Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020
30 October 2020, Shanghai, China

The SHNU System for the Blizzard Challenge 2020

Laipeng He[12], *Qiang Shi*[2], *Lang Wu*[2], *Jianqing Sun*[2], *Renke He*[1], *Yanhua Long*[1], *Jiaen Liang*[2]

[1]SHNU-Unisound Joint Laboratory of Natural Human-Computer Interaction, Shanghai Normal University, Shanghai, China
[2]Unisound AI Technology Co.,Ltd, Beijing, China

sunjianqing@unisound.com, yanhua@shnu.edu.cn

Abstract

This paper introduces the SHNU (team I) speech synthesis system for Blizzard Challenge 2020. Speech data released this year includes two parts: a 9.5-hour Mandarin corpus from a male native speaker and a 3-hour Shanghainese corpus from a female native speaker. Based on these corpora, we built two neural network-based speech synthesis systems to synthesize speech for both tasks. The same system architecture was used for both the Mandarin and Shanghainese tasks. Specifically, our systems include a front-end module, a Tacotron-based spectrogram prediction network and a WaveNet-based neural vocoder. Firstly, a pre-built front-end module was used to generate character sequence and linguistic features from the training text. Then, we applied a Tacotron-based sequence-to-sequence model to generate mel-spectrogram from character sequence. Finally, a WaveNet-based neural vocoder was adopted to reconstruct audio waveform with the mel-spectrogram from Tacotron. Evaluation results demonstrated that our system achieved an extremely good performance on both tasks, which proved the effectiveness of our proposed system.

Index Terms: Blizzard Challenge 2020, Tacotron, WaveNet, SPSS, Speech Synthesis, Text-to-speech

1. Introduction

Conventional statistical parametric speech synthesis (SPSS) [1] system usually consists of front-end module, duration model, acoustic model and vocoder. Front-end module takes text as input and generates phoneme sequence and linguistic features. Duration model predicts the number of frames a phoneme lasts, while acoustic model predicts the acoustic features of each frame. Finally, speech waveform was reconstructed from acoustic features using a source-filter vocoder such as STRAIGHT [2] and WORLD [3].

Another mainstream speech synthesis solution is unit selection based waveform concatenation approach, which selects suitable speech units before concatenating them together. Benefit from the direct use of natural speech segments, these systems can be exploited for the advantage of large scale corpus, and synthesize speech close to human level. The main disadvantages are the requirement of large speech corpus and expert fine-tuning [4]. Statistical models such as HMM, DNN and LSTM are usually used to guide unit selection [5, 6].

In recent years, with the development of deep learning, neural network-based text-to-speech (TTS) systems become the state-of-the-art techniques. In these systems, the sequence-to-sequence acoustic model such as Tacotron [7] and Transformer [8] are used to generate acoustic features. Efforts have also been made to improve stability [9, 10, 11]. Their vocoders are also neural network-based ones, such as WaveNet [12], WaveGlow [13], MelGAN [14], etc. Extensive experiments in literature showed that neural network-based TTS systems have significantly outperformed the conventional SPSS and unit selection based methods.

Blizzard Challenge has been held annually since 2005, and this is our first participation. In previous challenges, most of the best results were achieved by using the SPSS and unit selection based approaches. However, in this year's challenge, our neural network-based TTS systems worked very well, and achieved the best results on most aspects of both tasks. Our system is composed of three parts: a front-end module, a Tacotron-based spectrogram prediction network and a WaveNet-based neural vocoder. Initially, a multi-speaker system was trained on our own 90-hour speech data, mostly of which are in Chinese as well as small amount of English, by 4 female and 2 male native Mandarin speakers. Then the official released 9.5-hour Mandarin corpus and 3-hour Shanghainese corpus were used to perform model fine-tuning to obtain the final synthesis systems.

The rest of the paper is organized as follows: Section 2 introduces the framework of our system. Section 3 illustrates the construction process of our system in detail. The evaluation results and conclusions are given at the end of the paper.

2. Framework

2.1. Front-end

The front-end module for Mandarin TTS typically includes text normalization (TN), Chinese word segmentation (CWS), part-of-speech (POS) tagging, grapheme-to-phoneme (G2P) conversion, and prosodic boundary prediction [15]. The TN module converts special marks such as numbers or symbols to Chinese characters. Since there is no separator between Chinese characters, CWS module should be used to split the sentence into a series of lexical words. With clear word boundaries, POS can be predicted using statistical models like conditional random fields (CRF) [16]. The goal of G2P module is to convert Chinese characters into Pinyin sequence. Polyphone, tone sandhi [17] and Erhua are three special language phenomena in Chinese which significantly affect the intelligibility of synthesized speech.

Modern Chinese Dictionary collects over 1000 polyphones, resulting in various pronunciations for a single character. Our solution is the application of a series of CRF models for the most common polyphones. If a polyphone is part of a word, its pronunciation is certain and can be determined by looking up a dictionary. But if the word is a polyphone, alone, then its pronunciation would be determined by the CRF model. Word form and POS information are used as model input.

Tone sandhi is a phonological change occurring in tonal languages, in which the tones assigned to individual words or morphemes change based on the pronunciation of adjacent words or morphemes. The most famous tone sandhi in Mandarin is the third-tone sandhi, whereby a third tone (dipping)

10.21437/VCC_BC.2020-2

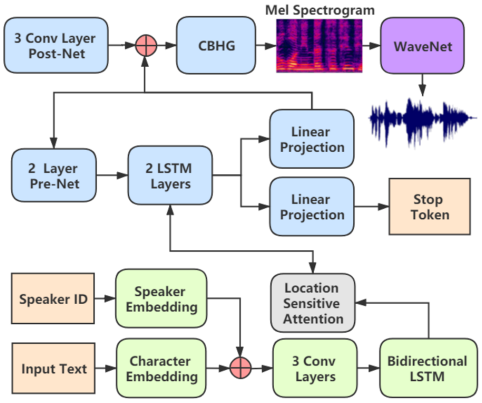

Figure 1: *Framework of Tacotron-based spectrogram prediction network.*

becomes a second tone (rising) in the context of a following third tone. Another important tone sandhi is tones on special syllables such as ”bu” and ”yi”. Our system uses grammatical rules [17] with the help of word and prosodic boundary information to handle these two types of tone sandhi.

In Mandarin, Erhua refers to the sound change phenomenon of the vowels due to the action of rolling tongue. In other words, when a syllable is followed by ”er”, the two syllables may be merged into one syllable, and the pronunciation will also change. Of course, there are exceptions. Therefore, a vocabulary list with the words that cannot be pronounced was built, words not in the list were pronounced as Erhua.

Prosodic prediction module is for boundary identification of prosodic word (L1) and prosodic phrase (L3). A CRF model was applied to predict L1 boundary with word form, POS and word length as input. Another one was for L3 boundary with word form, POS, word length and L1 boundary as input [18].

2.2. Tacotron-based spectrogram prediction network

A Tacotron-based sequence to sequence model was applied to predict mel-spectrogram from the input character sequence. The network is mainly composed of an encoder and a decoder with attention as shown in Figure 1. The encoder converts a character sequence into hidden feature representations. Decoder is an auto-regressive model which predicts mel-spectrogram from the encoded input sequence. The location-sensitive attention (LSA) [19] was adopted to mitigate potential failure modes where some sub-sequences were repeated or ignored by the decoder [20]. The CBHG [7] was used as a post-processing for decoder to achieve a higher prediction accuracy.

Chinese characters were first converted to Pinyin sequence by using a front-end module and then further represented by International Pronunciation Alphabet (IPA). Prosodic boundary representations were inserted between IPA characters, including prosodic word and prosodic phrase. For Shanghainese, however, the phonetic transcriptions were used straightly from the original training data as input of Tacotron.

In our system, a trainable embedding table was selected to store speaker embedding. Each paragraph was equipped with an embedding for style modeling, due to the multiple paragraphs contained in the original Mandarin training text as well as its different pronunciation styles for each paragraph respectively. We use the same embedding table to represent speaker and paragraph embedding for simplicity.

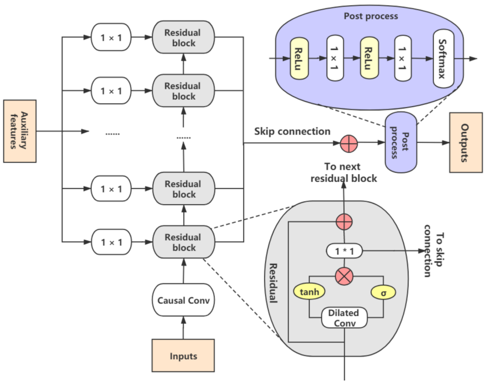

Figure 2: *Framework of WaveNet-based neural vocoder.*

2.3. WaveNet-based neural vocoder

WaveNet is a high-quality neural vocoder for generating audio waveform. Given a sequence of waveform $\mathbf{x} = \{x_1, ..., x_T\}$ and auxiliary feature $\mathbf{h}$, the joint probability of wave samples is represented as follows:

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p\left(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h}\right) \tag{1}$$

Global and local conditioning are two different ways to condition WaveNet on auxiliary inputs. In this paper, WaveNet model was conditioned on two extra inputs, speaker embedding and mel-spectrogram were used as global and local conditioning respectively. Figure 2 shows the framework of WaveNet-based neural vocoder.

In our system, 1024 possible values was mixed with quantized results of compressed audio samples, after applying a μ-law companding transformation initially. 30 stacked dilated convolutions grouped into 3 dilation cycles were adopted and filter width was set to 2. The dilation is doubled for every layer up to a limit of 512 and then repeated:

$$1, 2, 4, ..., 512; 1, 2, 4, ..., 512; 1, 2, 4, ..., 512 \tag{2}$$

3. System building

As indicated in Figure 3, our system was developed by three steps: data preparation, training phase and synthesis phase. Each step will be described in detail as follows.

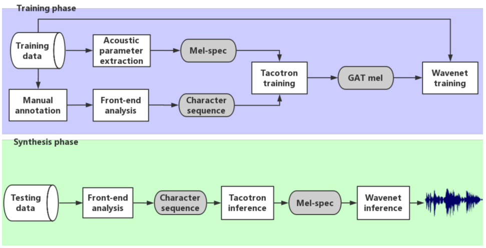

Figure 3: *Flowchart of our neural network-based speech synthesis system.*

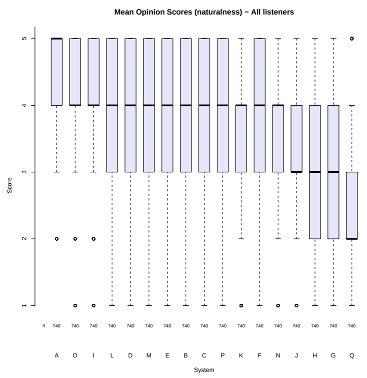

Figure 4: *Boxplot of naturalness scores of each submitted Mandarin system for all listeners.*

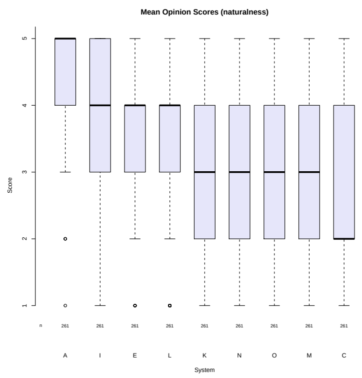

Figure 5: *Boxplot of naturalness scores of each submitted Shanghainese system for all listeners.*

3.1. Data preparation

Before model training, manual annotations were performed on the training data. Some extremely long sentences were divided into multiple clauses based on consistency of pronunciation and text. In addition, prosodic word and prosodic phrase boundary annotations were performed on Mandarin data. Silence at the beginning and end of sentence was trimmed to further improve the stability of attention. Training text was first sent to a front-end module to get Pinyin sequence and L1/L3 boundaries. $<$ Character, mel-spectorgram $>$ pairs were generated and used for training Tacotron. Sample rate of 24kHz was used for Tacotron and WaveNet modeling, making a trade-off between synthetic sound quality and modeling accuracy. Since the original sample rate of Shanghainese was 16kHz, we upsampled the audio to 24kHz by bandwidth extension. We built a fake spectrogram based on the high frequency band of the original 16kHz audio. Then we combined the original spectrogram (0-8kHz) and the fake spectrogram (8-12kHz) to generate a new audio signal with a sampling rate of 24kHz.

3.2. Training phase

At the training phase, a multilingual, multi-speaker Tacotron model was trained firstly, based on a large-scale training corpus. Mandarin and Shanghainese corpus were adopted to perform fine-tuning based on the pre-trained Tacotron model. It was worth noting that our large-scale training corpus does not contain Shanghainese data, but the experiments proved that the above training scheme was also effective for Shanghainese , which means that the training data of Mandarin was helpful for Shanghainese training. When the training procedure was finished, in-set character sequence was sent to the retrained Tacotron model to generate mel-spectrogram under ground truth align (GTA) mode. Then, the GTA mel-spectrogram was used for WaveNet fine-tuning from the pre-trained model. The mean and variance of ground truth (GT) and GTA mel-spectrogram were calculated respectively for global variance (GV) operation.

3.3. Synthesis phase

At the synthesis phase, Mandarin test sentences were sent to the front-end module to get Pinyin sequence and L1/L3 boundaries. Then, character sequence with prosodic boundary information was fed into the Tacotron-based spectrogram prediction network to generate mel-spectrogram. For Shanghainese, the phonetic transcriptions were used straightly as input of Tacotron, consistent with the training phase. Global variance conversion was performed on the generated mel-spectrogram after CBHG module to adjust the dynamic range of the generated mel-spectrogram. Speech waveform was synthesized by the WaveNet-based neural vocoder conditioning on the post-processed mel-spectrogram.

4. Evaluations

The official evaluation results of our system are presented in this section. System evaluation is divided into four dimensions: naturalness, similarity, intelligibility and paragraph performance. This year, 17 systems (16 participating teams and one natural speech) were evaluated for Mandarin task, and 9 systems (8 participating teams and one natural speech) for the Shanghainese task. Our system was marked as I and the original sound as A.

4.1. Naturalness test

Figure 4 and 5 show the boxplot of mean opinion scores (MOS) of each system on naturalness of synthesized speech. The MOS results showed a two-way tie for first place in all submitted systems for Mandarin tasks, one of which belongs to our team， that is the sore of 4.2. For the Shanghainese , our scores ranked the top, with 4.0 in MOS.

4.2. Similarity test

Figure 6 and 7 present the boxplot MOS of each submitted system on similarity between the synthesized speech and the reference recording. In the Mandarin task, our system obtained an average opinion similarity score of 4.2, one of the highest among all submitted systems. In the Shanghainese task, our system scored 3.6 points, ranking third among all submitted systems.

4.3. Intelligibility test

In this section, two different schemes were used to test Mandarin and Shanghainese. The Mandarin test process requires the listener to record the sound they hear after listening to the synthesized speech. The Shanghai dialect's process uses a scoring mechanism similar to the above-mentioned naturalness and

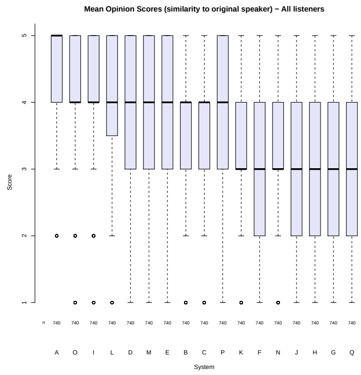

Figure 6: *Boxplot of similarity scores of each submitted Mandarin system for all listeners.*

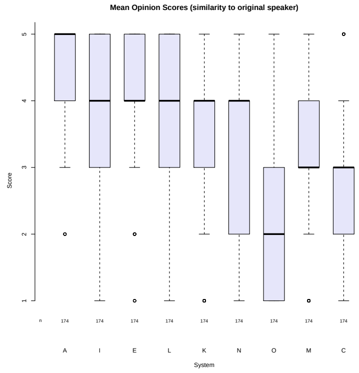

Figure 7: *Boxplot of similarity scores of each submitted Shanghainese system for all listeners.*

similarity. For Mandarin task, as shown in Figure 8, the pinyin error rate with tones of our system is 0.09. The lowest PER score is 0.086 of system L, but Wilcoxon signed rank tests indicate that the difference is not significant between system L and ours. The evaluation results show that our system does a very good job in terms of intelligibility, and solves most of the Erhua problems. Figure 9 presents the boxplot MOS of each submitted system on intelligibility in the Shanghainese task. From the results, the system score we submitted is 4.1 points, the highest score among all submissions.

4.4. Paragraph performance

This test was to evaluate synthesized speech from multiple aspects, including overall impression, pleasantness, speech pauses, stress, intonation, emotion, and listening efforts. In each part, listeners were asked to listen to one whole paragraph from news before choosing a score from 1 to 60 for each aspect. The mean opinion scores of our system are listed in Table 1. This results showed our system achieved the highest comprehensive scores among all submitted systems.

5. Conclusions

This paper introduces the details of the SHNU system for the Blizzard Challenge 2020. For both the Mandarin and Shanghainese tasks, we built a neural network-based TTS system respectively to generate the synthetic speech. Both systems have the same architecture: a front-end module, a Tacotron-based spectrogram prediction network and a WaveNet-based neural vocoder. Both the Mandarin and Shanghainese systems were fine-tuned from a multi-speaker Tacotron-WaveNet system which has been trained on our own large-scale corpus. From the released official evaluation results, we see that our system achieved an extremely well performance, which proves the effectiveness of our systems.

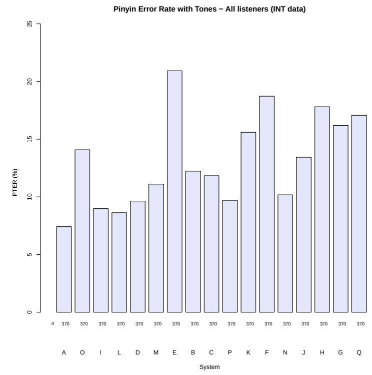

Figure 8: *Pinyin Error Rate with Tones(All listeners).*

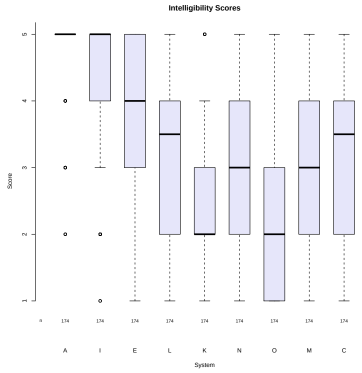

Figure 9: *Boxplot of intelligibility Scores of each Shanghainese submitted system for all listeners.*

Table 1: *Paragraph listening test scores of our system*

	MOS
overall impression	45
pleasantness	43
speech pauses	42
stress	42
intonation	42
emotion	43
listening effort	45

6. References

[1] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007, Honolulu, Hawaii, USA, April 15-20, 2007. IEEE 2007, ISBN 1-4244-0727-3*, 2007, pp. 1229–1232.

[2] 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, MAVEBA 2001, Florence, Italy, September 13-15, 2001. ISCA 2001*, 2001, pp. 59–64.

[3] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," in *IEICE Transactions on Information Systems, Volume 99-D*, 2016, pp. 1877–1884.

[4] Y. Jiang, X. Zhou, C. Ding, Y. jun Hu, Z.-H. Ling, and L.-R. Dai, "The ustc system for blizzard challenge 2018," 2018.

[5] Z.-H. Ling, H. Lu, G.-P. Hu, L.-R. Dai, and R.-H. Wang, "The ustc system for blizzard challenge 2008," 2008.

[6] V. Pollet, E. Zovato, S. Irhimeh, and P. D. Batzu, "Unit selection with hierarchical cascaded long short term memory bidirectional recurrent neural nets," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017. ISCA 2017*, 2017, pp. 3966–3970.

[7] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 4006–4010.

[8] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhouu, "Close to human quality tts with transformer," in *CoRR abs/1809.08895 (2018)*, 2018.

[9] 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 *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada. 2019*, 2019, pp. 3165–3174.

[10] 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 *CoRR, June 2020*, 2020.

[11] 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," in *CoRR, September 2019*, 2019.

[12] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016. ISCA 2016*, 2016, p. 125.

[13] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019. IEEE 2019, ISBN 978-1-4799-8131-1*, 2019, pp. 3617–3621.

[14] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Br é bisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada. 2019*, 2019, pp. 14 881–14 892.

[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," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020. IEEE 2020, ISBN 978-1-5090-6631-5*, 2020, pp. 6689–6693.

[16] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *18th ICML 2001: Williams College, Williamstown, MA, USA*, 2001, pp. 282–289.

[17] S. Duanmu, *The Phonology of Standard Chinese*. Oxford University Press, 2002.

[18] C. Ding, L. Xie, J. Yan, W. Zhang, and Y. Liu, "Automatic prosody prediction for chinese speech synthesis using blstm-rnn and embedding features," in *ASRU 2015: Scottsdale, AZ, USA*, 2015, pp. 98–102.

[19] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. 2015*, 2015, pp. 577–585.

[20] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R.-S. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018. IEEE 2018, ISBN 978-1-5386-4658-8*, 2018, pp. 4779–4783.