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

NUS-HLT System for Blizzard Challenge 2020

Yi Zhou[1], *Xiaohai Tian*[1], *Xuehao Zhou*[1], *Mingyang Zhang*[1], *Grandee Lee*[1], *Rui Liu*[1], *Berrak Sisman*[2,1] *and Haizhou Li*[1]

[1] National University of Singapore
[2] ISTD Pillar, Singapore University of Technology and Design, Singapore

yi.zhou@u.nus.edu, eletia@nus.edu.sg, xuehao.zhou@u.nus.edu, elezmin@nus.edu.sg, {grandee.lee,r.liu, berraksisman}@u.nus.edu, {haizhou.li}@nus.edu.sg

Abstract

The paper presents the NUS-HLT text-to-speech (TTS) system for the Blizzard Challenge 2020. The challenge has two tasks: Hub task 2020-MH1 to synthesize Mandarin Chinese given 9.5 hours of speech data from a male native speaker of Mandarin; Spoke task 2020-SS1 to synthesize Shanghainese given 3 hours of speech data from a female native speaker of Shanghainese. Our submitted system combines the word embedding, which is extracted from a pre-trained language model, with the E2E TTS synthesizer to generate acoustic features from text input. WaveRNN neural vocoder and WaveNet neural vocoder are utilized to generate speech waveforms from acoustic features in MH1 and SS1 tasks, respectively. 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

Blizzard Challenges have been held since 2005 to promote the research techniques in building corpus-based speech synthesis systems and provide a common platform with the necessary data [1]. This year Blizzard Challenge has two tasks: 1. Hub task 2020-MH1 (MH1) to synthesize Mandarin Chinese given 9.5 hours of speech data from a male native speaker of Mandarin; 2. Spoke task 2020-SS1 (SS1) to synthesize Shanghainese given 3 hours of speech data from a female native speaker of Shanghainese.

Various techniques have been proposed for text-to-speech (TTS) [2, 3] generation: concatenative speech synthesis [4, 5] and statistical parametric speech synthesis [6, 7, 8] are widely studied in the past decades. For concatenative speech synthesis, small speech segments are selected from the database and then stitched to construct a synthesized speech. Although it is able to produce high quality synthesized speech, the boundary artifacts remain a key issue to be addressed. Statistical parametric speech synthesis is another popular method, which parametrizes speech signals into acoustic features. Different modeling approaches have been applied to map the text information to the acoustic features. In the statistical parametric method, a vocoder, such as STRAIGHT [9] and WORLD [10], is employed to construct waveform from the generated features by the acoustic model. While the prosody is predicted by the duration model. Owing to its high flexibility and the advance of deep learning modeling techniques, the parametric method gains adequate interest in current TTS research field.

Recently, the sequence-to-sequence (seq2seq) models are proposed [11]. They learn to align the input linguistic sequence to the acoustic representation through the attention mechanism, which effectively refrains from possible alignment errors. Besides, such frameworks own a unified, entirely neural network architecture, that is desirable to model the complex feature transformation in a simple and flexible way [12]. Several end-to-end (E2E) TTS systems, e.g., Tacotron [13], Char2Wav [14], DeepVoice [12, 15], have demonstrated their superiority over the conventional structures.

Research has shown that injecting linguistic information at the input step can help the model to better utilize the acoustic features [16]. Inspired by the previous work, we include word embedding that is obtained from a pre-trained language model alongside with acoustic features to enhance the model. Furthermore, by fine-tuning the general-domain language model on a smaller task-domain dataset, it can alleviate the low-resource problem [17] posed by the Shanghainese language. Apart from modeling the linguistic to acoustic feature mapping, the vocoder is the other important functional block in a TTS system. To improve the synthesized speech quality, neural vocoders like WaveNet [18] and WaveRNN [19] are adopted to replace a conventional parametric vocoder.

Based on the recent appealing approaches, we adopt the recent E2E TTS architecture [20] to predict the acoustic features, and we adopt WaveRNN [19] and WaveNet [18] neural vocoders to generate waveform in time-domain from the predicted acoustic features.

This paper is organized as follows: Section 2 describes our system implementation. Section 3 demonstrates and discusses the evaluation results. Last, Section 4 concludes this paper.

2. System Architecture

2.1. End-to-End TTS Synthesizer

Our system is illustrated in Figure 1. We base our model on Tacotron2 [20], a seq-to-seq model using recurrent neural network (RNN). It is composed of an encoder and an attention-based decoder.

The encoder aims to generate textual representations from input sequences. We convert character sequences into phone sequences for MH1, while we use the provided phone sequences for SS1. Each phone is represented as one phone embedding via the embedding layer. We concatenate the tone information by tone embedding, and phone embedding together. Then we pass the concatenated representation to 3 convolutional layers, followed by a bi-directional LSTM layer to generate the encoder output.

The encoder output is attended by the attention-based decoder, which is a RNN-based network that predicts acoustic frames using encoder output. The attention mechanism is to compute a fixed-length context vector to provide additional input to the decoder network. Location-sensitive attention is able

10.21437/VCC_BC.2020-7

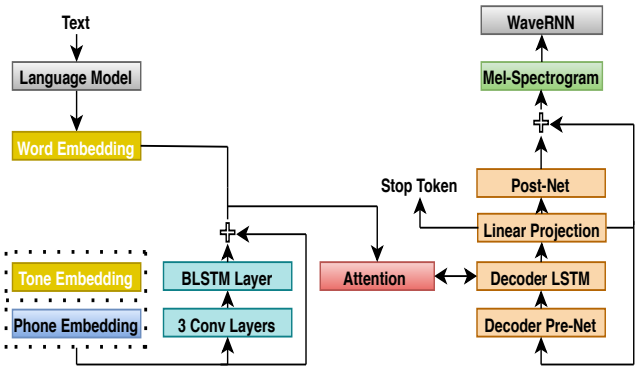

Figure 1: *An illustration of model architecture*

to reduce the frame prediction errors[21]. Guided-attention loss [22] is implemented to boost the alignment convergence. 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 post-net improves the general reconstruction. The stop token symbol is to predict when the model should stop during inference.

To provide an additional linguistic feature, which may benefit tokens sharing the same linguistic function to carry similar prosody, we utilize word embedding trained from large conversational domain corpora [23, 24]. The pre-trained language model [25] is able to process both English and Chinese text, of which only the Chinese word embedding is extracted from the embedding layers. To adapt to the task domain we fine-tune [26, 27] the language model using the task dataset for SS1, while we fine-tune the language model using [28] for MH1. For the MH1 task, we obtain the contextualized embedding from the 12 embedding layers of the language model. The concatenated embedding is projected into a 512-dim vector. For the SS1 task, since there is a lack of large Shanghainese text corpus for a proper fine-tuning, we adopt only the static embedding layer from the pre-trained language model, resulting in a 256-dim word embedding. These serve as an additional input to the end-to-end TTS model, similar to [29, 30, 31].

We adopt the residual encoder structure, proposed in [32], and we inject the word embedding to encoder output, the same as [31]. For MH1 task, the target acoustic feature is 80-dim log-mel spectrogram, extracted using 12.5ms frame shift and 4096-point Fourier transform, while 1024-point Fourier transform for SS1 task.

2.2. Neural Vocoder

For task MH1, WaveRNN [19] neural vocoder is utilized in our system owing to its efficiency in generating high-quality speech waveform. The WaveRNN is trained to map 80-dim mel-spectrogram input features to 10-bit waveform encoded by the μ-law. The WaveRNN model is mainly composed of a pre-processing network, two GRU layers, and an output layer. The preprocessing network is to upsample the mel-spectrogram to match the time-domain speech waveform resolution with the factor of [6, 10, 10]. Two GRU layers and the output linear layer all have 512 hidden neurons. The batch size is 32, and the learning rate is set to $1e^{-4}$. We trained the network for 1,000 epochs. WaveRNN is adopted for this task due to its fast waveform generation capability as there are a number of samples to be synthesized to within limited challenge time.

While for task SS1, WaveNet [33] neural vocoder is employed. WaveNet vocoder is also used to generate raw audio waveform from the 80-dimensional mel-spectrogram. The WaveNet vocoder contains 24 dilated convolution layers, and the k-th layer had a dilation size of $2^{mod(k-1,6)}$, where $mod(\cdot)$ was the modulo operation. The output of the dilated convolution layers and that of the skip channel had 30 and 128 dimensions, respectively.

3. Result

3.1. Challenge Participants

In total, there are 16 teams submitted their results for the MH1 task, which are denoted from B to Q. While 8 of them, system C, E, I, K, L, M, N, and O, additionally participated in the SS1 task. For both tasks, System A is natural speech. Our system is indicated as E.

3.2. Evaluation metrics

Subjective listening tests were designed to perceptually evaluate the synthetic samples for all systems in both MH1 and SS1 tasks. For the SS1 task, three sets of experiments were conducted to evaluate the synthetic samples, including naturalness, similarity, and intelligibility. While for the MH1 task, there is one more set of experiments that is conducted to evaluate the naturalness of synthetic paragraph. The detailed results will be presented in the next sections.

3.3. Perceptual evaluation for MH1 task

3.3.1. Naturalness of sentence

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 3.9 with 1.08 standard deviation and ranks at 6th position. While, as a reference, the natural speech has a score of 4.7 with 0.65 standard deviations.

3.3.2. Naturalness of paragraph

These sets of experiments were conducted to evaluate the naturalness of the synthetic paragraphs. Listeners should choose scores for one whole paragraph in seven aspects, e.g. overall impression, pleasantness, speech pauses, stress, intonation, emotion, and listening effort, e.g. 10 means "bad", while 50 means "excellent".

Figure 3 shows the boxplot of overall impression scores for the synthetic paragraph. Our system obtains an average score of 43 with 9.3 standard deviation and ranks at 5th position. As a reference, the average score of natural speech is 49 with a standard deviation of 7.8.

3.3.3. Similarity

During the experiments, listeners were asked to judge how similar the synthetic speech sounded to the reference sample of natural recordings with a score, which is scaled from 1 (Sounds like a totally different person) to 5 (Sounds like exactly the same person).

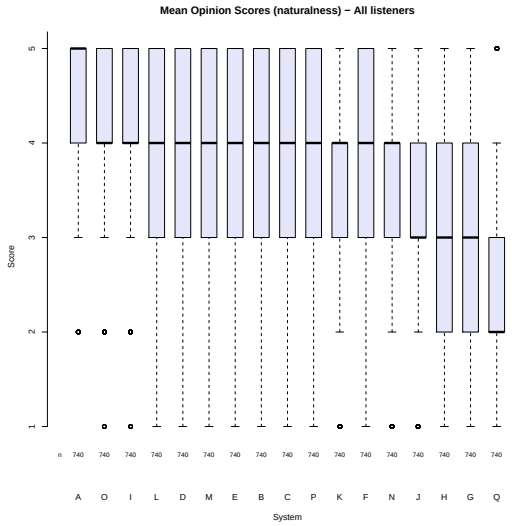

Figure 2: *Boxplot of naturalness scores of sentence synthesis for all listeners. E is our system.*

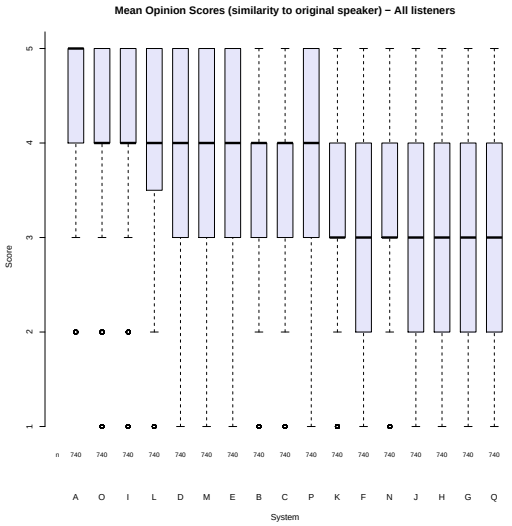

Figure 4: *Boxplot of similarity scores of each submitted system for MH1 task. E is our system.*

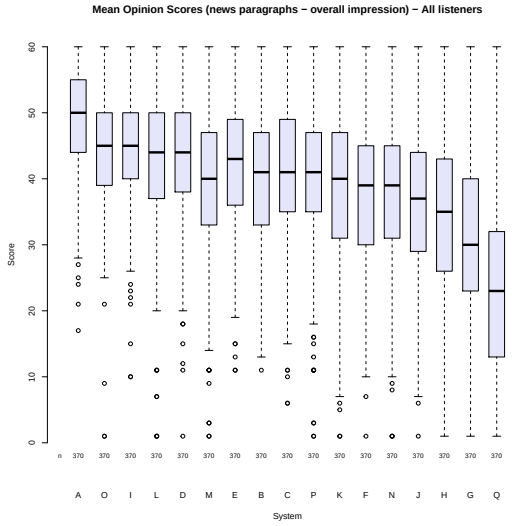

Figure 3: *Boxplot of overall impression scores of paragraph synthesis for MH1 task. E is our system.*

Figure 4 shows the boxplot of MOS of the speaker similarity. Our system obtains an average MOS of 3.8 with 1.16 standard deviation.

3.3.4. Intelligibility

The intelligibility evaluation of the MH1 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 Pinyin error rate with tones.

Figure 5 shows the Pinyin with tones error rate (PTER). It is observed that our system did not perform well in this evaluation with a 20.9% PTER and 0.21 standard deviation. As a reference, the natural speech obtains a 7.4% PTER and 0.13 standard deviation. An in-depth investigation could be considered as a research direction and a possible improvement in our future work.

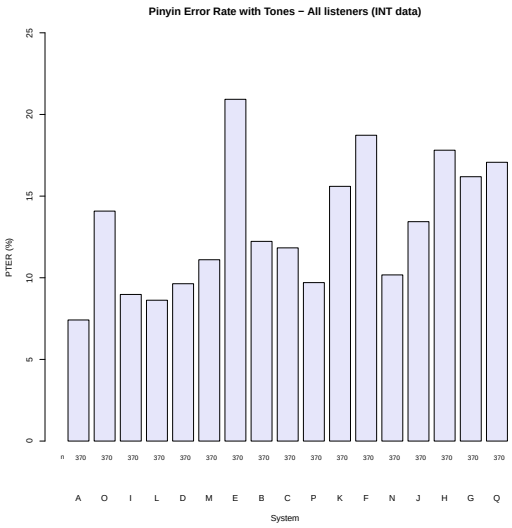

Figure 5: *Pinyin Error Rate with Tones for MH1 task. E is our system.*

3.4. Perceptual evaluation for SS1 task

3.4.1. Naturalness

A 5-scale MOS is used for naturalness evaluation of SS1 task, e.g. 1 indicates the speech sample is "Completely Unnatural" and 5 indicates that the speech sample is "Completely Natural". Figure 6 shows the boxplot of MOS of the naturalness for synthetic sentences. Our system achieves a mean MOS of 3.6 and a standard deviation of 1.06. This result is ranked at second place.

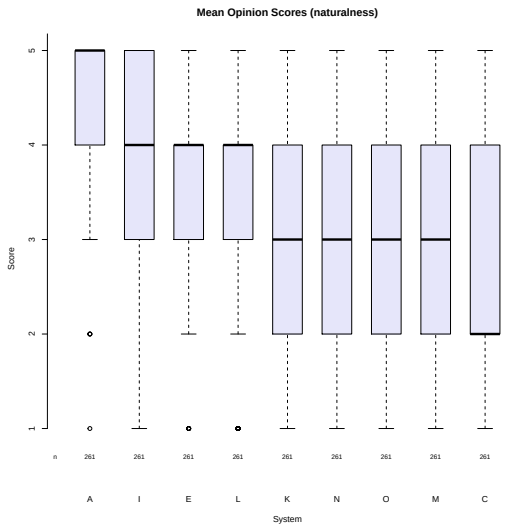

Figure 6: *Boxplot of naturalness scores of each submitted system for SS1 task. E is our system*

3.4.2. Similarity

A 5 scale MOS is used in similarity evaluation, where 1 means the synthetic sample sounds like an entirely different person, and 5 indicates the synthetic sample sounds like exactly the same person. Figure 7 shows the boxplot of MOS of the similarity for synthetic speech. It is observed our system outperforms all the submitted systems and obtains a mean MOS of 4.1 and standard deviation of 1.0.

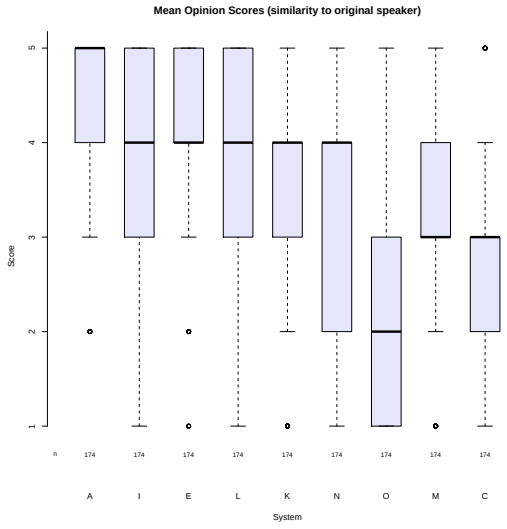

Figure 7: *Boxplot of similarity scores of each submitted system for SS1 task. E is our system*

3.4.3. Intelligibility

Different to the intelligibility evaluation of MH1 task, in this test, each listener should give a score to describe how intelligible or unintelligible of the speech sample. The score is scaled between 1 (Completely unintelligible) to 5 (Completely intelligible). Figure 8 shows the intelligibility scores for all submitted systems. Again, our system achieves a mean MOS of 4.0 and a standard deviation of 1.14. This result is just worse than the system "I" and ranked at second place.

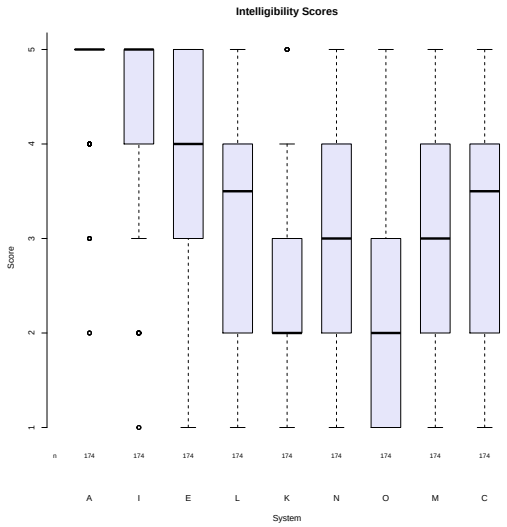

Figure 8: *Boxplot of intelligibility scores of each submitted system for SS1 task. E is our system*

4. Conclusions

This paper presents the NUS-HLT system submitted for Blizzard Challenge 2020. We built a TTS framework that first transforms text input to acoustic features combining with the word embedding and an end-to-end TTS synthesizer, followed by a neural vocoder to construct the audio waveform. The effectiveness of our system is successfully confirmed by the official evaluation results.

5. Acknowledgements

This work is supported by Human-Robot Interaction Phase 1 (Grant No. 19225 00054), National Research Foundation (NRF) Singapore under the National Robotics Programme; AI Speech Lab (Award No. AISG-100E-2018-006), NRF Singapore under the AI Singapore Programme; Human Robot Collaborative AI for AME (Grant No. A18A2b0046), NRF Singapore. Yi Zhou, Xuehao Zhou and Grandee Lee are also funded by the NUS research scholarship. Berrak Sisman is supported by SUTD Start-up Grant Artificial Intelligence for Human Voice Conversion (SRG ISTD 2020 158) and SUTD AI Grant, titled 'The Understanding and Synthesis of Expressive Speech by AI'.

6. References

[1] A. W. Black and K. Tokuda, "The blizzard challenge-2005: Evaluating corpus-based speech synthesis on common datasets," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[2] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[3] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *arXiv preprint arXiv:2008.03648*, 2020.

[4] 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.

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

[6] 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.

[7] H. Zen, 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.

[8] 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.

[9] 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.

[10] 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.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[12] S. O. Arik, 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," *arXiv preprint arXiv:1702.07825*, 2017.

[13] 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.

[14] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," 2017.

[15] 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," in *Advances in neural information processing systems*, 2017, pp. 2962–2970.

[16] X. Zhou, G. Lee, E. Yılmaz, Y. Long, J. Liang, and H. Li, "Self-and-mixed attention decoder with deep acoustic structure for transformer-based lvcsr," 2020.

[17] X. Yue, G. Lee, E. Yilmaz, F. Deng, and H. Li, "End-to-end code-switching asr for low-resourced language pairs," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2019. [Online]. Available: http://dx.doi.org/10.1109/ASRU46091.2019.9004035

[18] 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:1609.03499*, 2016.

[19] 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," in *ICML*, 2018, pp. 2415–2424.

[20] 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.

[21] 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.

[22] 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.

[23] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 2214–2218.

[24] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 2016, pp. 923–929.

[25] G. Lee and H. Li, "Modeling code-switch languages using bilingual parallel corpus," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 860–870. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.80

[26] G. Lee, X. Yue, and H. Li, "Linguistically Motivated Parallel Data Augmentation for Code-Switch Language Modeling," in *Proc. Interspeech 2019*, 2019, pp. 3730–3734. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1382

[27] G. Lee and H. Li, "Word and class common space embedding for code-switch language modelling," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6086–6090.

[28] G. Lee, T.-N. Ho, E.-S. Chng, and H. Li, "A review of the Mandarin-English code-switching corpus: SEAME," in *Asian Language Processing (IALP), 2017 International Conference on*. IEEE, 2017, pp. 210–213.

[29] 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)*. IEEE, 2019, pp. 6940–6944.

[30] 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.

[31] X. Zhou, X. Tian, G. Lee, R. K. Das, and H. Li, "End-to-end code-switching tts with cross-lingual language model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7614–7618.

[32] L. Xue, W. Song, G. Xu, L. Xie, and Z. Wu, "Building a mixed-lingual neural tts system with only monolingual data," *arXiv preprint arXiv:1904.06063*, 2019.

[33] 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 preprint arXiv:1609.03499*, 2016.