The DeepSound Text-To-Speech System for Blizzard Challenge 2019

Jianjin Xiao[1], *Donghai Wu*[1], *Xu Wang*[2], *Boxian Huang**[1]

[1]DeepSound, Guangzhou, P.R. China
[2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
wangxu@szu.edu.cn, tts@deepsound.cn

Abstract

In the Blizzard Challenge 2019, about 8 hours of speech data from an internet talk show by a well-known Chinese character were provided to build a speech synthesis system. We introduce our proposed **DeepSound text-to-speech system** for Blizzard Challenge 2019, which employs the VQVAE as the backbone of proposed acoustic model, for its efficiency on style transfer learning. Specifically, we implement both manual and automatic tagging operations on the raw BC data for preparing the datasets. Then, the proposed Mandarin TTS front-end transforms the input text sequences into prosody labels with the form of typical three-layer structure. To increase the pronunciation accuracy and strengthen the emotion, a embedding+prenet operation is introduced in the proposed VQVAE end-to-end speech synthesis back-end, which enriches the non-linear representation ability from texts. Besides, to further improve the quality of synthetic sounds and reduce pronunciation errors and other problems, extra multi-speaker datasets are used for the data augmentation. Finally, the proposed robust multi-speaker neural vocoder generates the high quality waves. Although this is the first time for our team to take part in the challenge, we got the second place in terms of grade in the final evaluation, which achieves a high degree of naturalness and low Pinyin Error Rate.

Index Terms: Blizzard Challenge 2019, DeepSound, Speech Synthesis, Deep Neural Networks

1. Introduction

Blizzard Challenge (BC) has been held every year to evaluate Text-to-Speech (TTS) systems since 2005 [1]. In Blizzard Challenge 2019, the MH1 task is to build an expressive TTS system based on a 8 hours speech dataset from an internet talk show by a well-known Chinese character. Speech synthesis technology has continued to produce amazing results since the held of Blizzard Challenge.

Existing speech synthesis techniques can be classified into statistical parametric speech synthesis (SPSS) approaches, unit-selection based waveform concatenation methods, hybrid methods and end-to-end methods. The most representative hidden Markov models (HMMs) based SPSS method [2] was proposed in 1999, which simultaneously modelled the spectrum, fundamental frequency and segment duration within a unified HMM framework. To reduce the influence of the unfavourable training data, Yamagishi *et al.* proposed a robust training method [3] for the HMM based SPSS method. With the emergence development of neural networks, deep neural network (DNN) [4] and long short term memory (LSTM) [5, 6] were widely used to replace the HMM module in SPSS system. These modules provided significant improvements on both naturalness and intelligibility. For example, bidirectional LSTM (BiLSTM) was applied in merlin [7] for grapheme-to-phoneme (G2P) and prosody prediction. Recently, a neural network based autoregressive model, namely, WaveNet [8, 9] was proposed. It can synthesize sounds similar to real person. Variants of WaveNet such as Parallel WaveNet [10], WaveRNN [11] and Deep Voice 1 [12] were proposed, which achieves higher quailty audio compared with the traditional vocoder like World [13]. Recently, to bridge the gap between real and synthetic audio, the generating adversarial networks (GAN) was applied to speech synthesis [14]. Tacotron [15, 16], a new end-to-end speech synthesis system was proposed in 2017. It predicts the spectrum directly from phonemes, which goes beyond the traditional SPSS method. Combining with WaveNet, it achieves state-of-art result. In addition to SPSS-based and end-to-end methods, unit selection based waveform concatenation [17, 18] methods were also widely used. Benefiting from the direct use of natural speech segments, they achieve great advantage in similarity, quality and expression.

This is the first time for our team to participate in Blizzard Challenge. The vector quantization variational autoencoder (VQVAE)-based acoustic model [19] is employed as the backbone of our proposed DeepSound TTS system, which is robust on the unclean dataset and can improve the emotional expression ability. In our proposed DeepSound TTS front-end, the pronunciation rule-based regularization method is proposed to improve the performance of text normalization, then the BiLSTM-based recurrent network (RNN) [20] is used in the G2P module for polyphone and prosody prediction. To extraction acoustic features, the outputs of front-end are then transformed into a non-linear representation and fed into the VQVAE based back-end [21]. Finally, the RNN based speaker independent vocoder is employed to generate the audio. The competition results demonstrate that our system still has excellent performance on both emotional and stylistic datasets.

The remainder of this paper is organized as follows: Section 2 describes the proposed DeepSound TTS system in details. Section 3 describes the evaluation results and our system performance. Conclusion are given in Section 4.

2. DeepSound Text-to-Speech Synthesis System

The TTS system aims to take the text as input and output the synthesized wave. In this section, we introduce the detailed structure of our proposed DeepSound TTS system, as shown in Fig. 1. Specifically, our proposed system consists of three components, namely, the TTS front-end, VQVAE based back-end, and the vocoder module. These three modules fully couple together, ensuring the subjective quality of synthesized wave. During the phase of training, we use additional Mandarin dataset, including polyphone and prosody data to train TTS front-end module. The output of information from front-

*corresponding author: tts@deepsound.cn

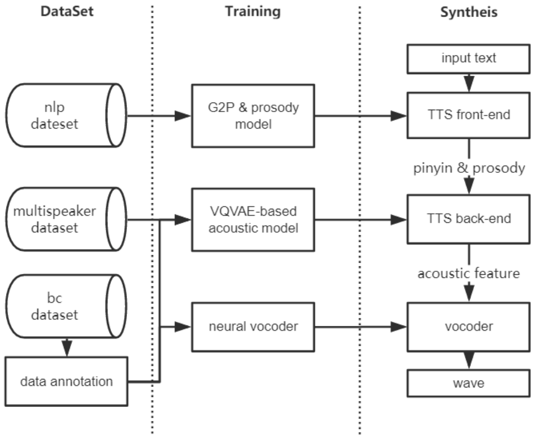

Figure 1: *Flowchart of proposed DeepSound TTS system.*

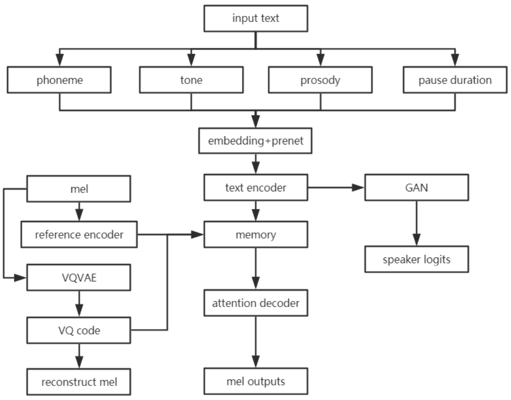

Figure 2: *VQVAE end-to-end speech synthesis back-end.*

end module is then fed into the following acoustic model. In addition, a multi-speaker dataset is employed to train the acoustic model and neural vocoder. During the phase of synthesis, neural vocoder generates audio from mel spectrograms produced by acoustic model. Detailed descriptions of our proposed system are provided as follows.

2.1. Data

2.1.1. Data description

Chinese data completing for 8 hours consists of 480 audio clips whose average duration is 1 minute. The source is from 60s talk show every morning and the host is a Chinese celebrity, called Luo Zhenyu. The audio sampling rate is $24khz$ and the quantization accuracy is 16 bits. They were recorded in a quiet room. Although official statistics contains the audio equivalent of the text but it is not accurate.

2.1.2. Data preprocessing

Based on our observation, the dataset released by the organizing committee contains audio clips and corresponding transcription. However, the transcriptions have not been manually checked. Meanwhile, the audio and text are inconsistent, *i.e*, audio consists of a number of colloquial sentences but text is formal. Thus, they can not be directly used for training our system. Furthermore, there is also a small amount of English in the data. To prepare the dataset for model training and validation, manual tagging methods such as text proofreading and noise processing are implemented on the raw BC data. Data annotation is a laborious and time-consuming task. Due to time constraints, we adopt some methods to speed up labeling, such as sentence level forced alignment using aeneas, phone level forced alignment using our automatic speech recognition (ASR) system. To ensure the stability of synthesized audio, all the raw audio clips are processed through noise reduction and volume normalization operation to filter out audio clips with low-quality.

2.2. TTS front-end

Our proposed Mandarin TTS front-end consists of three components, namely, text normalization, G2P and prosody prediction. Due to the complexity of mandarin language, the rule designing of G2P in our system is comparatively complex. For example, there are more than one hundred common polyphonic words that need to be predicted. Meanwhile, the characteristics such as qingsheng, sandhi and erhua, need to be considered in the system. During the implementation of proposed front-end system, the rule-based regularization method is proposed to improve the performance of text normalization sub-module. In our system, the typical three-layer prosody structure is employed as the label which consists of prosodic word (PW), prosodic phrase (PPH) and intonational phrase (IPH). The BiLSTM-RNN is adopted for polyphone and prosody prediction.

2.3. VQVAE end-to-end speech synthesis back-end

Because of the great differences between Mandarin and English, the simple text embedding method is difficult to efficiently represent the text feature of the input. To further improve the performance of back-end module, the text sequence is split into four parts, including phoneme, tone, prosody and pause duration, to enrich representations from texts. Each part is presented as an embedding respectively. The so-called ‘prenet‘ transforms each embedding to a non-linear representation. Then, these four representations are concatenated as inputs of encoder. We find these representations can increase the pronunciation accuracy, strengthen the emotion and improve pronunciation fluency.

During the training phase, multi-speaker datasets with varieties of tone and speech rate are used for boosting the speaker style representation ability. To fulfill the adaptively style representation of each speaker, we use reference encoder and VQ-VAE to extract a style vector. Reference encoder can extract high level information and VQVAE can stabilize the pronunciation. Besides, we observe that the generated audio may contain different timbre. We think that the network learn the style information from text. Therefore, an auxiliary GAN component is added to help disentangle the text information and style information.

2.4. Robust multi-speaker neural vocoder

The RNN based speaker independent vocoder is employed, which consists of conditional extraction module and autoregressive generative module. The concrete process of the vocoder is shown in Fig. 3. Conditional extraction module, composed by bidirectional Gate Recurrent Unit (GRU) [22], extracts local condition (hiding speaker’s information and acoustic informa-

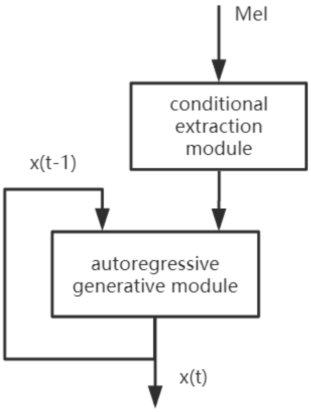
Mel
conditional extraction module
x(t-1)
autoregressive generative module
x(t)

Figure 3: *speaker independent vocoder.*

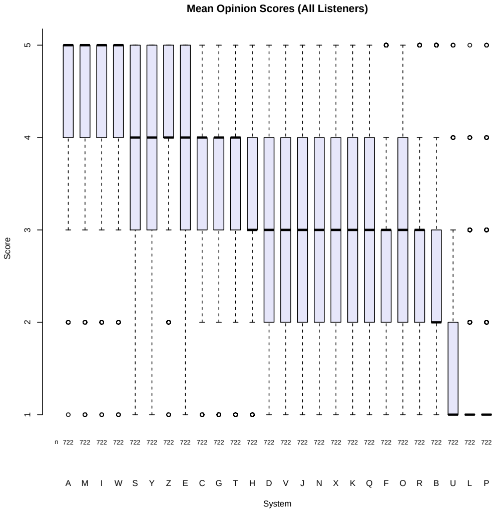

Figure 4: *MOS based on all the listeners' responses.*

tion) from mel spectrograms. Local condition after sampling is fed into the autoregressive generative module. Autoregressive generative module, composed by unidirectional GRU and DNN, generates a next sampling point $x(t)$ according to the local condition and the sampling point $x(t-1)$ of the preceding moment. For the purpose of balancing the model in the speed of convergence and the quality of synthesized audio, we implement a 10-bits μ-law quantification of the audio in advance. At last, the denoise operation is implemented on the synthesized audio to suppress white noise. As a result of comparatively poor quality of data provided by the challenge, we use multi-speaker data to train the vocoder to further improve the quality of synthesized audio.

3. Evaluation results

Blizzard Challenge officially provided evaluation results of 26 systems, including 24 systems submitted by participating teams and a benchmark system based on merlin and a natural speech. The 26 letters respectively represent these 26 systems, *e.g.*, **A** means Natural Speech, **B** means Merlin Benchmark and **I** is our system. The result shows that the overall performance of our system is only inferior to Natural Speech **A** and the participating team **M**, namely the second place in all participating teams. There are four indices of evaluation results, which are Naturalness, Similarity, Pinyin Error Rate (PER), Pinyin with Tone Error Rate (PTER). The MOS and Similarity results are based on all the listeners' responses, including paid listeners at Edinburgh, volunteers and experts. The PER and PTER are mainly based on paid listeners' responses.

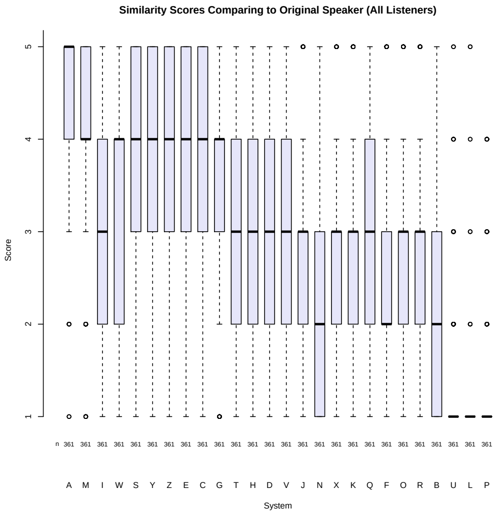

Figure 5: *Similarity based on all the listeners' responses.*

- **A**: Natural speech
- **B**: Merlin benchmark
- **I**: DeepSound TTS system

3.1. Naturalness test

Fig. 4 shows the boxplot of evaluation results of all systems on naturalness. As a reference, there is a very high quality pronunciation from the real person with 4.7 points in average naturalness of the original audio and the score of our system is 4.3. Although there is still a small gap compared with the real person, it has reached a level that is difficult to distinguish between original and system synthesis.

3.2. Similarity test

Speaker similarity scores are presented in Fig. 5. We only got 3.3 points in similarity, because our system adopted multi-speaker vocoder technology making the synthesized sounds deviate significantly from the real person. That is a crucial part that we need to improve in the future.

3.3. Pinyin Error Rate (PER) test

Pinyin Error Rate is shown in Fig. 6. The average error rate of our system is 0.092, performing well in all participating teams. Compared with high MOS performance System **M** and System **W**, the error rate of our system is much lower, which means the pronunciation of our system is much more clear and accurate while keeping high naturalness. This advantage comes from the accurate conversion results from the TTS front-end and our robust acoustic model.

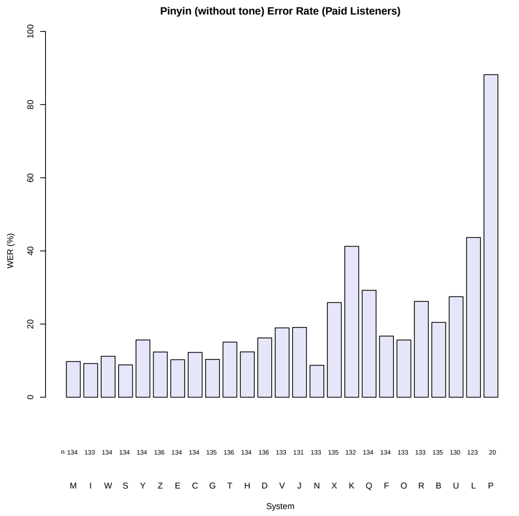

Figure 6: *Pinyin Error Rate (PER) based on paid listeners' responses.*

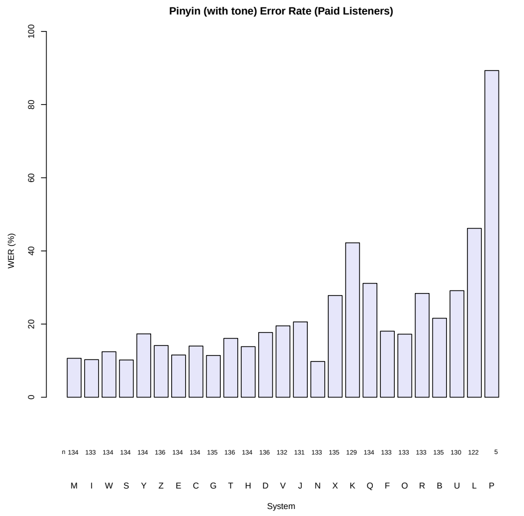

Figure 7: *Pinyin with Tone Error Rate (PTER) based on paid listeners' responses.*

3.4. Pinyin with Tone Error Rate (PTER) test

Pinyin with Tone Error Rate is shown in Fig. 7. It is almost the same as PER, so we won't go into details here.

4. Conclusions

This paper presents DeepSound text-to-speech system for Blizzard Challenge 2019. The BiLSTM based model is used in mandarin front-end to predict polyphone and prosody. We introduce VQVAE to back-end model to generate more stable and natural speech. Then GAN based postfiltering are used to compensate for the differences between natural and synthetic spectrum. Finally, robust multi-speaker neural vocoder conditioned on the mel spectrograms is used to obtain high quality audio. Evaluation results demonstrated the effectiveness and superiority of our system.

5. References

[1] K. Tokuda, Alan W Black, "The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. Interspeech*, 2005.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[3] J. Yamagishi, Z. Ling, and S. King, "Robustness of hmm-based speech synthesis," in *Proc. Interspeech*, 2008.

[4] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2013, pp. 7962–7966.

[5] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

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

[7] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system." in *SSW*, 2016, pp. 202–207.

[8] 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," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[9] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder." in *Proc. Interspeech*, 2017.

[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*, 2018, pp. 3915–3923.

[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*, 2018, pp. 2415–2424.

[12] S. Ö. 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," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 195–204.

[13] 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.

[14] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4910–4914.

[15] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017.

[16] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[17] Y. Qian, F. K. Soong, and Z.-J. Yan, “A unified trajectory tiling approach to high quality speech rendering,” *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 2, pp. 280–290, 2012.

[18] L.-J. Liu, C. Ding, Y. Jiang, M. Zhou, and S. Wei, “The iflytek system for blizzard challenge 2017,” in *The Blizzard Challenge 2017 Workshop, Stockholm*, 2017.

[19] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, “Learning latent representations for style control and transfer in end-to-end speech synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6945–6949.

[20] 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 *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 98–102.

[21] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, “Forward attention in sequence-to-sequence acoustic modeling for speech synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4789–4793.

[22] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, and V. Aggarwal, “Towards achieving robust universal neural vocoding,” in *Proc. Interspeech*, 2019.