American Journal of Neural Networks and Applications
2025, Vol. 11, No. 1, pp. 28-34
https://doi.org/10.11648/j.ajnna.20251101.13

Research Article

A Fast Acoustic Model Based on Multi-Scale Feature Fusion Module for Text-To-Speech Synthesis

Jin-Hyok Song, Song-Chol Jong, Thae-Myong Kim, Guk-Chol Kim, Hakho Hong*

Institute of Mathematics, State Academy of Sciences, Pyongyang, Democratic People's Republic of Korea

Abstract

In the end-to-end Text-to-speech synthesis, the ability of acoustic model has important effects on the quality of the speech generated. In the acoustic model, the encoder and decoder are critical components, and usually Transformer is used. The previous works have a lack of ability to model the essential features of speech signal, as they model fixed length of features. There is also limitation of slow inference speed of the acoustic model due to the characteristics of the transformer including the high-computational multi-head self-attention layer. This limits the application of TTS in the low performance of devices such as embedded devices or mobile phones. In this paper, we propose a novel acoustic model to model the different length of features and improve the speed of generating synthetic speech and naturalness as compared to the conventional Transformer structure. Through the experiment, we confirmed that the proposed method improves the naturalness of synthetic speech and operation speed in the low performance of devices.

Keywords

Text-To-Speech Synthesis, Encoder, Vocoder, Transformer

1. Introduction

The Text-to-speech (TTS) synthesis technology has been used to synthesis the natural speech for the human to understand from the given text [15, 10].

The neural network based TTS models have made much improvement in the speech quality and naturalness compared to traditional ones, and recently produced a speech which is very similar to the human speech. The non-autoregressive TTS models are accelerating their application in low performance devices such as mobile phones and embedded devices, while making inference faster than autoregressive models. Much of the progress in TTS has led to improved synthetic quality, but there are still many problems that require improvement compared to human speech [2, 6].

Typical TTS systems consist of 3 main components: text analysis, acoustic model, and vocoder. First, the input text is normalized and converted to phoneme level, syllable level, and word level through the various levels of linguistic features, i.e., the text analysis module. Then, the linguistic features are converted by acoustic models into the intermediate acoustic representations such as mel-spectrograms. Finally, the acoustic representation is converted to speech by a vocoder. Concated speech synthesis and statistical parametric speech synthesis, together with HMM/DNN/LSTM-based models, have been the most popular methods in the past [16, 20]. In recent years, there has been a wide study of end-to-end TTS. Tacotron 2 [14] improved the speech quality

*Corresponding author: hhhong@star-co.net.kp (Hakho Hong)

Received: 17 February 2025; **Accepted:** 26 March 2025; **Published:** 22 April 2025

Copyright: © The Author(s), 2025. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

using attention-based encoder-decoder acoustic model and vocoder based on the WaveNet [11] model structure. However, the robustness problems of utterance errors such as errors, omissions, and repetitions, and the slow speed in training and inferencing are problematic. A method to improve training speed and speech quality by applying Transformer [17] was proposed, which attempts to improve the robustness of speech synthesis by using linguistic features and phoneme duration. However, there is still a drawback of slow inference time due to the autoregressive structure.

FastSpeech [13] is a non-autoregressive TTS model that uses a duration predictor in addition to a parallel operable feed-forward converter-based encoder and decoder, which is much faster and more stable to utterance errors than autoregressive models such as Tacotron 2.

In DelightfulTTS [7], an improved Conformer [3] module was applied to the acoustic model for better modeling the global and local dependencies in the mel-spectrogram. Conformer is a model that combines convolutional neural networks (CNN) and Transformer. It was used in the initial end-to-end automatic speech recognition (ASR), which achieved better accuracy with fewer parameters than previous works on several public ASR datasets. Conformer suggests a new combination of self-attention and convolution, where self-attention learns global interactions and the convolution structure effectively captures local correlations.

The existing TTS models have large memory consumption and slow inference speed due to large model size and inefficient structure. For example, acoustic models of non-autoregressive structures, including FastSpeech, FastSpeech 2 [12], LightSpeech [9], DelightfulTTS, are mostly based on transformer or variants such as Transformer or Conformer [1, 4, 5, 7, 8, 12, 18]. The transformers all contain a multi-head self-attention (MHSA) layer, although their structure is slightly different. MHSA was first used in machine translation, which showed good performance in seq2seq modeling, and is capable of parallel operations in structure. However, MHSA incurs a higher computational cost as the sequence length increases. This limits the application of TTS in embedded devices or mobile phones.

In this paper, we propose an acoustic model to better model the characteristics of speech signals while reducing the computational cost in transformer variants including MHSA.

2. Proposed Method

First, we briefly review the previous acoustic models relevant to the proposed model and present the structure of our acoustic model.

2.1. FastSpeech 2

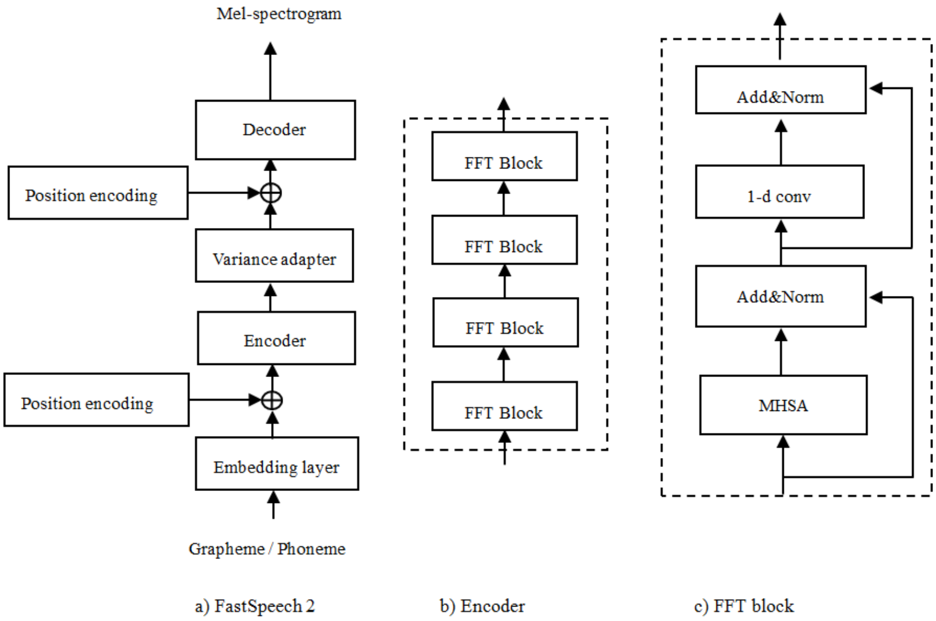

Figure 1. *The overall architechture for FastSpeech 2 and sub blocks.*

The FastSpeech 2 acoustic model consists of a grapheme/phoneme embedding layer, encoder, variance adapter, and decoder. Encoder and decoder have the same structure, which consists of feedforward transformer (FFT) blocks.

Figure 1 shows FastSpeech 2 acoustic model, encoder and FFT structure.

In the encoder and decoder of FastSpeech 2, the FFTs are cascade connected, which consists of MHSA, layer normalization, and convolutional layer.

2.2. DelightfulTTS

DelightfulTTS is structurally similar to FastSpeech 2, but instead of FFT block, it uses Conformer block to model the global and local dependences of the speech signal (Figure 2a). And instead of existing Conformer block, an improved Conformer block that swapped the order of self-attention and convolution module and replaced the linear layer with the convolution layer in the feedforward layer was used. (Figure 2c).

DelightfulTTS model and structure of the improved Conformer are shown in Figure 2. The basic structure is the same as FastSpeech 2, and only the difference is to use an improved Conformer instead of FFT block in encoder.

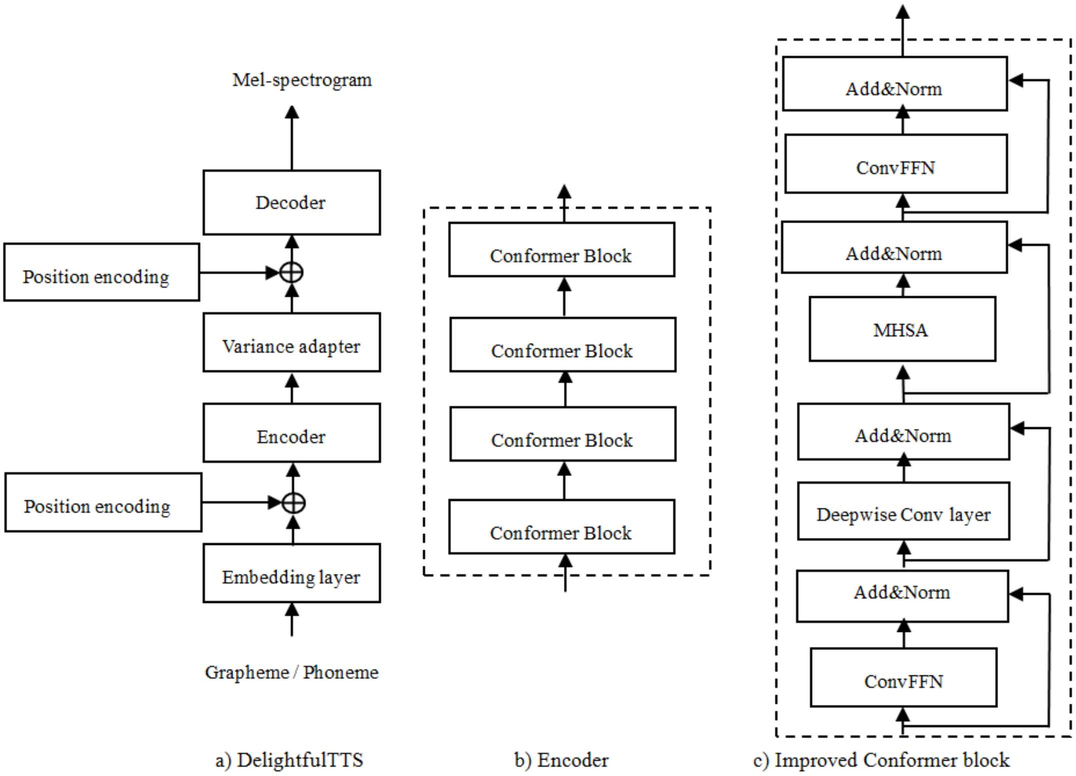

Figure 2. The architecture for DelightfulTTS and sub blocks.

2.3. Proposed Model

In the previous acoustic models which is non-autoregressive structures including FastSpeech 2 or DelightfulTTS, a variance adapter to model the variance information such as duration, pitch and energy was applied to solve the problem of one-to-many mapping of TTS. Here, we will modify only the encoder and decoder structures, while retaining the variance adapter used in the previous models.

2.3.1. An improved Encoder-Decoder

FastSpeech 2 contains 4 FFT blocks in the encoder and decoder, and DelightfulTTS used a Conformer block instead of the FFT block. However, such structure can only deal with fixed feature sequences for the input.

In the tasks with large length differences between input and output sequences, such as TTS and ASR, it would be effective to analyze the different length of input features, select the essential features and then generate the output sequence. This is because in the fixed length of feature, it is only available to model features in the fixed regions, but at different length of features, it is also available to model global and local contexts in the different regions. Hence, we propose an improved structure to process the different length of features (Figure 3). Transformer and Conformer can be used as a basic block in Figure 3.

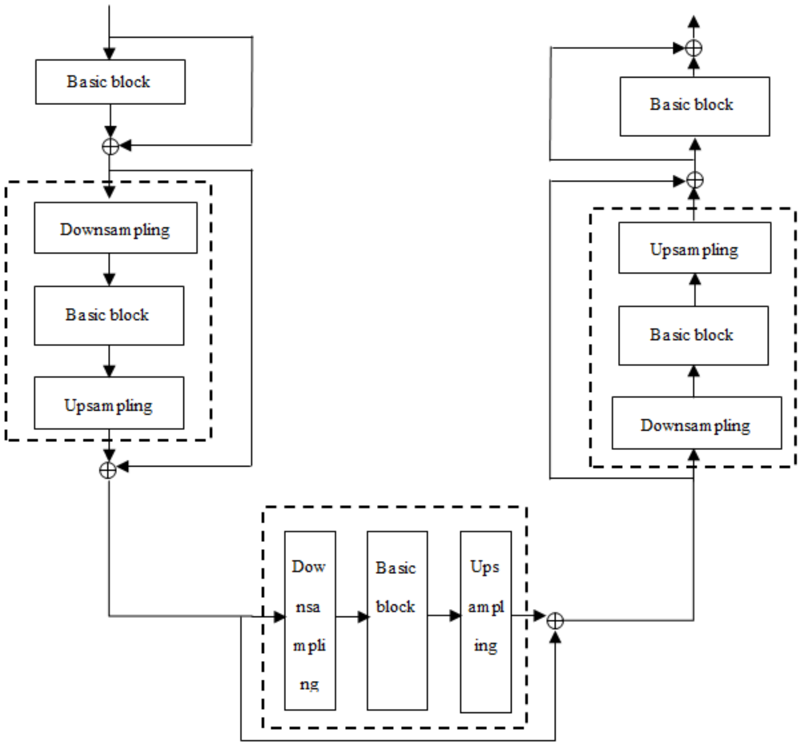

Figure 3. Proposed encoder.

The encoder and decoder are structurally similar but only differ in hyperparameters. The detailed description of the hyperparameters is given in the experimental section.

Downsampling and upsampling in the dashed square play a role in decreasing or increasing the sequence length, and its implementation is done in a very simple way. For example, when the sequence length is reduced by half, 2 subsequent features are averaged. When the sequence length is increased by double, each frame is repeated twice and linked. Note that the amount of operation is very small because the upsampling and downsampling is performed by simple method.

Through the downsampling module, the basic blocks receive as input features that is reduced in length by downsampling rate. As the length of input feature sequence decreases, the amount of operation in the basic block will be decreased. By applying different downsampling rates on the dashed square, the basic blocks can process the different length of features. We use the improved Conformer as a basic block.

2.3.2. The Overall Architecture of Proposed TTS

The overall architecture of the proposed TTS system follows as Figure 4. Given a grapheme or phoneme sequence, it is entered into the encoder through the embedding layer. Positional encoding is used to consider the order of sequences. The encoder receives a Positional encoding input so that the order can be considered.

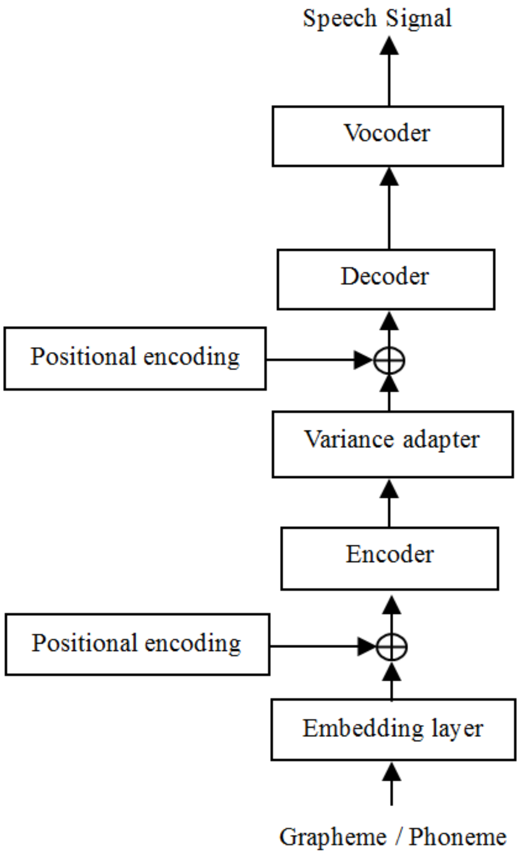

Figure 4. Proposed TTS system.

The output feature through the variance adapter that has a function to model the variance information including energy, pitch, and duration is entered to the decoder. The decoder generates mel-spectrogram. The decoder has the same structure as encoder, but the hyper parameter is different. The last layer of the decoder is the feedforward layer, whose output dimension is equal to the mel-spectrogram dimension. And the mel-spectrogram is converted to speech via a vocoder.

3. Experiment

3.1. Experiment Condition

Training dataset

We train and evaluate proposed method in the dataset for Korean TTS. The dataset consists of approximately 20,010 audio clips (about 25 hours) and corresponding text transcripts. The sampling frequency is 22.05 kHz, the quantization bit rate is 16, and the single-channel speech. The dataset is divided into 3 parts: training subset (12,900 data), test subset (100 data), and validation subset (100 data).

Text sequence is converted into phonetic sequence.

The speech data are converted into 80-dimensional mel-spectrograms using short-time fourier transform (STFT). When performing STFT, we used a Hanning window, and set frame size and hop size to 1024 and 256 respectively.

Training and inference environment

The proposed model is trained on a single NVIDIA Tesla P100 GPU computer. Adam optimizer is used.

The inference is performed in mobile phone with AArch64 (1.8GHz) processor, and for the vocoder, MB-MelGAN [19] model is used. The MB-MelGAN model is trained with the speech dataset used to train the acoustic model. The model structure is taken the same as in the MB-MelGAN model.

3.2. Model Structure

Table 1. Hyperparameters of our model.

Hyperparameter	Value
Phoneme embedding size	256
Number of layers in encoder	5
Number of attention heads in encoder	2, 2, 2, 2, 2
Attention dimension	96, 96, 96, 96, 96
Encoder dimension	128, 128, 128, 128 128
Downsampling rates in encoder	1, 2, 4, 2, 1
Filter sizes	31, 31, 31, 31, 31
Nummber of layers in decoder	4
Number of attention heads in decoder	2, 2, 2, 2
Downsampling rates in decoder	1, 2, 4, 2
Dropout rate	0.1

Proposed model consists of 5 basic blocks as shown in Figure 3.

Table 1 shows the hyperparameters of our model. Downsampling rates in the encoder are 1, 2, 4, 2, 1 which means that downsampling and upsampling are not applied in the first and last blocks. Likewise downsampling rates of 1 in decoder means that downsampling and upsampling are not applied.

The hyperparameters of the variance adapter such as duration, pitch and energe are same as the values used in FastSpeech 2.

3.3. Result

3.3.1. The Evaluation of the Speech Quality

Table 2 gives the MOS for the speech generated by the previous methods, FastSpeech 2 and DelightfulTTS, and the proposed method. “GT” means ground-truth audio.

Table 2. MOS evaluation between the previous methods and proposed method.

Method	MOS
GT	4.46
FastSpeech 2 Mel+MB-MelGAN	4.18 ±0.08
DelightfulTTS Mel+MB-MelGAN	4.31 ±0.05
Proposed model Mel+MB-MelGAN	4.36 ±0.06

Compared with the MOS 4.46 of the recorded speech, it is shown that the speech generated by the proposed model is almost similar to human speech.

The results of the table show that the proposed method has a higher performance compared to the previous method.

3.3.2. The Evaluation of the Model Size and Inference Speed

Table 3. Inference speed of the previous and proposed methods.

Method	RTF	Model size (M)
FastSpeech 2	1.62	27.4
DelightfulTTS	1.89	29.5

Method	RTF	Model size (M)
Proposed method	0.85	18.2

The real-time factor (RTF) is the ratio of the time taken to generate and the duration of the generated speech.

$$RTF = \frac{T_{gen}}{T_{len}} \quad (1)$$

You can see that the model size is much smaller than the previous two methods by 18.2 M. Here 1M means 10^6 parameters.

The experiment results in the Table 2 and Table 3 show that model size is smaller and operation speed is faster of the proposed method than previous methods.

3.4. Experimental Analysis

The proposed encoder structure has a downsampling and upsampling module to process the different length of features, which differs from previous encoders.

In this experiment, we evaluate the performance of the CMOS and RTF when downsampling and upsampling are eliminated.

Table 4. CMOS evaluation of the proposed methods with and without downsampling and upsampling rates.

Method	CMOS	RTF
Proposed method (Applied downsampling and upsampling)	4.36 ±0.06	0.85
Proposed method (Eliminated downsampling and upsampliing)	4.26 ±0.09	1.21

The absence of downsampling and upsampling means that downsampling rate in encoder and downsampling rate in decoder are all 1.

Experimental results show the advantages of method to process the different length of features. When downsampling and upsampling are applied, the MOS is higher than when downsampling is not applied, while the real-time factor is reduced by 0.7 times. This is due to the reduced sequence length of input entering the basic block when downsampling is performed.

The experimental results show that applying downsampling and upsampling gives good results.

If you use an improved Conformer block as a basic block, and downsampling or upsampling module is not used, the model is similar to DelightfulTTS. But if you compare the MOSs in the Table 2 and Table 4, there is a little differences which is related to the implementation of the programming.

4. Conclusion

In this paper, we propose an encoder that can process the different length of features instead of fixed length of feature to improve the performance of acoustic model in TTS. Through experiments, we confirmed that the proposed method gives an improvement in both audio quality and operation speed compared to previous acoustic models using the traditional encoder.

The proposed acoustic model is not unique for Korean, so it can be easily applied to other languages, including English and Chinese. In the future, we will investigate the TTS with lighter model and human-like level of speech quality.

Abbreviations

ASR	Automatic Speech Recognition
CNN	Convolutional Neural Network
FFT	Feed Forward Transformer
MHSA	Multi-Head Self-Attention
MOS	Mean Opinion Score
RTF	Real Time Factor
STFT	Short Time Fourier Transform
TTS	Text-to-Speech

Author Contributions

Jin-Hyok Song: Conceptualization, Investigation, Project administration, Supervision, Writing – original draft, Writing – review & editing

Song-Chol Jong: Formal Analysis, Investigation, Validation, Visualization, Writing – review & editing

Thae-Myong Kim: Data curation, Formal Analysis, Investigation, Resources, Software, Validation, Visualization, Writing – review & editing

Guk-Chol Kim: Methodology, Resources, Software, Validation, Visualization

Hakho Hong: Data curation, Formal Analysis, Investigation, Validation, Writing – review & editing

Funding

The authors declare that no fund and no support were received during the preparation of the research paper.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Dai, Z., Yang. Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

[2] Fan, Y., Qian, Y., Xie, F.-L., & Soong, F. K. (2014). Tts synthesis with bidirectional lstm based recurrent neural networks. *In Fifteenth annual conference of the international speech communication association.*

[3] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. (2020). Conformer: Convolution augmented transformer for speech recognition. In *Proceedings of Interspeech*.

[4] Han, C.-J., Ri, U.-C., Mun, S.-I., Jang, K.-S., Han, S.-Y. (2022). An end-to-end TTS model with pronunciation predictor, *International Journal of Speech Technology*. https://doi.org/10.1007/s10772-022-10008-7

[5] Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019) Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, (pp. 6706– 6713).

[6] Li, N., Liu, Y., Wu, Y., Liu, S., Zhao, S., & Liu, M. (2020). Robutrans: A robust transformer-based text-to-speech model. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, (pp. 8228–8235).

[7] Liu. Y., Xu. Z., Wang. G., Chen. K., Li. B., Tan. X., Li. J., He. L., & Zhao. S. (2021). Delightfultts: The microsoft speech synthesis system for blizzard challenge 2021. In *Proceedings of the blizzard challenge 2021*.

[8] Liu, Y., Xue, R., He, L., Tan, X., & Zhao, S. (2022). DelightfulTTS 2: End-to-end speech synthesis with adversarial vector-quantized auto-encoders. In *Proceedings of Interspeech*.

[9] Luo, R., Tan, X., Wang, R., Qin, T., Li, J., Zhao, S., Chen, E., & Liu, T. Y. (2021). Lightspeech: Lightweight and fast text to speech with neural architecture search. In *Proceedings of international conference on acoustics, speech and signal processing (ICASSP).*

[10] Mu, Z., Yang, X., & Dong, Y. (2021). Review of end-to-end speech synthesis technology based on deep learning. *International Conference on Secure Cyber Computing and Communications (ICSCCC).*

[11] Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. In 9th *ISCA speech synthesis workshop* (pp. 125–125).

[12] Ren, Y., Hu C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2021). Fastspeech 2: Fast and high-quality end-to-end text to speech. In *Proceedings of international conference on learning representations (ICLR 2021)*.

[13] Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). Fastspeech: Fast, robust and controllable text to speech. In *NeurIPS*.

[14] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proceedings of international conference on acoustics, speech and signal processing (ICASSP 2018)*. (pp. 4779– 4783).

[15] Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. arXiv preprint arXiv: 2106.15561.

[16] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for hmm-based speech synthesis. In *Proceedings of international conference on acoustics, speech and signal processing (ICASSP 2000)* (pp. 1315–1318).

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

[18] Xu, Z., Zhang, S., Wang, X., Zhang, J., Wei, W., He, L., Zhao, S. (2023). MuLanTTS: The Microsoft Speech Synthesis System for Blizzard Challenge 2023. In *Proceedings of the blizzard challenge 2023*.

[19] Yang, G., Yang, S., Liu, K., Fang, P., Chen, W., & Xie, L. (2020). Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *Proceedings of Spoken Language Technology Workshop (SLT).*

[20] Ze, H., Senior, A., & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proceedings of international conference on acoustics, speech and signal processing (ICASSP 2013)* (pp. 7962–7966).