantly reduces the model size, which is very helpful to increase the cache hit rate. Moreover, the quantized parameters help to accelerate the calculations using the avx2 instruction of the Intel CPU. The combination of quantization and avx2 instruction could achieve 4x faster inference speed than floating-point calculations.

3.2. Multi-band Processing

To further reduce the computational complexity of WaveRNN model, we propose general multiband processing framework for neural vocoders. One important characteristic of mulitband processing is that when signal is splited into N subbands, the signal in each subband can be downsampled by N times without loss of information. In other words, if we model each subband with a separate neural vocoder, the inference step in each subband can be reduced by N time. Existing approaches [23, 24] utilizing such charactertistics of multiband processing is to model each subband with separatly trained neural vocoder which will then be inferenced in parallel using mulitple CPUs or GPUs. However, such parallization based approach while improving the inference speed, it does not reduce the fundamental computational complexity of neural vocoders. On the other hand, our proposed multi-band WaveRNN algorithm exploits the sparseness of the neural network model and uses a single shared WaveRNN model for all subband signal predictions. More specifically, the shared WaveRNN model takes all subband samples predicted from the previous step as input and predicts next samples in all subbands in one inference step as illustrated in Figure. 3. We modify the original WaveRNN model to take inputs from multiple subbands and predict samples for all subbands simultanously through multiple output (and softmax) layers. As the total inference steps reduces by a factor of N (the number of frequency bands), the total computational complexity can also be significantly reduced. The predicted audio signals in each frequency band are upsampled and then passed to synthesis filters. The signals from each frequency band after synthesis filter are summed to single audio signal.

3.3. Filter Design For Multiband WaveRNN

A stable yet more efficient low cost filter bank, called Pseudo Quadratue Mirror Filter Bank (Pseudo-QMF), is employed for our multi-band processing. The prototype filter is designed to have a linear phase, leading to a phase-distortion-free analysis/synthesis system. The sampling frequency of each subband waveform is f_s/N Hz, where N is the number of filter channels/subbands and f_s is the desired sampling rate for the fullband signal. With the property of aliasing cancellation of Pseudo-QMF, the critical downsampling is applied after the fullband signal is decomposed into subbands by the analysis filterbank.

4. Experiments

4.1. Speech Synthesis

We evaluated the naturalness and robustness of the proposed system using two different datasets. The first dataset is based on 18 hours of male speech and the other one is based on 7 hours

of female speech recording. Both are Mandarin speech datasets. All the training data has a sampling rate of 16KHz. Mean Opinion Score (MOS) of the naturalness of generated speech utterances are rated by human subjects participated in the listening tests. We use 40 unseen sentences for evaluating the models trained from the male speaker, and 20 relatively longer out-of-domain sentences for evaluating the models trained from the female speaker. In all the experiments, 20 native Mandarin speakers participated in the listening test. We compared our model with the traditional BLSTM-based parametric system [20] and the Tacotron-2 system. As shown in Table. 1, DurIAN and Tacotron 2 perform significantly better than the traditional parametric system. In both tests DurIAN and Tacotron-2 perform on-par with each other. No statistically significant difference can be observed. These results tell us that the superior naturalness in Tacotron-2 is likely a result of all other components in Tacotron other than the end-to-end attention mechanism.

Table 1: *5-scale mean opinion score evaluation.*

	Male	Female
Parametric	3.54	3.47
Tacotron 2	4.10	4.28
DurIAN	4.11	4.26

As the design goal of DurIAN is to achieve the naturalness comparable to Tacotron 2 while avoiding the artifacts observed in the Tacotron 2 system, We further compared two systems in robustness of generated speech. In this evaluation, we mainly focused on the word skipping and repeating errors commonly occur in the Tacotron 2 systems. Both DurIAN and Tacotron 2 systems were used to generate 1000 unseen utterances. The occurrence rate of word skipping and repeating errors are listed in Table. 2. These results clearly indicate that DurIAN is much more robust than Tacotron-2 and generated no error in this category.

Table 2: *Word skipping or repeating errors rate.*

	skip/repeat
Tacotron 2 [7]	1%
Deep Voice 3 [11]	4%
Tacotron 2	2%
DurIAN	0%

4.2. Multi-band WaveRNN

We evaluated the naturalness of generated speech and the speed of the Multi-band WaveRNN.

4.2.1. Speed

The main complexity of WaveRNN comes from two GRUs and four fully-connected layers. We ignore the overhead of additive operations and focus only on the complexity of multiplication operations for each sample generated, which is

$$\mathrm{C} = 2*(2 * \mathrm{N_G} * \mathrm{N_G} * 3 + \mathrm{N_G} * \mathrm{N_F} + 256 * \mathrm{N_G} * \mathrm{N_B}) * \mathrm{S_R} / \mathrm{N_B}, \quad (1)$$

where $\mathrm{N_G}$ is the size of the two GRUs, $\mathrm{N_F}$ is the width of affine layer connected with final fully-connected layer, $\mathrm{N_B}$ is the number of frequency band, and $\mathrm{S_R}$ is the sampling rate. Using $\mathrm{N_G} = 192$, $\mathrm{N_F} = 192$ and $\mathrm{N_B} = 1$ (fullband WaveRNN) for $\mathrm{S_R} = 16000$, we obtain a total complexity around 9.8 GFLOPS. When we set $\mathrm{N_B} = 4$, the total complexity is 3.6 GFLOPS.

We also measured the Real Time Factor (RTF) for Multi-band WaveRNN systems listed in Table 3. All the RTF values were measured on a single Intel Xeon CPU E5-2680 v4 core. The results show that with quantization and avx2 speedup, the RTF can be reduced from 1.337 to 0.387 for the baseline WaveRNN model. With the 4-band model, the RTF can be further reduced to 0.171, which is 2x times faster than quantized WaveRNN model.

Table 3: *Real Time Factor (RTF) evaluation of proposed Multi-band WaveRNN.*

RTF	fullband	4band
float	1.337	0.503
int8	0.387	0.171

4.2.2. Quality

The Mean Opinion Scores (MOSs) of proposed multi-band WaveRNN were obtained through subjective listening tests. The female dataset used in Sec. 4.1 was used for training both the DurIAN and WaveRNN models. Three WaveRNN systems, the baseline WaveRNN model without quantization and the 4-band WaveRNN model with and without quantization, were compared. Experimental results in Table. 4 indicate that the three systems evaluated are on-par with each other. No statistically significant difference was observed. If fact, most of the subjects participated in the listening tests cannot feel any difference between utterances generated from these three different WaveRNN systems. We can conclude that the proposed multi-band synthesis approach and the 8-bit quantization technique can effectively reduce the computational cost without deteriorating the quality of the generated speech.

Table 4: *5-scale mean opinion score (MOS) evaluation of the proposed Multi-band WaveRNN.*

Systems	MOS
Fullband WaveRNN (float)	4.53
4-band WaveRNN (int8)	4.58
4-band WaveRNN (int8)	4.56

5. Conclusions

In this paper, we presented a speech synthesis framework that is capable of generating highly natural and robust speech. Our experimental results indicate that the proposed DurIAN system could synthesize speech with the naturalness and quality on par with the current state of the art end-to-end system Tacotron 2, at the same time effectively avoid the word skipping and repeating errors in generated speech. We also proposed multi-band speech generation algorithm which effectively reduce the computational complexity of WaveRNN model from 9.8 to 3.6 GFLOPS without deteriorating the quality of generated speech. Finally, the proposed DurIAN model is a general synthesis framework which we have successfully extended for other generation tasks such as singing [25, 26], multimodal synthesis [27], and fine-grained style controlled speech synthesis [27].

6. References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1. IEEE, 1996, pp. 373–376.

[2] A. W. Black and P. A. Taylor, "Automatically clustering similar units for unit selection in speech synthesis." 1997.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 3. IEEE, 2000, pp. 1315–1318.

[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *speech communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[5] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7962–7966.

[6] 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," *arXiv preprint arXiv:1703.10135*, 2017.

[7] 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.

[8] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality tts with transformer," *arXiv preprint arXiv:1809.08895*, 2018.

[9] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[10] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," 2017.

[11] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[12] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *arXiv preprint arXiv:1905.09263*, 2019.

[13] Y. Qian, Z. Wu, X. Ma, and F. Soong, "Automatic prosody prediction and detection with conditional random field (crf) models," in *2010 7th International Symposium on Chinese Spoken Language Processing*, 2010, pp. 135–138.

[14] Zhiwei Ying and Xiaohua Shi, "An rnn-based algorithm to detect prosodic phrase for chinese tts," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 809–812 vol.2.

[15] J. Pan, X. Yin, Z. Zhang, S. Liu, Y. Zhang, Z. Ma, and Y. Wang, "A unified sequence-to-sequence front-end model for mandarin text-to-speech synthesis," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020. [Online]. Available: http://dx.doi.org/10.1109/icassp40776.2020.9053390

[16] 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.

[17] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[18] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[19] 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.

[20] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks," *INTERSPEECH*, pp. 1964–1968, Singapore, September, 2014.

[21] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, "Fast, compact, and high quality lstm-rnn based statistical parametric speech synthesizers for mobile devices," *arXiv preprint arXiv:1606.06061*, 2016.

[22] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Advances in neural information processing systems*, 2015, pp. 2773–2781.

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

[24] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5654–5658.

[25] L. Zhang, C. Yu, H. Lu, C. Weng, Y. Wu, X. Xie, Z. Li, and D. Yu, "Learning singing from speech," 2019.

[26] Y. Wu, S. Li, C. Yu, H. Lu, C. Weng, L. Zhang, and D. Yu, "Synthesising expressiveness in peking opera via duration informed attention network," 2019.

[27] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, D. Su, and D. Yu, "Durian: Duration informed attention network for multimodal synthesis," 2019.