The Sogou System for Blizzard Challenge 2020

Fanbo Meng, Ruimin Wang, Peng Fang, Shuangyuan Zou,
Wenjun Duan, Ming Zhou, Kai Liu, Wei Chen

Sogou, Beijing, P.R. China
{liukaios3228, mengfanbosi0935}@sogou-inc.com

Abstract

In this paper, we introduce the text-to-speech system from Sogou team submitted to Blizzard Challenge 2020. The goal of this year's challenge is to build a natural Mandarin Chinese speech synthesis system from the 10-hours corpus by a native Chinese male speaker. We will discuss the major modules of the submitted system: (1) the front-end module to analyze the pronunciation and prosody of text; (2) the FastSpeech-based sequence-to-sequence acoustic model to predict acoustic features; (3) the WaveRNN based neural vocoder to reconstruct waveforms. Evaluation results provided by the challenge organizer are also discussed.

Index Terms: speech synthesis, FastSpeech, neutral network vocoder

1. Introduction

The Blizzard Challenge has been held once a year since 2005, in order to better understand and compare research techniques in building corpus-based speech synthesizers on the same data. The basic challenge is to take the released speech database, build a synthetic voice from the data and synthesize a prescribed set of test sentences. The sentences from each synthesizer are then evaluated through listening tests.

The HMM-based statistical parametric speech synthesis (SPSS) method was first proposed and applied successfully in 1999 [1]. In this method, spectrum, pitch and duration are modeled simultaneously in a framework of decision trees and HMMs. Then, many techniques such as MGE-training [2] and phone duration modeling [3] were proposed to improve the framework. And post-filter methods, such as global variance (GV) [4], variance scaling (VS) [5] and modulation spectrum (MS) [6], were also helpful to improve the quality of synthesized speech.

Deep Neural Networks models have been applied successfully to SPSS in 2013 [7-9]. DNN-LSTM models have achieved greater performance in both the frontend text processing [10] and backend acoustic modeling [11]. Recently, a post-filter based on a generative adversarial network (GAN) was proposed to compensate for the differences between natural speech and speech synthesized by statistical parametric speech synthesis [12].

The traditional frameworks need an extra module to align the linguistic and acoustic features, and model the duration of linguistic units and acoustic features separately. The inaccurate align errors and the disagreement between the duration model and acoustic model may degrade the synthesis effect [13]. To address this problem, the attention-based sequence-to-sequence (seq2seq) models [14, 15] have been proposed and obtain superior performance [16-22].

After acoustic models, vocoder is used to generate waveforms from acoustic features predicted by acoustic models. The synthesized waveforms by traditional vocoders, such as STRAIGHT [23], WORLD [24] and so on, are found distortion compared to real speech. Many neural network vocoders are proposed to address the problem. Van den Oord et al. proposed WaveNet [25], a fully probabilistic and autoregressive deep neural network, with the predictive distribution for each audio sample conditioned on all previous ones. In Blizzard 2017, the WaveNet system had a good performance [26]. Deep Voice 1 and 3 [27, 28], the Parallel WaveNet [29], WaveRNN [30] and WaveGlow [31] have done more attempts and optimizations.

The paper is organized as follows. Section 2 introduces the details of the English task in Blizzard 202. Section 3 describes our system, including text analysis system, acoustic model and vocoder. Section 4 presents the results of the benchmark systems and all the participation. Finally, the conclusion is given in Section 5.

2. The task in Blizzard 2020

There are two tasks this year:

2020-MH1: The Mandarin data for the hub task 2020-MH1 is provided with text transcriptions only. About 10 hours of speech data of a single Chinese male in news style. The data is on several different channels.

2020-SS1: The Shanghainese data for the spoke task 2020-SS1 is provided with both text and phonemic transcriptions. Neither are time-aligned.

We participated in the task 2020-MH1 of Blizzard 2020.

3. Sogou speech synthesis system

As shown in Figure 1, our system consists of two parts, training and synthesis for task MH1. At the training phase, we first preprocess the original data, and then train our acoustic model and vocoder model based on this preprocessed data. At the synthesis phase, we first do the text analysis which converts the test manuscript to phoneme sequence with prosody boundary, and then the acoustic model and the vocoder model are used to get the final synthesized waveforms. We will introduce the details as follows.

3.1. Data processing

The data provided by the organizer are 4365 audio files at 48 kHz sampling rate and the corresponding texts. Considering the texts may disagree with the audios, firstly we check all the texts based on the audios. Secondly, we annotate the pinyin and the boundaries of the syllables, and the prosody boundaries in three types: syllable, prosody word and prosody phrase. Thirdly an HMM-based forced-alignment system is used to get the boundaries of the initials and the finals. Finally, we convert the

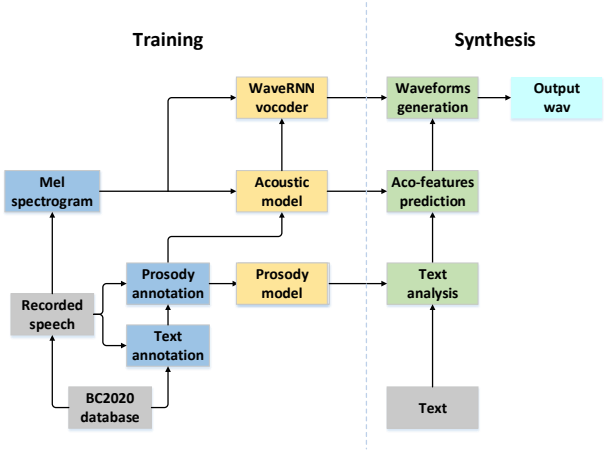

Figure 1: *The flowchart of Sogou TTS system*

sampling rate of the data to 32 kHz and reduce the energy of the waveforms at silence. Based on the processed data, we extract 160-dimension Mel spectrogram at 32 kHz using a 50ms frame length, 5ms frame hop and a Hann window function.

3.2. Text analysis system

The text analysis system mainly contains text normalization (TN), Chinese word segmentation (CWS), part-of-speech (POS) tagging, polyphone prediction and prosodic boundary prediction.

Text normalization is an essential procedure to normalize unreadable numbers, it contains a lot of ambiguity. For example, "11:23" can be read as time or the score of a game, "2020" can be read as a year or a number. We combine the advantages of a rule-based model and maximum entropy (ME) model to resolve such ambiguity problems and convert all symbolic chars into Chinese characters.

For word segmentation, we train a Bi-LSTM model with 140k external data to predict the word boundaries from the normalized texts. Besides, an extra user-defined dictionary is adopted to reduce the generation of the out-of-vocabulary word. Specifically, we abstract the word dictionary as features, and then merge these features and char information into the model for prediction.

For polyphone prediction or G2P, we use the phonetic information of words contained in the dictionary directly. And, we use the BERT-DNN model to do polyphone prediction by more than 300k external data. We use a pre-trained BERT[1] to extract char-level features, then merge the POS features and send them to the DNN network to obtain pinyin classification.

Finally, for the prosodic boundary prediction, we predict two-level boundaries using a BERT-LSTM model. We use BERT to get char-level features, and merge POS tags and several numeric features of the word context, then send them to the Bi-LSTM network. More than 200k annotated sentences are used to finetune the BERT-LSTM model.

3.3. Fastspeech-based acoustic model

We adopt FastSpeech [22] as our acoustic model to predict mel-spectrogram from the phoneme sequence. Besides, we also made some optimizations on basic FastSpeech model.

[1] https://github.com/google-research/bert

FastSpeech is composed of a multi-layer transformer encoder-decoder, and CNN duration model with a length regulator. Our first optimization is to use hierarchical variational autoencoders (VAE) [32], which includes sentence-level VAE and phoneme-level VAE. We use sentence-level VAE to model the channel difference information in the training data. Phoneme-level VAE is used to model the local prosodic information of sentences. In the synthesis stage, the sentence VAE is set to zero. We use an additional model to predict the value of phoneme-level VAE.

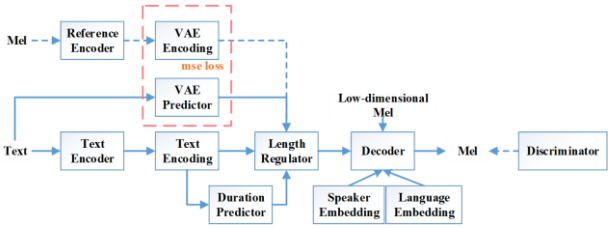

Figure 2: *Overview of the proposed acoustic model. Dashed lines are only valid during training.*

Secondly, we adopt a multi-band decoder. The motivation is that high-dimensional Mel spectrograms are more difficult to predict and are correlated with low-dimensional ones. So we use several different decoders to model different frequency band's mel-spectrogram. Each decoder additionally uses real (training stage) or predicted (synthesis stage) low-dimensional mel-spectrogram as input. For example, the decoder which predicts 101-160 dimensional Mel uses the real or predicted 1-100 dimensional Mel. All decoders synthesize in order from low frequency to high frequency in the synthesis stage. Our internal experiments show that this method can improve the accuracy of high-dimensional Mel.

Thirdly, we introduced GAN. After the basic model converges, we regard the basic acoustic model as a generator G, and introduce a discriminator D to distinguish real and predicted mel-spectrogram. We perform additional adversarial training to make the distribution of predicted Mel closer to the real distribution.

Specifically, our encoder, decoder, and the model to predict phoneme-level VAE all use 6-layer 384-node transformers. 4 decoders are used to model 160-dimensional mel-spectrogram extracted from the waveforms with 32KHz sample rate, respectively modeling 1-20 dimensions, 21-40 dimensions, 41-100 dimensions, 101-160 dimensions. The hierarchical VAE reference encoder uses 3 layers of 2-d convolution and a GRU, and the discriminator D uses 3 layers of 2-d convolution.

3.4. WaveRNN-based vocoder

First of all, we adopt WaveRNN [30] as our base vocoder structure. In the originally WaveRNN, a 16-bits speech signal is split into two 8-bits parts which represent the coarse fine of the sample. In our WaveRNN-like vocoder we do not build the model in two split sample parts, but directly predict a 16-bits wave with two GRU layers instead. Since training a 16-bits wave by categorical distribution would be prohibitively costly, we instead modeled the samples with the discretized mixture of logistics (MOL) distribution like the Parallel-WaveNet[29]. We found that the generated speech quality in MOL distribution is better than the original WaveRNN model.

Secondly, we introduce subband technology, which is widely used in speech signal processing, e.g., audio codec and speech enhancement. In 2018, the technology was applied to neural vocoder for the first time. Then, Chengzhu Yu et al. find an efficient way to combine subband and vocoder by employ the Pseudo Quadrature Mirror Filter Bank (PQMF) to multi-band processing [33][34]. Considering the efficiency and speech quality, we choose PQMF as our subband technology too. The result of our system display that subband processing can improve the generated speech quality.

Finally, we found that the generated speech quality degraded with the acoustic model generated mel-spectrogram as a vocoder input. This is caused by the mismatch between the real mel-spectrogram and the mel-spectrogram generated by the acoustic model. In order to fix this problem, we train a Tacotron-like acoustic model to process all the training mel-spectrogram by teaching force. Eventually, the speech quality gets a great promotion.

4. Results

In this year's challenge, there are 17 systems in total including natural speech (system A). Our submitted system is annotated as D. There are three criteria in the evaluation of sentences: Pinyin+Tone Error Rate (PTER), Mean opinion score (naturalness) and Similarity. And there are seven criteria in the evaluation of paragraphs: Overall, Pleasantness, Speech Pauses, Stress, Intonation, Emotion and Listening effort. We will discuss the details as follows.

4.1. Evaluation of sentences

Figure 3 shows the PTER of sentences. As expected, the original natural speech achieves the best score of 0.074. System L achieves the best score of 0.086 among all the submitted systems. The PTER of our system is 0.096. In doing the listening test, we found one system split the test manuscript by words, and inserted pauses between words, e.g., "农业/辅导/生词/检查/对话/和尚". We think this method may improve the PTER.

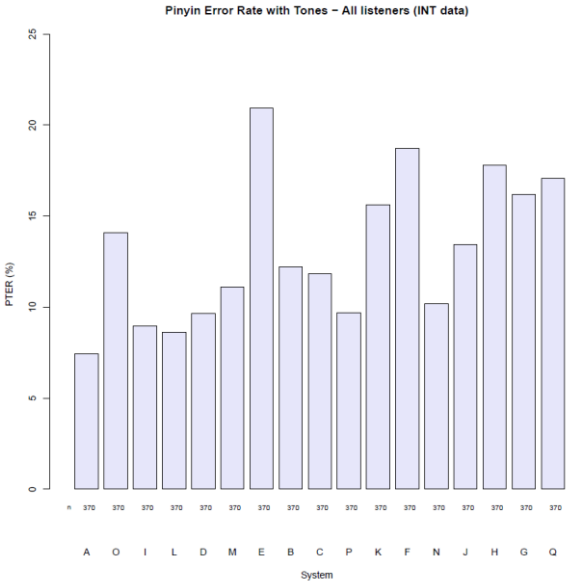

Figure 3: *the PTER of the evaluation of sentences.*

Figure 4 shows the MOS of the naturalness of sentences, multiple systems' median scores are relatively close. The original natural speech achieves the highest score of 4.7. System I and system O achieve the highest score of 4.2 among all the submitted systems. The MOS of our system is 3.9. We think it may be helpful that providing more information to the FastSpeech acoustic model, such as stress.

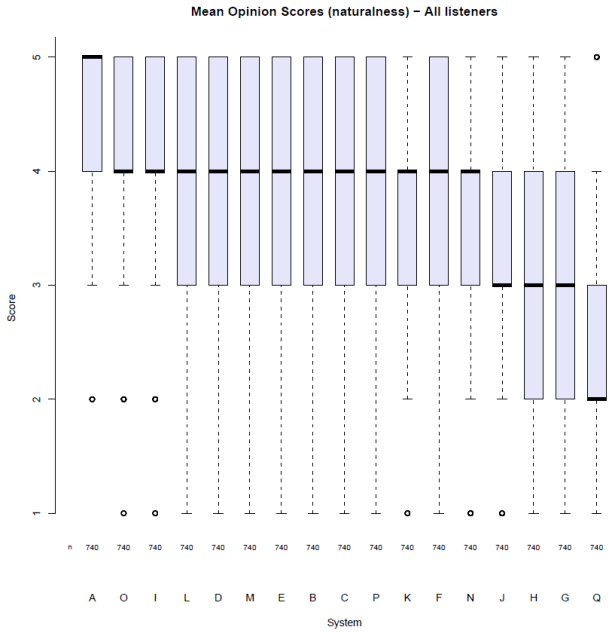

Figure 4: *the MOS of naturalness of sentences.*

Figure 5 shows the MOS of similarity of sentences. It can be seen from the median score that the effects of multiple systems are relatively close. The original natural speech achieves the highest score of 4.4. The system I achieves the highest score of 4.2 among all the submitted systems. The MOS of our system is 3.9. Firstly there is still a big gap between the predicted acoustic mel- spectrogram and the real ones, and we use GAN to improve the quality of synthesized speech, which may disagree with the perception of humans and hurt the similarity.

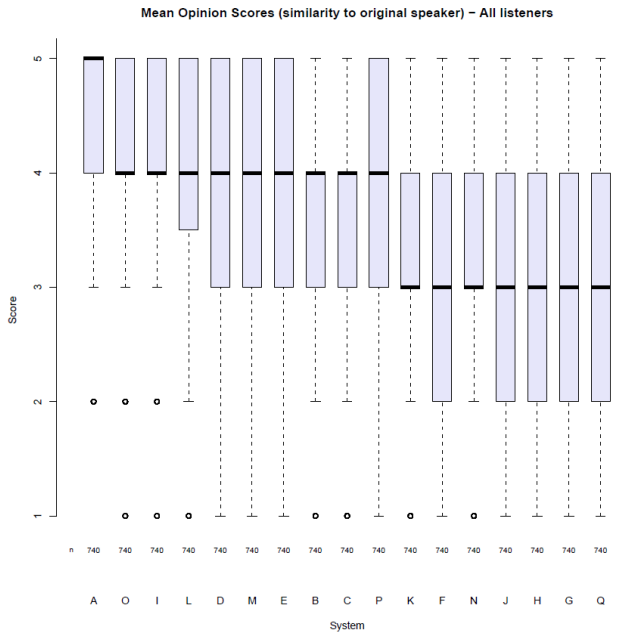

Figure 5: *the MOS of similarity of sentences.*

4.2. Evaluation of paragraphs

Figure 6 shows the MOS of overall impression of paragraphs. The distance between the score of the original natural speech and the score of the submitted systems is bigger than the distance of the scores of sentences. While the distance between the results of our system and the results of the best submitted system is not as big as the distance of the scores of sentences. This indicates the submitted system still can't make full use of paragraph information. There is still a long way to synthesize paragraphs.

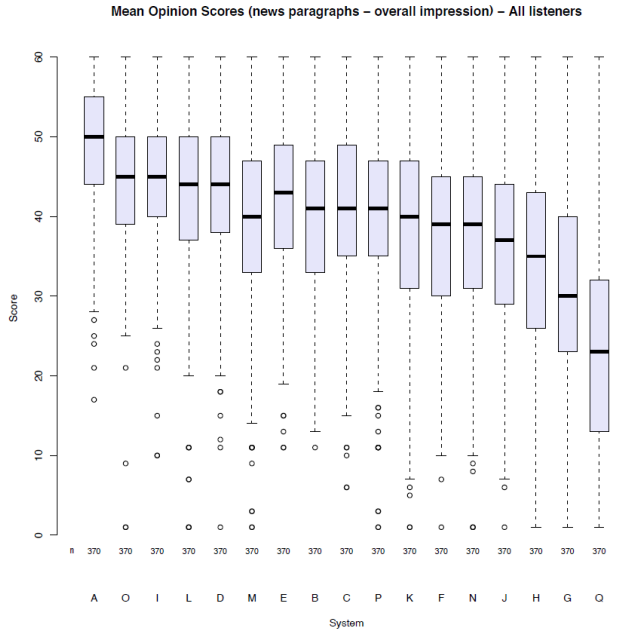

Figure 6: *the MOS of overall impression of paragraphs.*

5. Conclusions

This paper presents the details of our submitted system and summarizes the results in Blizzard Challenge 2020. In our system, improved FastSpeech and WaveRNN vocoder are used in order to achieve natural and high-fidelity speech.

6. References

[1] Yoshimura T, Tokuda K, Masuko T, et al. "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Sixth European Conference on Speech Communication and Technology, 1999.

[2] Wu, Yi-Jian, and Ren-Hua Wang, "Minimum generation error training for HMM-based speech synthesis," Acoustics, Speech and Signal Processing, 2006.

[3] Yi-Jian Wu, "Research on HMM-based Speech Synthesis," Ph.D Thesis, University of Science and Technology of China, 2006. [in Chinese]

[4] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in IEICE TRANSACTIONS on Information and Systems, 2007, 90(5): 816-824.

[5] Siln H, Helander E, Nurminen J, et al., "Ways to implement global variance in statistical speech synthesis," Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[6] Takamichi S, Toda T, Neubig G, et al., "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," in Proc. ICASSP 2014, 2014, pp. 290C294.

[7] Ze, Heiga, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 7962-7966.

[8] Ling Z H, Kang S Y, Zen H, et al., "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends" IEEE Signal Processing Magazine, 2015, 32(3): 35-52.

[9] Wu Z, Valentini-Botinhao C,Watts O, et al., "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis" Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015: 4460-4464.

[10] Ding C, Xie L, Yan J, et al., "Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features" in Automatic Speech Recognition and Understanding, 2015, pp. 98C102.

[11] Zen, Heiga, and Ha?im Sak., "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis." Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015.

[12] Kaneko T, Kameoka H, Hojo N, et al., "Generative adversarial network-based postfilter for statistical parametric speech synthesis" Proc. ICASSP, 2017, 2017: 4910-4914. S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.

[13] M. Li, Z. Wu, and L. Xie, "On the impact of phoneme alignment in dnn-based speech synthesis," in 9th ISCA Speech Synthesis Workshop, Sunnyvale (USA), 2016, pp. 196–201.

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

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[16] W.Wang, S. Xu, and B. Xu, "First Step Towards End-to-End Parametric TTS Synthesis: Generating Spectral Parameters with

Neural Attention," in Proc. INTERSPEECH, 2016, pp. 2243–2247.
[17] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in Proc. ICLR workshop, 2017.
[18] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang et al., "Tacotron: Towards end-to-end speech synthesis," in Proc. INTERSPEECH, 2017, pp. 4006–4010.
[19] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," arXiv preprint arXiv:1710.07654, 2017.
[20] 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.
[21] S. Yang, H. Lu, S. Kang, L. Xie, and D. Yu, "Enhancing hybrid self-attention structure with relative-position-aware bias for speech synthesis," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6910–6914.
[22] 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 e-prints, arXiv:1905.09263, arXiv:1905.09263, 2019.
[23] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds", Speech Commun., vol. 27, no. 34, pp. 187-207, 1999.
[24] 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. E99-D, no. 7, pp. 1877-1884, 2016.
[25] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, " Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.
[26] Hu Y J, Ding C, Liu L J, et al., "The USTC system for blizzard challenge 2017" Proc. Blizzard Challenge Workshop, 2017.
[27] Arik S O, Chrzanowski M, Coates A, et al., "Deep voice: Real-time neural text-to-speech" arXiv preprint arXiv:1702.07825, 2017.
[28] PingW, Peng K, Gibiansky A, et al., "Deep voice 3: 2000-speaker neural text-to-speech" arXiv preprint arXiv:1710.07654, 2017.
[29] Oord A, Li Y, Babuschkin I, et al., "Parallel WaveNet: Fast highfidelity speech synthesis" arXiv preprint arXiv:1711.10433, 2017.
[30] 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," arXiv:1802.08435, 2018.
[31] R. Prenger, R. Valle and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis", Proc. ICASSP, May 2019.
[32] Kingma, D.P.; Welling, M. An Introduction to Variational Autoencoders. Found. Trend. Mach. Learn. 2019, 12, 307–392.
[33] 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), pp. 304–311, IEEE, 2018.
[34] 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 multimodal synthesis," arXiv preprint arXiv:1909.01700, 2019.