The OPPO System for the Blizzard Challenge 2020

Yang Song, Min Liang, Guilin Yang, Kun Xie, Jie Hao

OPPO, Beijing, P.R. China

songyang@oppo.com

Abstract

This paper presents the OPPO text-to-speech system for Blizzard Challenge 2020. A statistical parametric speech synthesis based system was built with improvements in both frontend and backend. For the Mandarin task, a BERT model was used for the frontend, a Tacotron acoustic model and a WaveRNN vocoder model were used for the backend. For the Shanghainese task, the frontend was built from scratch, a Tacotron acoustic model and a MelGAN vocoder model were used for the backend. For the Mandarin task, evaluation results showed that our proposed system performed best in naturalness, and achieved near-best results in similarity. For the Shanghainese task, we got poor results in most indicators.

Index Terms: speech synthesis, BERT, Tacotron, WaveRNN, MelGAN

1. Introduction

The text-to-speech (TTS) system has attracted the attention of many researchers, due to its importance in human-computer interaction. The statistical parametric speech synthesis (SPSS) [1] has become the most mainstream method of speech synthesis. SPSS is essentially composed of two parts, frontend and backend. In the frontend, linguistic features are extracted from the text input. In the backend, the acoustic model is used for transforming linguistic features to acoustic features, and the vocoder model is used for transforming acoustic features to audio signal. Recently, deep neural networks (DNNs) have been utilized as the acoustic model [2][3], and vocoder model [4][5] for TTS. Techniques for training DNNs to generate high-quality speech are widely studied.

Considering the rapid development of speech synthesis technologies, the Blizzard Challenge has been devised to tasks with more challenging data. In this year's challenge, the task is to build a speech synthesis system from Mandarin corpus about 9.5 hours and Shanghainese corpus about 3 hours.

For the Mandarin task, we developed the frontend based on BERT [6]. For the acoustic model and neural vocoder, we firstly trained a multi-speaker Tacotron2 [7] model and WaveRNN [8] model, and then the corpus provided by the organizer was used for adaptive training.

For the Shanghainese task, we built the Shanghainese grapheme-to-phoneme (G2P) module based on Shanghainese pronunciation dictionary. The module of word segmentation and prosody prediction remained the same as Mandarin. For the acoustic model, we used the corpus provided by the organizer to train a Tacotron2 model. For the neural vocoder, we trained a multi-speaker MelGAN [9] model, and then the corpus provided by the organizer was used for adaptive training.

2. TTS System

In this section, we will introduce the framework of our SPSS system. As indicated in Figure 1, our SPSS system consists of two parts, the training phase and the testing phase. We followed this flowchart and constructed our submitted system this year. A detailed description of the training and testing procedure will be presented as follows.

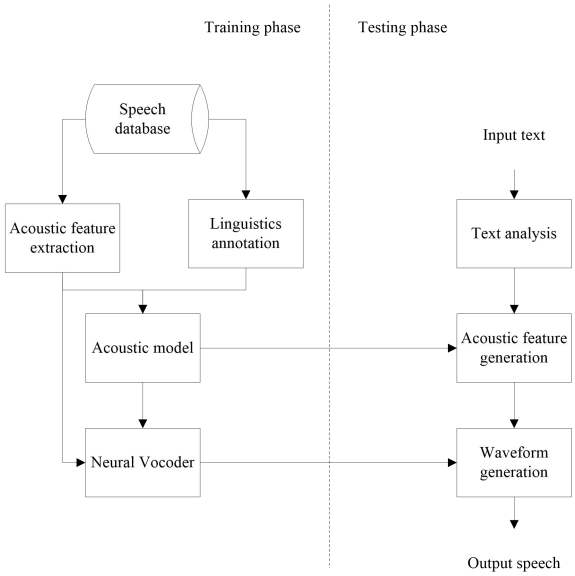

Figure 1: *Pipeline of our SPSS system.*

2.1. Data processing

For the Mandarin task, the data provided by the organizer are 4365 audio files at 48kHz sampling rate and the corresponding texts. For the Shanghainese task, the data provided by the organizer are 1900 audio files at 16kHz sampling rate and the corresponding phonetic transcriptions. First, we performed manual annotation, including Pinyin (with tone), prosodic word boundary and prosodic phrase boundary. Then, we extracted 80-dimensional Mel spectrogram.

2.2. Frontend

The goal of the frontend is to extract phonetic and prosodic information from the input text. Figure 2 shows the overview of our frontend. Frontend contains some components, including text normalization, word segmentation, part-of-speech (POS) tagging, prosodic boundary prediction and G2P. We used BERT to extract features for improving prediction

10.21437/VCC_BC.2020-3

accuracy. We designed the architecture of each module to fit specific tasks.

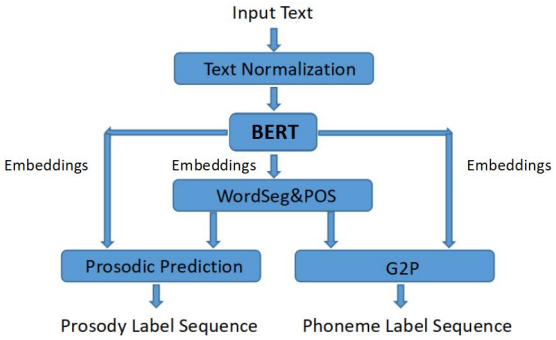

Figure 2: *Pipeline of our Frontend.*

Prosody structure is important for naturalness and intelligibility of speech. In Mandarin, a hierarchical prosodic structure [10], including prosodic word (PW), prosodic phrase (PPH), and intonational phrase (IPH), was widely employed to distinguish different levels of pauses in sentences.

For the task of prosody prediction, there are currently two basic modeling units. One is based on characters, and the other is based on words. Both methods have advantages and disadvantages. The character-based method is simple and flexible, and is not affected by the result of word segmentation. But when the amount of data is small, its performance may be worse than the word-based method. The word-based approach is more precise because words naturally carry certain prosodic boundary information. But it is easy to be affected by the result of word segmentation.

Inspired by WC-LSTM [11], we used the character-based method, and appended the word information to the corresponding characters at the same time. In this way, we retained the word information and alleviated the problem that the result was affected by the word segmentation result. Figure 3 shows the architecture of concatenated embedding. We still adopted hierarchical prosodic structure as our main approach. But in practice, we found that the probability of IPH appearing in the middle of the sentence is quite low, so we removed this part of the prediction of IPH and used the punctuations to determine it.

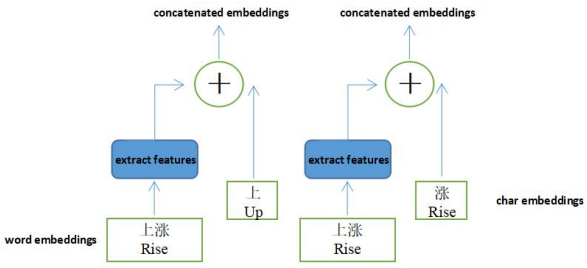

Figure 3: *The architecture of concatenated embeddings.*

For Shanghainese, we used word matching to determine the corresponding pronunciation based on our Shanghainese pronunciation dictionary. The module of word segmentation and prosody prediction remained the same as Mandarin. We used our own set of phonemes made up of smaller phone units, containing initial-consonants, medial-vowels, nucleus-vowels, and coda-consonants. There are totally 100 Shanghainese phonemes in our system.

2.3. Acoustic model

For both Mandarin and Shanghainese we used the Tacotron2 model for predicting the Mel spectrogram. GMM attention [12] with 8 mixtures was used for improving the robustness in synthesizing long sentences. At the same time, we used the guided attention loss [13] for the alignment.

For Mandarin, we built a multi-speaker model with five male Mandarin Chinese corpus and one male English corpus. The available data contained one Mandarin speaker about 13 hours, four Mandarin speakers about 5 hours, and one English speaker about 12 hours. In order to preserve the characteristics of each speaker, we added a 16-dimensional speaker embedding table, and concatenated the speaker embedding with the original output of encoder part to form new input for the decoder part. We used the converged multi-speaker model as the initial model, and then fine-tuned the model with the Mandarin corpus provided by the organizer.

For Shanghainese, we trained a Tacotron2 model only with the official data.

2.4. Neural Vocoder

In our Mandarin system, we adopted 10-bits μ-law WaveRNN for converting Mel spectrogram to waveform. A multi-speaker corpus about 25 hours was trained to get the initial model. The released corpus was used for updating the model. In order to eliminate the mismatch between training and inference, we used the ground truth aligned (GTA) [7] Mel spectrogram for fine-tuning our vocoder.

In our Shanghainese system, we chose MelGAN as the vocoder to generate the Shanghainese speaker's voice. We firstly trained an initial model with multi-speaker dataset, and then fine-tuned the model with the Shanghainese data. We found the performance of the original MelGAN was relatively poor, so we used the following methods for improving quality.

1. Expanding receptive field was helpful to improve the quality of speech generation and we tried 3, 4, and 5 residual layers. Finally, 5 residual layers were used.

2. Our generator loss and discriminator loss were the same as the original MelGAN, but we replaced the feature matching loss with multi-resolution STFT loss.

The work of Parallel WaveGAN [14] proved that STFT loss can make results more stable. A single STFT loss includes the spectral convergence and log STFT magnitude loss. By combining multiple STFT losses with different analysis parameters, it greatly helps the generator to learn the time-frequency characteristics of speech.

3. Results

For Mandarin evaluation, 17 systems including 16 submitted systems were evaluated. For Shanghainese evaluation, 9 systems including 8 submitted systems were evaluated. The identifier of natural speech is A, and our system is O.

3.1. Naturalness evaluation

Figure 4 shows the boxplot of evaluation results of all systems on Mandarin naturalness. Our system O and system I achieved the highest MOS of 4.2. But the variance of our system is smaller, indicating more stability and better consistency. Figure 5 shows the boxplot of evaluation results of all systems on Shanghainese naturalness. Our system's performance on Shanghainese naturalness is quite poor. There may be many semantic errors in our Shanghainese frontend. We did not use the official phonetic transcriptions, which may be the reason for our poor performance.

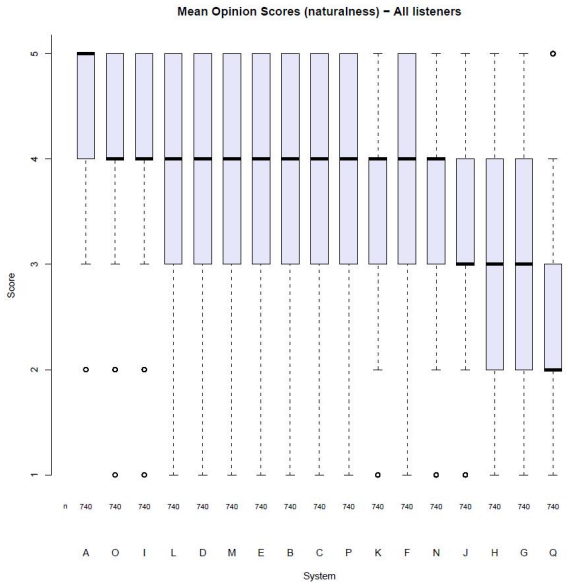

Figure 4: *Naturalness MOS (Mandarin).*

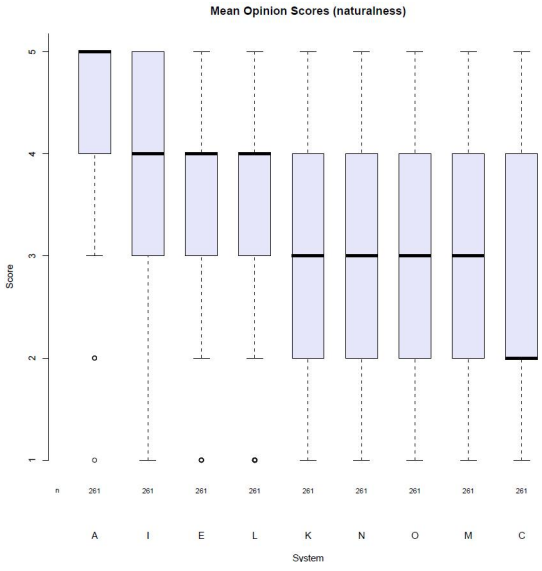

Figure 5: *Naturalness MOS (Shanghainese).*

3.2. Similarity evaluation

Figure 6 shows the boxplot of evaluation results of all systems on Mandarin similarity. Our system O achieved the MOS of 4.1. The gap between system O and system I is small. Figure 7 shows the boxplot of evaluation results of all systems on Shanghainese similarity. Our system's performance on Shanghainese similarity is quite poor.

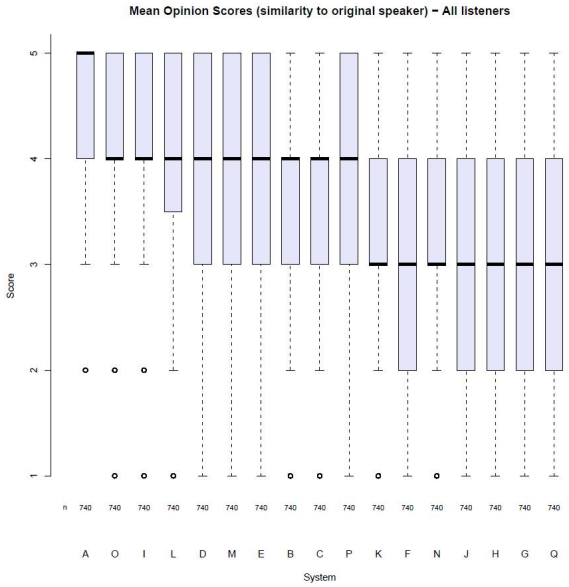

Figure 6: S*imilarity MOS (Mandarin).*

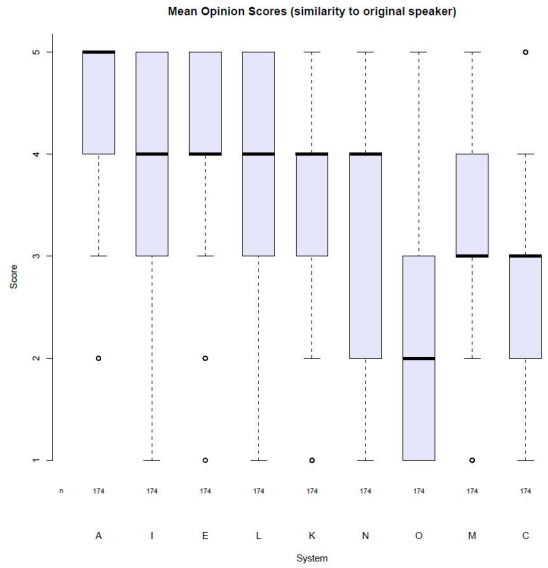

Figure 7: S*imilarity MOS (Shanghainese).*

4. Conclusions

This paper presents the details of our submitted system and summarizes the results in Blizzard Challenge 2020. In our system, the DNN-based frontend and backend models were used in order to achieve natural speech. From the results, we

believe that there is still substantial space for performance improvement in Mandarin and Shanghainese text-to-speech systems.

5. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," Speech Communication, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. M. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," IEEE Signal Processing Magazine, vol. 32, no. 3, pp. 35–52, 2015.

[3] 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: A fully end-to-end text-to-speech synthesis model," arXiv preprint arXiv: 1703.10135, 2017.

[4] 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," CoRR abs/1609.03499, 2016.

[5] W. Ping, K. N. Peng, and J. T. Chen, "ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech," arXiv preprint arXiv: 1807.07281, 2018.

[6] D. J, C. M.W, and e. a. Lee K, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[7] 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.

[8] 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.

[9] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Br'ebisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in Advances in Neural Information Processing Systems, 2019, pp. 14881-14892.

[10] Chu M, Qian Y, "Locating boundaries for prosodic constituents in unrestricted Mandarin texts," International Journal of Computational Linguistics & Chinese Language Processing, Volume 6, Number 1, February 2001: Special Issue on Natural Language Processing Researches in MSRA. 2001, pp. 61-82.

[11] Liu W, Xu T, Xu Q, et al., "An Encoding Strategy Based Word-Character LSTM for Chinese NER," 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. 2379-2389.

[12] A. Graves, "Generating sequences with recurrent neural networks," arXiv preprint arXiv:1308.0850, 2013.

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

[14] Yamamoto, Ryuichi, Eunwoo Song, and Jae-Min Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6199-6203.