SUTD-NUS System for Blizzard Challenge 2021

Mingyang Zhang[1], *Xuehao Zhou*[1], *Kun Zhou*[1], *Rui Liu*[1,2], *Perry Lam*[2], *Berrak Sisman*[2], *Haizhou Li*[1]

[1]National University of Singapore
[2]Singapore University of Technology and Design

elezmin@nus.edu.sg, xuehao.zhou@u.nus.edu, zhoukun@u.nus.edu, liurui_imu@163.com,
perry_lam@mymail.sutd.edu.sg, berraksisman@u.nus.edu, haizhou.li@u.nus.edu

Abstract

This paper describes our effort to build the SUTD-NUS system for Blizzard Challenge 2021. The challenge has two tasks: 1) Hub task 2021-SH1: to build a Spanish text-to-speech (TTS) system using about 5 hours data from a European Spanish female speaker, and 2) Spoke task 2021-SS1: to build a TTS system that is able to synthesize the Spanish text containing a small amount of English words, using the same training data as Hub task 2021-SH1. Our submitted system is an end-to-end TTS structure that can generate acoustic features from text input. MelGAN neural vocoder are utilized to generate speech waveforms from acoustic features for both SH1 and SS1 tasks. Evaluation results provided by the challenge organizers demonstrate the effectiveness of our submitted TTS system.

Index Terms: speech synthesis, text-to-speech, Blizzard Challenge

1. Introduction

The Blizzard Challenge, held annually since 2005 [1], aims to promote research techniques in speech synthesis and provides a common platform with the necessary data. The challenge consists of two tasks this year: 1) Hub task 2021-SH1: to generate Spanish speech from the texts that only contains Spanish words; 2) Spoke task 2021-SS1: to generate the speech from Spanish texts that also contains a small amount of English words in each sentence.

Text-to-speech (TTS) aims to synthesize the voice from the text, and can be broadly divided into two types: concatenative speech synthesis [2, 3] and statistical parametric speech synthesis [4, 5, 6]. In concatenative speech synthesis, the synthesized speech is constructed by speech segments selected from a database. Even though concatenative approaches can produce high-quality speech, boundary artifacts still remain to be a key issue. Statistical parametric approaches parameterize speech signals into acoustic features, and map text features to acoustic features with an acoustic model. Thus, acoustic model training has been the focus for statistical parametric speech synthesis. With the advent of deep learning, neural network (NN)-based TTS, such as deep neural network (DNN) based [5, 7] and recurrent neural network (RNN) based methods [8, 9], have advanced the state-of-the-art in acoustic modelling. NN-based methods also greatly improve the performance of vocoders, such as WaveNet [10] and WaveRNN [11]. Thanks to its high flexibility and low data cost, parametric speech synthesis has gained widespread interests in previous TTS research.

Recent sequence-to-sequence (seq2seq) methods in TTS include Tacotron [12], Deep Voice 3 [13] and FastSpeech [14]. These methods learn to associate the text sequence and the acoustic features in an end-to-end manner. Compared to conventional TTS systems, seq2seq-based speech synthesis frameworks can synthesize highly intelligible and natural speech, and are less dependent on human label preprocessing and feature engineering.

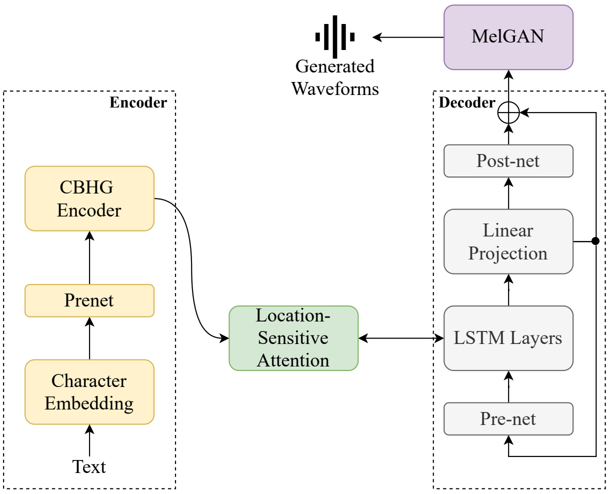

Figure 1: *TTS synthesizer system architecture*

We build our system based on Tacotron2 [15], which is an end-to-end TTS system with a neural vocoder. Considering that both Spanish and English share the same alphabet, we utilize the characters to represent the linguistic information as the input to the encoder. We also concatenate the character embedding with the tone embedding to improve the expressiveness of the synthesized speech. MelGAN [16] is adopted as the neural vocoder to reconstruct the waveform from the acoustic features.

This paper is organized as follows: in Section 2, we describe our system implementation for these two tasks; in Section 3, we report the experimental results; Section 4 concludes this paper.

2. System Architecture

2.1. End-to-end TTS synthesizer

Our system architecture is illustrated in Figure 1. The model is a modified Tacotron2 [15], an end-to-end TTS system that has a seq2seq encoder-decoder architecture with attention mechanism. It comprises (1) an encoder that converts input characters to a fixed-dimension text embedding, (2) an attention-based decoder that reconstructs acoustic features from the text embedding, and (3) a neural vocoder that generates human-like speech from the acoustic features.

The encoder aims to generate linguistic representations from the input text sequence. For the SH1 task, the input texts only contain Spanish words, but for the SS1 task there will be some English words in the input sentences. Since the Spanish alphabet is a superset of the English alphabet, we directly use the characters as the input for both tasks. In addition, Spanish uses ”¿” and ”¡” to mark questions and exclamations at the beginning of an utterance, for example:

- ¿Cuál es la peor película que has visto?
- ¡Cubrimos los regalos de todos!

To make our synthesis model more expressive, we regard ”¿” and ”¡” as tone information in the input representation of our TTS system. The tone IDs are first converted to an 8-dimensional tone embedding through a look-up table, then concatenated with the corresponding character embedding. The concatenated input vectors are passed to a 2 layer pre-net, followed by the CBHG [12] text encoder to generate the latent text embedding.

The encoder output is attended by the attention-based decoder, which is a RNN-based network that predicts acoustic frames from encoder output. The attention mechanism computes a fixed-length context vector to provide additional input to the decoder network. Location-sensitive attention is able to reduce the frame prediction errors [17]. The decoder input is passed through 2 layer pre-net and 2 LSTM layers, followed by a linear projection layer to predict acoustic features. The residual structure of a 5 layer convolutional postnet improves the general reconstruction. The stop token symbol is to predict when the model should stop during inference.

For both SH1 and SS1 tasks, we use the same input representation and model architecture. The output acoustic features are mel-spectral features extracted from 48kHz raw audios, with 12.5ms frame shift and 25ms frame length.

2.2. Neural Vocoder

For rapid inference and high quality waveform generation, we use MelGAN as the neural vocoder to synthesize audio from mel spectra [16]. MelGAN is a generative adversarial network (GAN) based waveform generation model [18] with non-autoregressive feed-forward convolutional architecture. It consists of a fully convolutional feed-forward network as the generator and a multi-scale discriminator. The model is trained by jointly optimizing multi-resolution short-time Fourier transform (STFT) loss and adversarial loss that enables the model to capture the time-frequency distribution of the realistic waveform effectively.

The waveform generator is a fully convolutional feed-forward network that transforms mel-spectral features to the output waveform. We followed the model architecture that described in [16]. To make the time resolution of the Mel-spectral features and the waveform consistent, the strides of the transposed convolutional upsampling layers are [6, 5, 5, 4] respectively.

For the discriminator, it has multiple discriminators operating on different audio scales. This is because the audio signal has different levels of characteristics. Each discriminator then learns the corresponding feature from different frequency ranges. The architecture of the discriminator is illustrated in Table 1.

Multi-resolution STFT loss is introduced to improve the stability and efficiency of the adversarial traning procedure. The loss is computed by the sum of three different STFT losses with different FFT size, frame shift and window size. The three sets of STFT parameters are illustrated in Table 2.

Table 1: *Discriminator architecture*

15 × 1, stride=1, conv 16, LeakyReLU
41 × 1, stride=4, group=4, conv 64, LeakyReLU
51 × 1, stride=5, group=16, conv 320, LeakyReLU
51 × 1, stride=5, group=80, conv 1600, LeakyReLU
61 × 1, stride=6, group=400, conv 1600, LeakyReLU
5 × 1, stride=1, conv 1600, LeakyReLU
3 × 1, stride=1, conv 1

Table 2: *Parameters for STFT losses*

FFT size	Frame shift	Window size
1024	120	600
2048	240	1200
512	50	240

3. Results

3.1. Challenge Participants

In total, 12 teams (A/B/C/D/E/F/G/I/J/K/L/N) submitted their results for the SH1 task and 11 teams (A/C/D/E/H/I/K/L/M/N) participated in the SS1 task. For both tasks, system R is natural speech. Our system is labelled as A.

3.2. Evaluation metrics

Subjective listening tests were designed to perceptually evaluate the synthetic samples for all systems in both SH1 and SS1 tasks. For the SH1 task, three sets of experiments were conducted to evaluate the synthetic samples, including naturalness, similarity to original speaker, and intelligibility. For the SS1 task, naturalness, similarity to original speaker and acceptability of English words were reported to evaluate the performance. The detailed results will be presented in the next sections.

3.3. Perceptual evaluation for SH1 task

3.3.1. Naturalness

These sets of experiments were conducted to evaluate the naturalness of the synthetic sentences. The listeners were asked to assign a score to represent how natural or unnatural of the speech sample, where a score 1 indicates the speech sample is ”Completely Unnatural”, while a score 5 indicates that the speech sample is ”Completely Natural”.

Figure 2 shows the boxplot of mean opinion scores (MOS) of the naturalness for synthetic sentences. Our system obtains an average MOS of 2.17 ± 1.08 standard deviation. For reference, thenatural speech has a score of 4.21 ± 0.93.

3.3.2. Similarity to original speaker

These sets of experiments were conducted to evaluate the similarity between the synthetic sentences and the original speaker. The listeners were asked to assign a score to represent how similar the synthetic voice sounded to the voice in the reference samples, where a score of 1 indicates the speech sample ”Sounds like a totally different person”, while a score of 5 is ”Sounds like exactly the same person”.

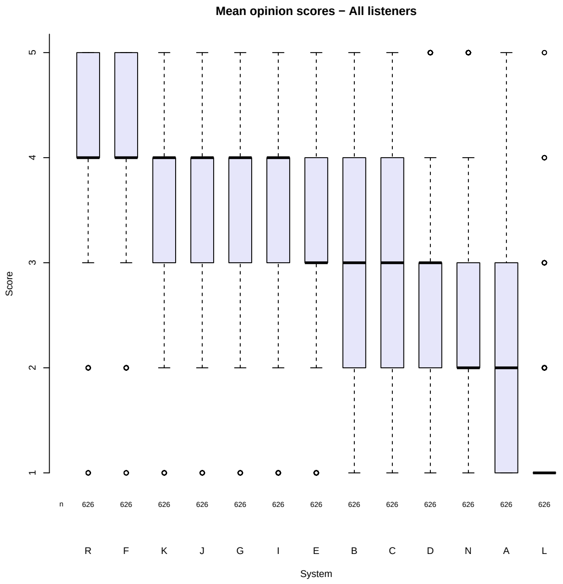

Figure 2: *Boxplot of naturalness scores of sentence synthesis for all listeners for SH1 task. A is our system.*

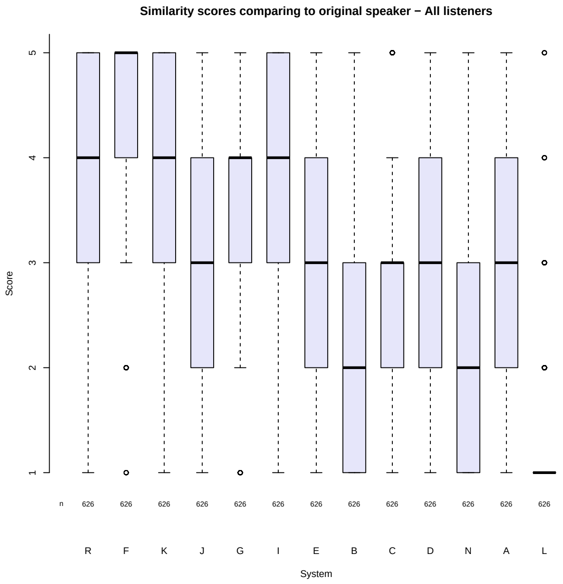

Figure 3: *Boxplot of similarity scores of sentence synthesis for all listeners for SH1 task. A is our system.*

Figure 3 shows the boxplot of similarity scores of the synthetic sentences. Our sytem obtains an average MOS of 2.81 with 1.28 standard deviations.

3.3.3. Intelligibility

The intelligibility evaluation of the SH1 task is performed by dictation, where listeners were asked to write down the contents they heard from the given samples. The performance is evaluated by calculating the word error rate.

Figure 4 shows the word error rate for the INT-Sharvard section. In this section, our system obtains a 6.1% word error rate with 0.19 standard deviations. As a reference, the natural speech has a 4.6% word error rate with 0.14 standard deviations. Figure 5 shows the word error rate for the INT-SUS section. In this section, our system obtains a 14% word error rate and 0.21 standard deviations.

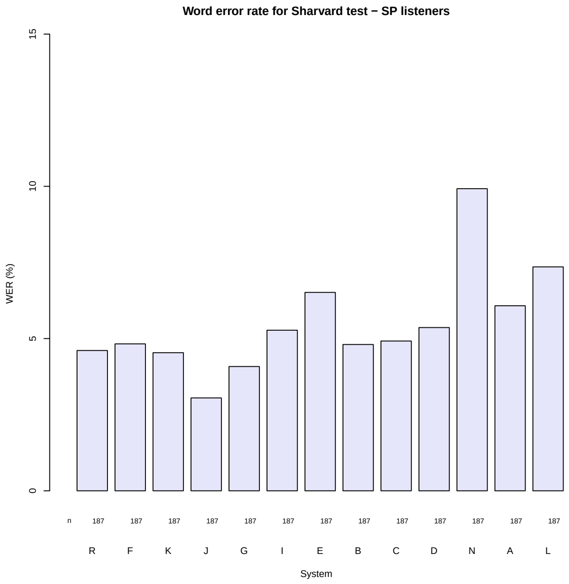

Figure 4: *Boxplot of SHARVARD intelligibility scores of sentence synthesis for all listeners for SH1 task. A is our system.*

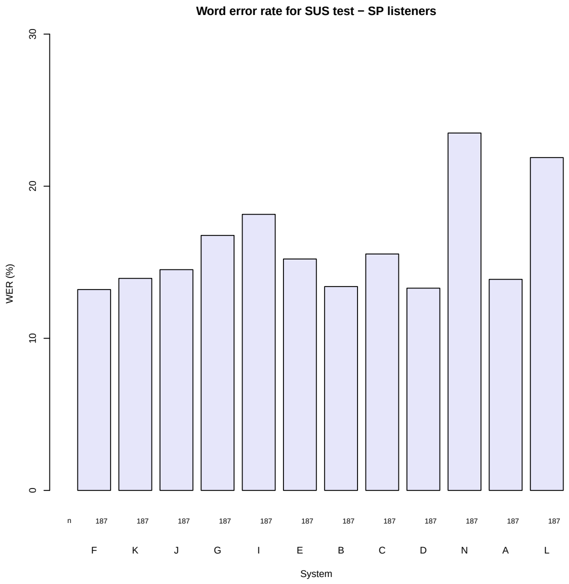

Figure 5: *Boxplot of SUS intelligibility scores of sentence synthesis for all listeners for SH1 task. A is our system.*

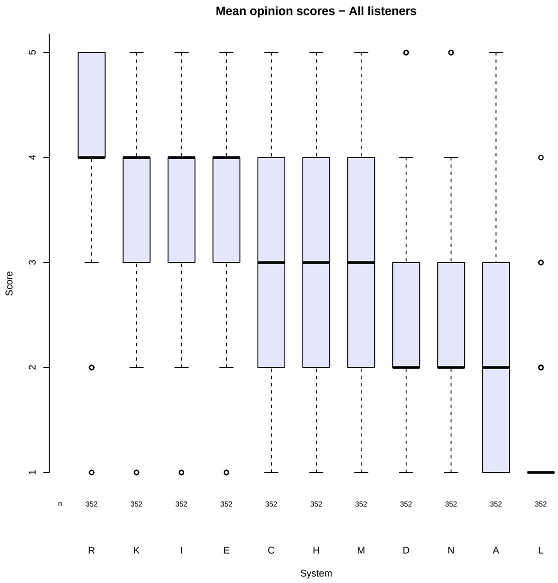

Figure 6: *Boxplot of naturalness scores of sentence synthesis for all listeners for SS1 task. A is our system.*

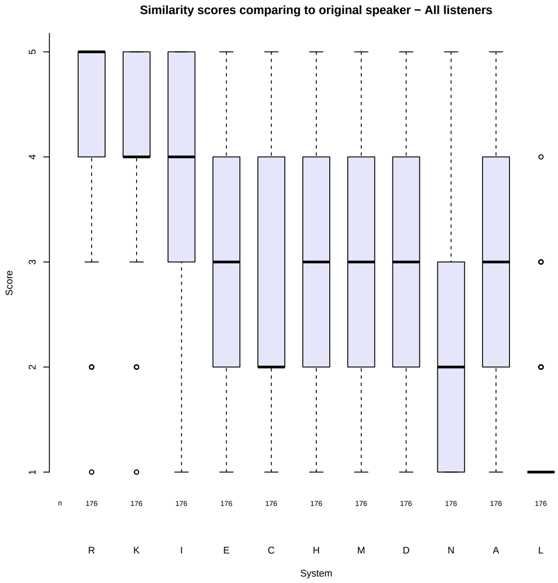

Figure 7: *Boxplot of similarity scores of sentence synthesis for all listeners for SS1 task. A is our system.*

3.4. Perceptual evaluation for SS1 task

3.4.1. Naturalness

These sets of experiments were conducted to evaluate the naturalness of the synthetic sentences. The listeners were asked to assign a score to represent how natural or unnatural of the speech sample, where a score 1 indicates the speech sample is ”Completely Unnatural”, while a score 5 indicates that the speech sample is ”Completely Natural”.

Figure 6 shows the boxplot of mean opinion scores (MOS) of the naturalness for synthetic sentences. Our system obtains an average MOS of 2.28 with 1.08 standard deviation. While, as a reference, the natural speech has a score of 4.29 with 0.88 standard deviations.

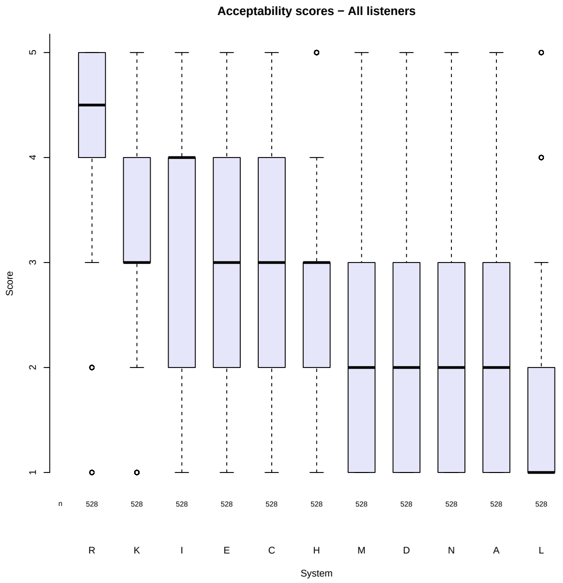

Figure 8: *Boxplot of acceptability scores of sentence synthesis for all listeners for SS1 task. A is our system.*

3.4.2. Similarity to original speaker

These sets of experiments were conducted to evaluate the similarity between the synthetic sentences and the original speaker. The listeners were asked to assign a score to represent how similar the synthetic voice sounded to the voice in the reference samples, where a score 1 indicates the speech sample is ”Sounds like a totally different person”, while a score 5 indicates that the speech sample is ”Sounds like exactly the same person”.

Figure 7 shows the boxplot of similarity scores of the synthetic sentences. Our system obtains an average MOS of 2.76 with 1.25 standard deviations.

3.4.3. Acceptability of English words

These sets of experiments were conducted to evaluate how acceptable or unacceptable the English words were in the synthetic sentences. The listeners were asked to assign a score to represent the acceptability of the English words, where a score 1 indicates the English words are ”Not Intelligible”, while a score 5 indicates the English words are ”Perfect”.

Figure 8 shows the boxplot of the acceptability scores of the synthetic sentences. Our system obtains an average MOS of 2.20 ± 1.08.

4. Conclusions

This paper presents the SUTD-NUS system submitted for Blizzard Challenge 2021. We built a TTS framework that first transforms text input to acoustic features using an end-to-end TTS synthesizer, followed by a MelGAN neural vocoder to construct the audio waveform. The effectiveness of our system is successfully confirmed by the official evaluation results.

5. References

[1] A. Black and K. Tokuda, "The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common databases," in *Proceedings of interspeech*, 2005, pp. 77–80.

[2] 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.

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

[4] 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.

[5] H. Ze, 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] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[7] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (dnn) for parametric tts synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3829–3833.

[8] H. Zen, "Acoustic modeling in statistical parametric speech synthesis-from hmm to lstm-rnn," 2015.

[9] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4470–4474.

[10] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 3918–3926.

[11] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *arXiv preprint arXiv:1910.06711*, 2019.

[17] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *arXiv preprint arXiv:1506.07503*, 2015.

[18] 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.