The TAL Speech Synthesis System for Blizzard Challenge 2021

Shaotong Guo[1], *Shuaiting Chen*[1], *Dao Zhou*[1], *Gang He*[1], *Changbin Chen*[1]

[1]TAL Education Group, Beijing, China

{guoshaotong,chenshuaiting1,zhoudao1,hegang1,chenchangbin}@tal.com

Abstract

This paper introduces the TAL speech synthesis system for Blizzard Challenge 2021 which aims to synthesize voice as similar as the provided target speaker. We built a Spanish speech synthesis system based on the pre-trained BERT model, GST and HiFi-GAN for task 2021-SH1. First, we use a modified open source Spanish front-end to generate Spanish phoneme sequences from the input Spanish text. Then, we constructed a modified GST model which condition the encoder on linguistic features. The acoustic model is trained on two speakers, and then fine-tune on the target speaker from provided corpus. To speed up the synthesis process and maintain the speech quality, we use HiFi-GAN, an efficient and high fidelity GAN-based vocoder, to synthesize mel-spectrogram into speech waveform. The evaluation results shows that our system performs well especially in the word error rates evaluation.

Index Terms: Blizzard Challenge 2021, Speech Synthesis, BERT, HiFi-GAN

1. Introduction

Text-to-speech (TTS) is a kind of technology that transform text into speech signal. The goal of the TTS system is to generate natural and intelligent speech like humans. The conventional statistical parametric speech synthesis (SPSS) system and concatenate based synthesis system can generate acceptable speech. However, these systems need expert knowledge or huge amount of speech segment corpus. In recent years, deep learning based TTS system can generate high quality speech, such as Tacotron [1], Tacotron2 [2], Deep Voice series [3, 4, 5] and FastSpeech series [6, 7]. Correspondingly, neural network based vocoder, such as WaveRNN [8], WaveNet [9] and HiFi-GAN [10], are also perform better than digital signal processing based vocoder, such as WORLD [11], STRAIGHT [12] and Griffin-Lim [13].

For most TTS system, the input is phoneme or character sequence, however, the model can not learn enough semantic information from these fine-grained features. It has been proved that the linguistic information derive from the language model, such as text embedding and word embedding, is helpful for TTS system on naturalness [14, 15] and prosody [16, 17]. The pre-trained BERT model [18] is a state-of-the-art language model and has been used in most of natural language processing (NLP) tasks [19, 20] and some TTS system [21]. The BERT model is a Transformer [22] based language model trained on huge amount of unlabeled text. The deep structure learn the internal relationship between words in the same sentence, the output is a deep representation of context.

In this paper, we use the linguistic features sequence derived from the pre-trained BERT model as the condition on GST model [23] to improved the naturalness and intelligence of synthesized speech. Firstly, we use an open-sourced Spanish front-end from a unit-selection based TTS system[1] to generate Spanish phoneme sequence. Secondly, a pre-trained multi-lingual BERT model is used to generate sub-word level BERT embedding, a post processing method of up-sampling sub-word level embedding into longer length according to phoneme sequences is proposed. The up-sampled BERT embedding were conditioned on the encoder of the model. Lastly, HiFi-GAN, an efficient and high fidelity GAN-based vocoder, is used to convert mel-spectrogram into speech signal.

In the following sections, the more details of data processing, model structure and evaluation results will be introduced.

2. The task in Blizzard Challenge 2021

In Blizzard Challenge 2021, the task is to build a Spanish speech synthesis system and synthesize voice as similar as a target European Spanish female speaker. The provided data is about 9.7 hours and consist of 4920 audio files of which type is 'wav', sample rate of audio files is 48kHz. All text transcripts are also provided. Blizzard Challenge 2021 have two optional tasks, we choose the Hub task, 2021-SH1, which request participants build a voice from the provided speaker and synthesise text containing only Spanish words.

3. Proposed method

Our model consists of three parts, a Spanish front-end, a GST-based acoustic model with a linguistic features processing module and a neural vocoder. We will introduce the data processing and the structure of each module in the following parts.

3.1. Data processing

The data set provided by the committee contains 4920 audio files, 48kHz sample rate and 16-bit format. Total duration of these data is about 9.7 hours, we remove some audios and use about 4.06 hours for our system training. Corresponding texts transcripts is provided. We use about 95 hours external for the system training, including 9.7 hours data from M-AILABS[2] and 85 hours data from AISHELL-3[3]. The subset of AISHELL3 used for vocoder training, the subset of M-AILABS and processed provided data used for both acoustic model training and vocoder training.

Each line of the provided text transcriptions is correspond to one audio file. Because all text were manually checked by the committee, we did not perform any further text processing. All text were directly convert into phoneme sequences by our front-end and without any auxiliary manually annotate.

For audios provided by committee, the processing is as bellow. First, we remove audios that containing English words, because 2021-SH1 task only needs synthesize Spanish words.

[1]https://github.com/pilarOG/unit_selection_tts

[2]https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/

[3]http://www.aishelltech.com/aishell_3

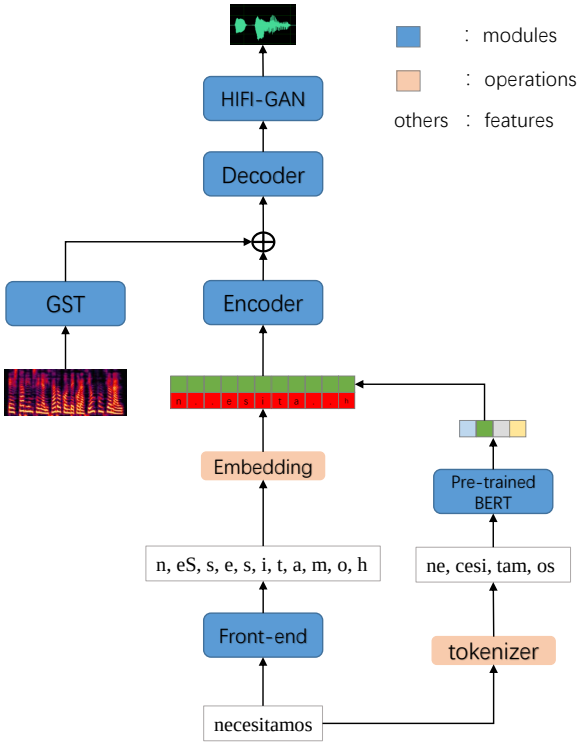

Figure 1: *The structure of our system. One word is converted into phoneme sequence by front-end and converted into token sequence by tokenizer of the BERT model respectively. We use the embedding of the middle sub-word to represent the whole word and concatenate with phoneme embedding after up-sampling.*

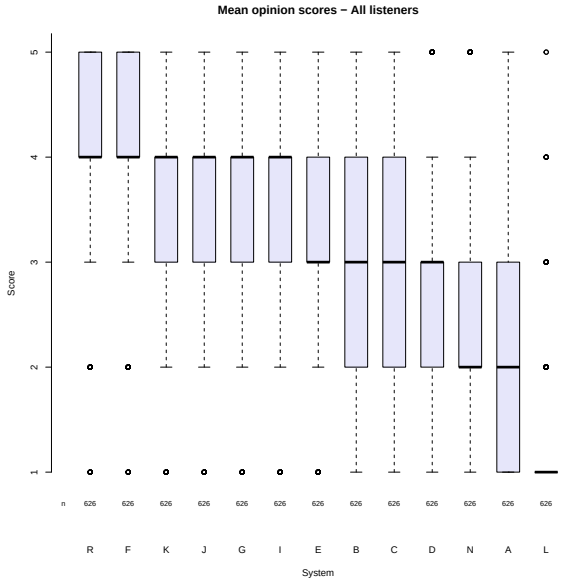

Figure 2: *Boxplot of naturalness scores of each submitted system of all listeners (2021-SH1).*

Meanwhile, as non-Spanish speakers, we can not completely ensure whether the phoneme sequence generated by our front-end is correct. Then, we manually pruned silence at the start and the end of audios, and replace the with fixed length silence about 300ms respectively. Because most audio contains more than 1 second of silence and that is too long for TTS system training. In order to avoid the influence of the noise in audio on model training, we also remove audios that contains too much noise. Finally, we down-sample remaining processed 3742 (about 4.06 hours) audio files to 16kHz as training data. The training data were de-noised by an internal tool, and extracted mel-spectrogram for training.

Research [24] have shown that at least 10 hours of data is enough to train a neural TTS system. we use a Spanish female subset of M-AILABS as external data set (about 9.7 hours), and combine with about 4.06 hours processed provided audios to train the acoustic model. We process external data follows the processing mentioned above.

For the vocoder, we use about 99 hours data totally. We use a subset of AISHELL-3, about 85 hours, to extract mel-spectrogram. The audios for training acoustic model (about 13.76 hours) also been used for vocoder training.

3.2. Front-end

The rule based front-end in our system is mainly to convert the input text into a unified phonemes representation which as the input for subsequent speech synthesis. The front-end module in our system is mainly composed of the three parts:

Firstly, convert all uppercase letters into lowercase letters in order to facilitate the subsequent text conversion work. Then, all Spanish letters is divided into 11 categories which will be used in the conversion rules.

Secondly, each letter in the input Spanish sentence form a window with the previous letter and the next letter, these three letters is translated into phoneme according to their categories and the conversion rules. For instance, if current letter is 's', and the combination of these three letters belongs to 'fricatives', a 'h' will be insert into phoneme sequence.

Thirdly, the stress mark will be added into phoneme sequence generated from previous step. In Spanish, syllable stress directly affects the meaning of words, and our front end is marked according to stress rules. All phonemes in current sequence will be split into 6 syllable categories which are 'V', 'VV', 'IUT', 'C', 'CC' and 'H'. For every phoneme in this sequence, if the adjacent phonemes belongs to a Specific combination of rules, these phonemes will regarded as a syllable. For instance, if the combination of categories of current phoneme, previous one and the next one is 'V-C-V', these phoneme can be regarded as one syllable. Then the syllable will be translated into stressed form, depending on the syllable category and the position in this word. Stressed syllables will decompose into phoneme sequences. Finally, the stressed phoneme sequence will output and used as the input of acoustic model. Phonemes of different words are separated by a space. Phonemes in the same word have no separators.

3.3. Acoustic model conditioned on BERT

The acoustic model of our system is based on GST, but the encoder is conditioned on up-sampled linguistic features. As shown in Figure 1, the whole acoustic model consists of the linguistic features process module, the GST module, the encoder and the attention based decoder.

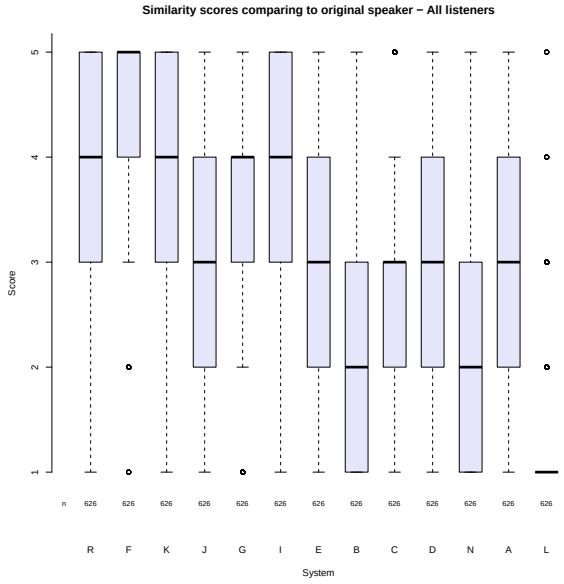

Figure 3: *Boxplot of similarity scores of each submitted system of all listeners (2021-SH1).*

The linguistic features process module is based on a pre-trained multi-lingual BERT model[4], we use a Spanish vocab to generate Spanish sub-word embedding. The input of pre-trained model is Spanish words sequence, we use the output of 12st decoder layer as the linguistic embedding which has a dimension of 768. A post processing were used to up-sample linguistic embedding to align with the phoneme sequence. The phoneme sequence generated by front-end contains the symbol ' ' (space) to indicates the word boundaries, so the word-to-word aligning between phoneme sequence and linguistic embedding is conveniently. Generally, one Spanish word can be separated into more than two sub-words by the tokenizer of the BERT model, we use the embedding of the middle sub-word to represent the whole word, because the embedding of each sub-word contains context information. This sub-word embedding will be up-sampled according to the length of phoneme sequence of the corresponding word, and named as the linguistic features.

The GST module components by a reference encoder and a multi-head attention. The input of GST module is mel-spectrogram of reference audio, the output is the global style embedding. During the training stage, the input mel-spectrogram is the same as the target mel-spectrogram. During the inference stage, the input mel-spectrogram is from the target speaker. The global style embedding will be up-sampled into the length of encoder outputs and named as GST embedding.

The encoder module components by a stack of convolution layers and a Bi-LSTM layer. The phoneme sequence are encoded as phoneme embedding by an embedding layer. the linguistic features reduced the dimension from 768 to 256 by a dense layer and concatenate with phoneme embedding, then as the input of the encoder. The output of the encoder will concatenate with GST embedding and fed into the attention based decoder.

The decoder reference from the structure from GST, but we add guided-attention loss [25] to speed up the convergence. In autoregressive process, 1 frame mel-spectrogram generated per-time step. The autoregressive process will stop when stop token value satisfied the threshold. The output mel-spectrogram will be synthesize by a GAN-based vocoder into waveform.

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

3.4. Vocoder

There are two mainstream approaches mapping acoustic features to raw audio, which are Wavenet-based autoregressive vocoder and MelGAN-based GAN [26] vocoder. The former is better in audio quality, while the latter is faster in synthesis speed. To balance these two aspects, we used HiFi-GAN as vocoder to synthesize raw waveforms from the acoustic features which produced by acoustic model, while HiFi-GAN is the state-of-the-art GAN vocoder that contains one generator and two discriminators. Compared with the previous MelGAN [27] vocoder, multi-period discriminator was applied in HiFi-GAN, it plays an important role in improving the robustness of the model. Besides, the designed multi-receptive field fusion module is meaningful in extracting rich information.

The training steps are mainly divided into two steps in our system, 1). we used a subset of AISHELL-3, a subset of Spanish M-AILABS data and the Spanish data set that offered by Blizzard Challenge to train the base model, the input of base model is mel-spectrogram, 2) then we fine-tune the model by using the acoustic GTA features from acoustic models.

4. Evaluation results

For task 2021-SH1, 12 systems were evaluated. The identifier of natural speech is R, and our team is B. The evaluation results will be discussed in the following sections.

4.1. Naturalness evaluation

The boxplot of naturalness evaluation results is shown in Figure 2. Our system has an average MOS score of 3.01, which is in a middle level in all evaluated systems. Our system performs not very well mainly from on two aspects. First is prosody, such as continuous pronounce. The front-end of our system outputs Spanish phoneme sequences, in which only space were used as separator between words, the lack of prosody symbols and continuous pronounce marks make the performance reasonable. Second is speech quality, the synthesized speech of our system lost some energy in high frequency level. This may because of the de-noise processing in data pre-process stage, to as possible as remove the noise in training audios, we use a high de-noise value and may damaged the quality of the original audio.

4.2. Similarity evaluation

The boxplot of similarity evaluation results is shown in Figure 3. Our system didn't have a good performance and lags behind most of systems. Under our analysis, the main reason is the imbalance of the number of two speakers in the training data of acoustic model, and we didn't fine-tune the acoustic model sufficiently with the target speaker. Furtherly, the de-noise processing mentioned in previous section may also have a negative impact on speaker similarity.

4.3. Word error rates evaluation

There are two kinds of Word error rates (WER) evaluation. One is Sharvard test in which the sentences and the reference natural recordings come from Sharvard corpus, in which the reference recording is a different speaker from the speaker of the data

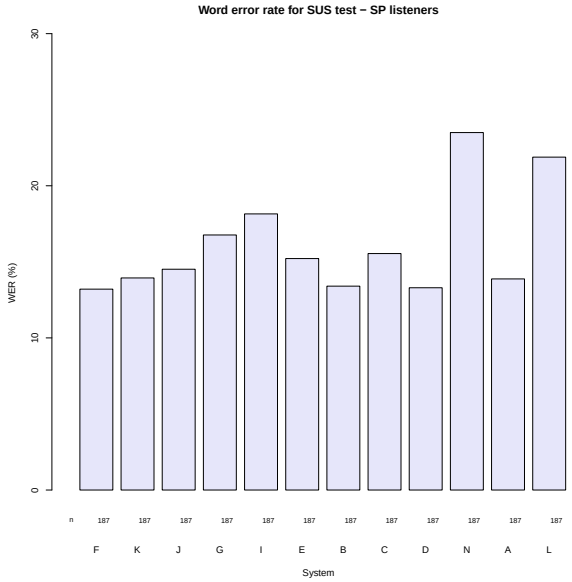

Figure 4: *Word error rates for SUS test of each submitted system for paid listeners(2021-SH1).*

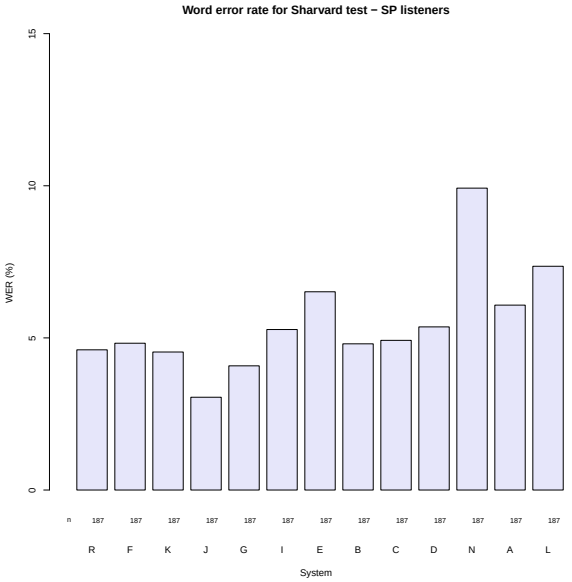

Figure 5: *Word error rates for Sharvard test of each submitted system for paid listener (2021-SH1).*

provided by committee. The other one is SUS test.

The evaluation results for SUS test is shown in Figure 4. The results shows that our system, together with system D and system F, achieved the lowest WER in the SUS test. Its performance has no significant differences with all other systems except system L and system N. The evaluation results for Sharvard test is shown in Figure 5. Our system still performs well in Sharvard test and is close to the WER of natural speech.

These two evaluation shows that, our system performs good in intelligence, which prove that the linguistic features derived from the pre-trained BERT model has a good effect on the system performance, even we only use the sub-word BERT embedding to represent the whole word.

5. Conclusion

In this paper, the details of our submitted system and summarized result in Blizzard Challenge were discussed. In our system, a linguistic feature process module based on the pre-trained BERT model with post processing were proposed. The main aim of post processing is to solve the time resolution mismatch between BERT embedding and phonemes sequence. A GST based model were used as acoustic model, The encoder is conditioned on up-sampled linguistic features to improve the prosody and intelligibility. HiFi-GAN as the vocoder to generate high quality speech. In the future, we will continue to investigate how the linguistic information, such as word embedding, sentence structure and prosody information, influence the result of TTS system and how to make better use of them.

6. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," 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] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *International Conference on Machine Learning*. PMLR, 2017, pp. 195–204.

[4] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *Advances in neural information processing systems*, vol. 30, 2017.

[5] W. Ping, K. Peng, A. Gibiansky, S. Ö. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning." in *ICLR (Poster)*, 2018.

[6] 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 *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 3171–3180.

[7] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2020.

[8] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Real-time neural text-to-speech with sequence-to-sequence acoustic model and waveglow or single gaussian wavernn vocoders." in *INTERSPEECH*, 2019, pp. 1308–1312.

[9] 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," *arXiv e-prints*, pp. arXiv–1609, 2016.

[10] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[11] 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.

[12] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[13] 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.

[14] T. Hayashi, S. Watanabe, T. Toda, K. Takeda, S. Toshniwal, and K. Livescu, "Pre-trained text embeddings for enhanced text-to-speech synthesis." in *INTERSPEECH*, 2019, pp. 4430–4434.

[15] H. Ming, L. He, H. Guo, and F. K. Soong, "Feature reinforcement with word embedding and parsing information in neural tts," *arXiv preprint arXiv:1901.00707*, 2019.

[16] Y. Xiao, L. He, H. Ming, and F. K. Soong, "Improving prosody with linguistic and bert derived features in multi-speaker based mandarin chinese neural tts," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6704–6708.

[17] T. Kenter, M. K. Sharma, and R. Clark, "Improving prosody of rnn-based english text-to-speech synthesis by incorporating a bert model," 2020.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, and N. A. Smith, "Linguistic knowledge and transferability of contextual representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 1073–1094.

[20] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4593–4601.

[21] G. Xu, W. Song, Z. Zhang, C. Zhang, X. He, and B. Zhou, "Improving prosody modelling with cross-utterance bert embeddings for end-to-end speech synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6079–6083.

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

[23] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[24] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6940–6944.

[25] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4784–4788.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[27] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in Neural Information Processing Systems*, vol. 32, 2019.