The Ximalaya TTS System for Blizzard Challenge 2020

Zhiba Su, Wendi He*, Yang Sun†*

Ximalaya FM

{zhiba.su, cloris.he, youngsuen}@ximalaya.com

Abstract

This paper describes the proposed Himalaya text-to-speech synthesis system built for the Blizzard Challenge 2020. The two tasks are to build expressive speech synthesizers based on the released 9.5-hour Mandarin corpus from a male native speaker and 3-hour Shanghainese corpus from a female native speaker respectively. Our architecture is Tacotron2-based acoustic model with WaveRNN vocoder. Several methods for preprocessing and checking the raw BC transcript are implemented. Firstly, the multi-task TTS front-end module transforms the text sequences into phoneme-level sequences with prosody label after implement the polyphonic disambiguation and prosody prediction module. Then, we train the released corpus on a Seq2seq multi-speaker acoustic model for Mel spectrograms modeling. Besides, the neural vocoder WaveRNN[1] with minor improvements generate high-quality audio for the submitted results. The identifier for our system is M, and the experimental evaluation results in listening tests show that the system we submitted performed well in most of the criterion.

Index Terms: Speech Synthesis, Deep Neural Networks, Tacotron2, WaveRNN, Blizzard Challenge 2020

1. Introduction

This is the first time for Himalaya to participate in this Blizzard Challenge (BC) as it has held every year since 2005 for speech synthesis techniques enhancement and communication.[2] The task of Blizzard Challenge 2020 is to build synthetic voices in Mandarin and Shanghainese. Testing set of sentences are also provided. Participates are required to submit the corresponding speech files generated by their own model. Then these files will be evaluated by paid participants, volunteers and speech experts.

Speech synthesis, also called TTS (text-to-speech) is a technology aiming to generate human-like speeches from texts. End-to-end TTS system is a type of system that can be trained on (text,audio) pairs. It usually contains 2 components: an acoustic model and a vocoder. Acoustic model predicts acoustic intermediate features from texts. As for vocoder, e.g. Griffin-Lim [3], WORLD [4], WaveNet [5] or WaveRNN synthesizes speeches with generated acoustic features.

In this paper, we propose a end-to-end Mandarin & Shanghainese TTS system based on Tacotron2 and WaveRNN. As Chinese texts aren't suitable for TTS, before feeding texts into acoustic model, we convert texts into Chinese pinyin (phoneme) sequences with tone firstly. With the given input pinyins, our system could generate high fidelity speech. Specifically, we follow the commonly used encoder-decoder structure with attention structure to get mel-spectrogram as acoustic features, then feed them into neural vocoder to generate speech. Moreover, polyphonic disambiguation and prosody prediction modules have been applied as the front-end part for better waveform quality.

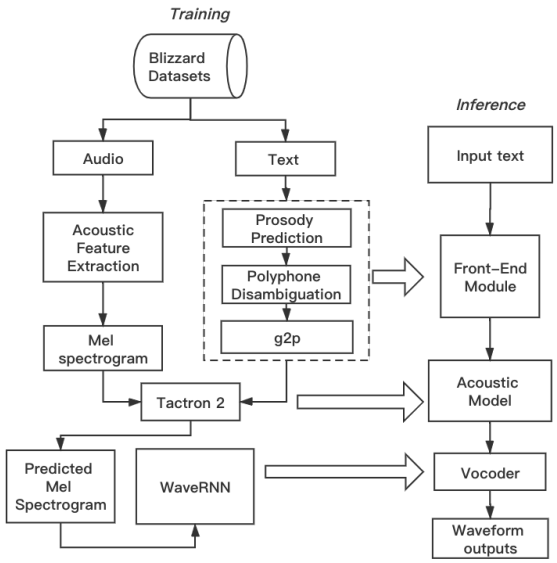

Figure 1: *Block diagram of the our system architecture*

The rest of the paper is structured as follows. Section 2 introduces Blizzard Challenge Tasks and the given datasets. Section 3 describes front-end structure and the details of our submitted system, followed by the evaluation results in section 4. Finally, the conclusion is given in section 5.

2. Task and Datasets in BC2020

2.1. Tasks

The two tasks this year are to build a voice from in-domain and out-domain evaluated sentences by making use of the given two training datasets:

- Mandarin Chinese (Task1) - About 9.5 hours of speech data from a male native speaker of Mandarin.
- Shanghainese (Task2) - About 3 hours of speech data from a female native speaker of Shanghainese.

2.2. Datasets Description

Datasets the committee provided including waveform files and corresponding Chinese-character text transcripts.

Provided audio data have mono channel, 16-bit depth with almost the same quality. The directory in Task1 containing 4365 audio files with 48 kHz sampling rate (around 9 hours 36 minutes) but only contain 1900 audio files with 16kHz sampling rate (around 2 hours 56 minutes) in Task2.

[1]Zhiba Su, Wendi He have equal contributions. Yang Sun is the corresponding author.

10.21437/VCC_BC.2020-10

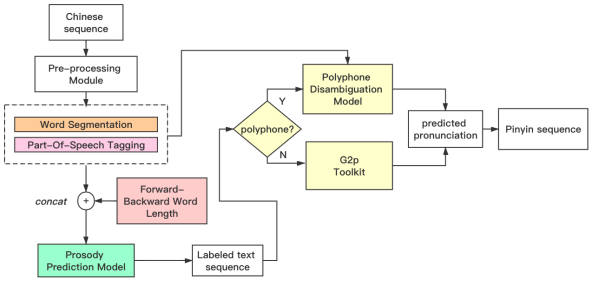

Figure 2: *Front-end architecture*

It is worth mentioning that several wrong phoneme results of some single words in the offered Shanghainese transcription have manually corrected. Besides, as the external datasets are allowed to use for data augmentation in this challenge, we use extra four-speaker datasets (Ximalaya®) in Task1.

3. System Description

As the goal for text-to-speech synthesis system is to get the waveform reconstructed from the text input, in this section, we will describe the detailed system structure we used in this challenge. See Figure 1.

3.1. Data Preprocessing

Because of the sampling rates of provided training data in two tasks are different, all the audio data from each task have been downsampled and upsampled by FFmpeg from 48kHz and 16kHz respectively to consistent 22kHz. The coefficient 0.97 is used for pre-emphasis process, then the acoustic features are extracted as 80-band mel-scale spectrogram from waveforms through STFT with 50ms frame size and 12.5ms frame hop. Text is converted into syllable sequence through front-end modules mentioned below, see Figure 2.

3.1.1. Polyphonic Disambiguation

In Mandarin speech synthesis, due to the existance of homographs, it is common to see that some Chinese characters have more than one pronunciation which called polyphonic ambiguity. Since the Grapheme-to-phoneme (G2P) conversion serves as an essential component in Chinese Mandarin text-to-speech system, the pronunciation of polyphonic characters directly affect the listening effort and naturalness. A variety of approaches have been proposed to address the polyphonic disambiguation problem. Some classification tasks have been proposed like DT-guided TBL[6] or Maximum Entropy model[7, 8, 9]. Besides, End-to-end framework have been served in recent years such as Bi-directional LSTM framework[10, 11], pre-trained BERT[12] etc.. In our front-end system, we input a sequence of characters and train the network in the end-to-end manner. Firstly, we get the polyphonic character in each sentence with its previous and following words as contexts. Then, we generate the POS tag sequence as token sequence, thus each feature vector includes a combination information of character itself, a polyphonic identity and a POS tag. Finally, these feature sequences are fed into a slightly optimized BiLSTM-CNN framework with 0.5 dropout rate has been used for training. The training sets including 255 polyphonic character and their corresponding sentences with different pronunciations about 530,000 in total.

3.1.2. Prosody Boundaries Prediction

Not only model can learn rhythm by text-audio pair, but also can learn it by prosodic break. The strength of the break is usually means distinguishing duration between characters, which classified as three-level symbols including prosodic words (PW), prosodic phrases (PPH) and intonation phrases (IPH)[13]. Therefore, a multi-task learning model have been proposed for addressing this issue through a Seq2seq framework which mainly consist of bidirectional LSTM network. When the labeled prosody training data are ready, we firstly recover then into original chinese sentences. In our cases, for the pre-trained model, 350,000 boundary-labeled training sentences[1] are pre-processed by the existing open-sourced toolkit PKUSEG[14], which provides word segmentation (each word are converted into 'BMES' labels[15]) and part-of-speech tagging. Then, we can easily extract the length of each word in sentence forward and backward. The last step of data processing is to align multiple semantic information including word segmentation, part-of-speech tagging, forward-backward word length are compressed into the fixed-length embedding. It is fed into a bidirectional LSTM encoder with 0.5 dropout rate. The output comes from the encoder is decoded through a RNN layer with a following single fully connected layer.

3.2. Attention-based Acoustic Model

Nowadays, most end-to-end Text-to-Speech systems have an encoder-decoder structure with attention, which is significantly helpful for alignment learning. Tacotron2[16] uses an auto-regressive attention[17] structure to predict alignment, with CNNs-LSTM-based architecture [18, 19]. The phoneme-level sequences of raw text is fed into an encoder which are converted into hidden states. Then the attention determines the alignment of each phoneme, from which the number of frames that attend on that phoneme can be induced. Then the features generated by attention module are fed in every step of the decoder.

In order to improve rhythm and pronunciation, we use polyphonic disambiguation and prosody prediction module to correct the Pinyin and generate prosody break in Mandarin Dataset. As Blizzard Challenge holder provides a 9.5 hours single-speaker Mandarin Dataset which is not enough for us to train a state-of-art model, our system modify the data preprocessing step and add a 512-dimension random initialized trainable speaker embedding which concatenate with the Encoder outputs for speaker identification (which is not shown in Figure 1.). Before training the provided Mandarin dataset, we pre-trained a four-speaker Tacotron2 model (which the datasets includes over 15 hours bilingual text-audio pairs for each speaker) for fine-tuning on Blizzard data. Finally, the audio is reconstructed by the Neural Vocoder introduced in the following part. Because of time constraints, for dealing with Shanghainese Dataset, we only train the model in the normal way without fine-tuning and any other methods.

3.3. WaveRNN Neural Vocoder

Due to the lack of phase prediction and the artifacts caused by traditional vocoders like Griffin-Lim[3] and STRAIGHT[20] which face great difficulties in producing high fidelity speech, we use neural WaveRNN as our vocoder for waveform generation. WaveRNN is a simple and powerful recurrent network for the sequential modeling of high fidelity audio. It matches the

[1] Dataset belongs to Ximalaya®

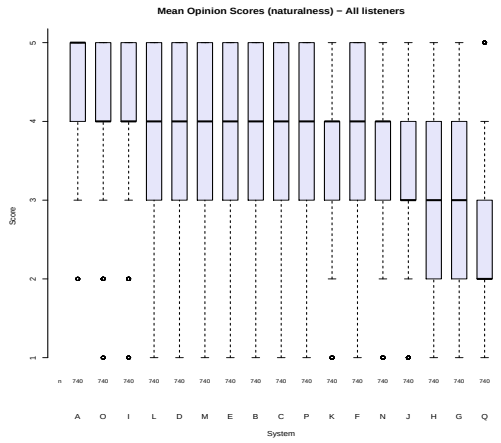

Figure 3: *Mean Opinion Scores (Naturalness ratings). System M is the proposed system and A is the natural speech in MH1.*

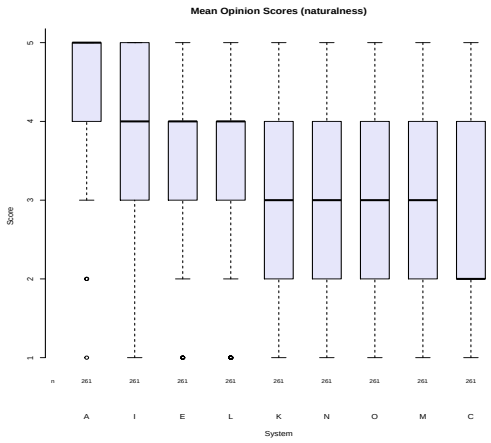

Figure 4: *Mean Opinion Scores (Naturalness ratings). System M is the proposed system and A is the natural speech in SS1.*

quality of the latest WaveNet model, but has much lower costs. The overall computation in the WaveRNN is as follows:

$$
\begin{aligned}
x_t &= [c_{t-1}, f_{t-1}, c_t] \\
u_t &= \sigma(R_u h_{t-1} I_u^* x_t) \\
r_t &= \sigma(R_r h_{t-1} + I_r^* x_T) \\
e_t &= \tau(r_t \cdot (R_e h_{t-1} + I_e^* x_t) \\
h_t &= u_t \cdot h_{t-1} + (1 - u_t) \cdot e_t \\
y_c, y_f &= split(h_t) \\
P(c_t) &= softmax(O_2 relu(O_1 y_c)) \\
P(f_t) &= softmax(O_4 relu(O_3 y_f))
\end{aligned}
\tag{1}
$$

where the * indicates a masked matrix whereby the last coarse input ct is only connected to the fine part of the states u_t, r_t, e_t and h_t and thus only affects the fine output y_f . The coarse and fine parts ct and ft are encoded as scalars in [0, 255] and scaled to the interval [-1, 1]. The matrix R formed from the matrices R_u, R_r, R_e is computed as a single matrix-vector product to produce the contributions to all three gates u_t, r_t and e_t. and are the standard sigmoid and tanh non-linearities.

In our system, the configuration details are almost the same as the origin WaveRNN. The model in mandarin datasets is trained in a multi-speaker Mandarin model way in order to improve fidelity but is trained in single-speaker datasets in SH1 as the limitation in Shanghainese datasets. In addition, noise is added into model to help model against outliers.

4. Results and Evaluation

In this section, we will discuss our evaluation results in detail. There are 17 system including 1 benchmark with 16 submitted systems evaluated in MH1, and 8 submitted systems evaluated in SS1. Our designated system identification letter is "M", whereas system "A" is the natural speech. The subjects who are involved in the listening test are speech experts, paid listeners and online volunteers and so on. We ranked the 5th in task-MH1 among 16 participating teams and ranked the 7th in task SS1. The content below will demonstrate the official evaluation results in Blizzard Challenge 2020 of two tasks respectively. In Task1 (MH1), we mainly focus on analysing the indicators of INT (Intelligibility of sentences), SIM (Similarity),

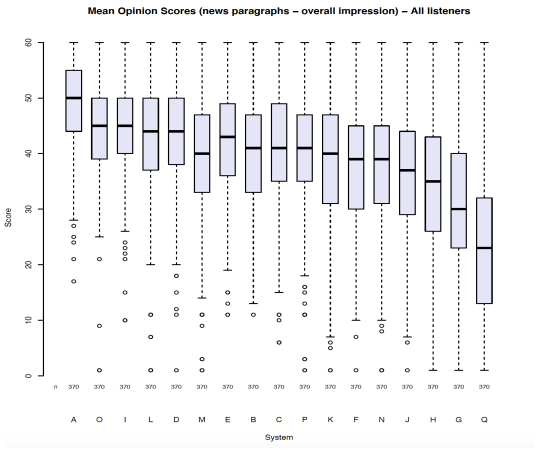

Figure 5: *MOS - various criteria. M Scores evaluated the proposed system as well as A Scores evaluated the natural speech in MH1. The scale of the score is from 1 to 60 for the following aspects (the scale of 1 to 50 is represented the results from bad to excellent. All the scores are evaluated based on MEAN scores.*

MOS (Mean Opinion Scores) and the overall impression. Besides, we will discuss the INT, SIM and Mos in Task2 (SS1).

4.1. Results for Mandarin and Shanghainese

4.1.1. MOS on sentences

Figure 3. and Figure 4. both show the boxplot evaluation of MOS on naturalness of all systems for MH1 evaluated on news and PSC sentences and SS1 evaluated on chat and news sentences respectively. In general speaking, difference of tasks is not significant among the participating systems ranking well. In task MH1, System A performs a high-quality natural speech from real person as a reference, and the score of our system is 3.9 which ranked the 5th among all the team. However, in task SS1, we ranked the 7th among 8 participants that we only got 2.7 points which need to be improved as we did not take any methods on low-resourced data in this task owing to the time constraints.

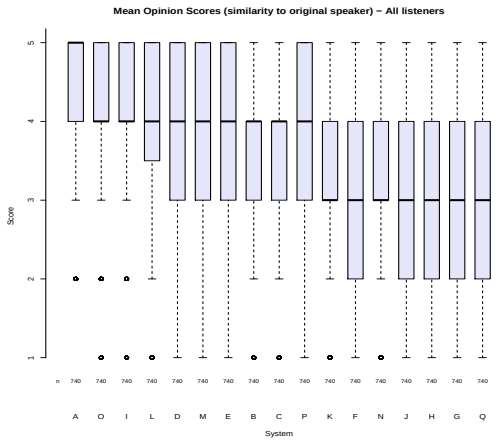

Figure 6: *Speaker similarity ratings. System M is the proposed system, A is the natural speech in MH1.*

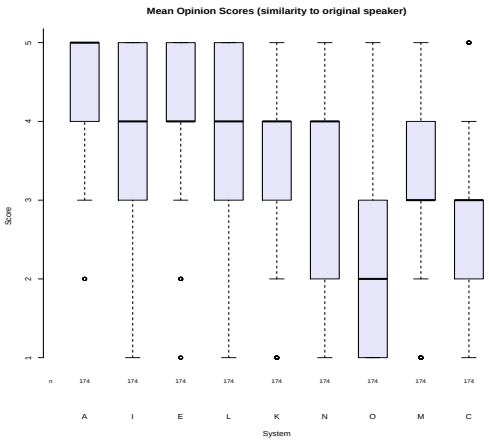

Figure 7: *Speaker similarity ratings. System M is the proposed system, A is the natural speech in SH1.*

4.1.2. MOS on paragraph

The results of the MOS tests on news paragraph sentences is consists of six section including emotion, intonation, listening effort, pleasantness, speech pauses, stress and overall impression are shown in Figure 5.

The results of speech pause and stress could comes from the prosody prediction module in the front-end network which need to be more domain-adaptive with more accurate hierarchical level. Moreover, although there are several methods we have tried, like VAE, β-VAE and some hierarchical generative model etc., that have been proven to be feasible on emotion and intonation control, we did not applied on the blizzard challenge as considering about the robust and stability of the model which could affect the quality of the results. Furthermore, pitch and duration model could benefit for intonation and pleasantness which could be applied and enhanced in the future. As we can see, our system have a moderate-level performance in MOS evaluation of various criteria. It indicates that the overall impression of our system which got marks 40 means that there is still a gap compared with natural speech which got 50 marks.

4.1.3. Similarity test

The two boxplot results of speaker similarity scores evaluated compared to the real speaker are showed in Figure 6. and Figure 7. We are slightly above the average among all participants as we achieved the scores of 3.8 and 3.3 in tasks 1 and 2 respectively.

4.1.4. Intelligibility test

In this part, the test sentences are derived from random combinations of Chinese phrases in the PSC. As it can be seen, Figure 8. displayed the evaluation results in this section. Compared with real speaker system A and system I and L who performance well in intelligibility test, we still have room for improvements. Besides, in comparison to system O who received the highest MOS score, we got 0.111 of the average PERT (Pinyin Error Rate with Tones) which shows the pronunciation prediction of our system is respectively accurate while maintaining good naturalness.

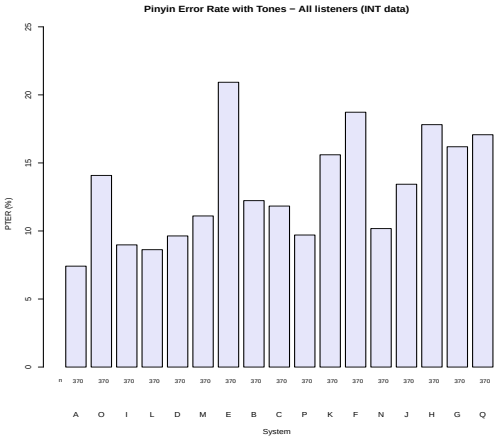

Figure 8: *Overall pinyin error rate with tones in MH1.*

5. Conclusions

In this paper, we present the details of our TTS system, including polyphonic disambiguation, prosody boundary prediction module, attention-based acoustic model and neural vocoder. We built an end-to-end speech synthesis system based on Tacotron2 and WaveRNN and investigate the effect of multi-speaker, transfer learning and front-end processing application. In general, the final evaluation results in both tasks indicates that our system has a middle performance slightly above the average. It seems that we still have much room for improvement, especially in the aspects of low-resourced data synthesize, style control, acoustic feature extracting, the front-end accuracy, speed and quality of vocoder and achieve better performance in all criterion in the future.

6. References

[1] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

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

[3] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast griffin-lim algorithm," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[4] 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.

[5] A. v. d. 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 preprint arXiv:1609.03499*, 2016.

[6] Fangzhou Liu, Qin Shi, and Jianhua Tao, "Tree-guided transformation-based homograph disambiguation in mandarin tts system," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4657–4660.

[7] R. T. Tsai and Yu-Chun Wang, "A maximum entropy approach to chinese grapheme-to-phoneme conversion," in *2009 IEEE International Conference on Information Reuse Integration*, 2009, pp. 411–416.

[8] F. Z. Liu and Y. Zhou, "Polyphone disambiguation based on maximum entropy model in mandarin grapheme-to-phoneme conversion," in *Materials Engineering for Advanced Technologies*, ser. Key Engineering Materials, vol. 480. Trans Tech Publications Ltd, 10 2011, pp. 1043–1048.

[9] X. Mao, Y. Dong, J. Han, D. Huang, and H. Wang, "Inequality maximum entropy classifier with character features for polyphone disambiguation in mandarin tts systems," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, 2007, pp. IV–705–IV–708.

[10] C. Shan, L. Xie, and K. Yao, "A bi-directional lstm approach for polyphone disambiguation in mandarin chinese," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2016, pp. 1–5.

[11] K. Park and S. Lee, "g2pm: A neural grapheme-to-phoneme conversion package for mandarinchinese based on a new open benchmark dataset," *ArXiv*, vol. abs/2004.03136, 2020.

[12] D. Dai, Z. Wu, S. Kang, X. Wu, J. Jia, D. Su, D. Yu, and H. Meng, "Disambiguation of Chinese Polyphones in an End-to-End Framework with Semantic Features Extracted by Pre-Trained BERT," in *Proc. Interspeech 2019*, 2019, pp. 2090–2094. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2292

[13] A. Li, "Chinese prosody and prosodic labeling of spontaneous speech," in *Speech Prosody 2002, International Conference*, 2002.

[14] R. Luo, J. Xu, Y. Zhang, X. Ren, and X. Sun, "Pkuseg: A toolkit for multi-domain chinese word segmentation." *CoRR*, vol. abs/1906.11455, 2019. [Online]. Available: https://arxiv.org/abs/1906.11455

[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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6689–6693.

[16] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017. [Online]. Available: http://arxiv.org/abs/1712.05884

[17] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[18] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[19] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.

[20] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.