SZ-NPU Team's Entry to Blizzard Challenge 2019

Shan Yang, Wenshuo Ge, Fengyu Yang, Xinyong Zhou, Fanbo Meng, Kai Liu, Lei Xie

Northwestern Polytechnical University, Xian, China, 710129

{syang, xyzhou}@nwpu-aslp.org, y522659913@gmail.com, liukaios3228@sogou-inc.com, lxie@nwpu.edu.cn

Abstract

In this paper, we introduce the entry from the SZ-NPU team submitted to Blizzard Challenge 2019. The goal of this year challenge is to build a natural Mandarin Chinese speech synthesis system from 8-hours single-speaker stylistic speech data in talk shows. We will discuss the major modules of the submitted Tacotron-Wavenet system: (1) The front-end module to analyze the pronunciation and prosody of text; (2) The GMM-attention based sequence-to-sequence acoustic model to predict speech features; (3) The Wavenet based neural vocoder to reconstruct waveforms; (4) A bandwidth extension module to up-sample the generated speech. Evaluation results provided by the challenge organizer are also discussed.

Index Terms: speech synthesis, end-to-end, wavenet

1. Introduction

Text-to-speech (TTS) has achieved significantly improved performance from hidden Markov models (HMMs) to neural networks (NNs). The HMM-based frameworks use the Gaussian mixture model (GMM) to model the hidden states of speech observations [1, 2, 3]. Considering the limitations of the HMM-GMM framework [4], deep neural networks (DNN) are applied to acoustic modeling and produce much better performance [4, 5, 6]. Based on the DNN framework, more novel architectures or variants has come into the NN family to improve the performance of synthesized speech [7, 8]. Note that the traditional frameworks need an extra module to align the linguistic and acoustic representations, and the inaccurate align errors may propagate to the latter synthesis model [9]. The attention-based sequence-to-sequence (seq2seq) models [10, 11] have been proposed to migrate from this problem and thus simplify the pipeline of traditional systems. Several end-to-end speech synthesis systems have recently shown superior performance over the conventional structures [12, 13, 14, 15, 16, 17].

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 for a well-known Chinese talk-show speaker, Zhenyu Luo, using about 8 hours real talk-show recordings given by the organizer. Thus the data is stylistic and imperfect with inevitable background noise.

Our submission to the challenge is based on the seq2seq acoustic model [18], where an independent Wavenet is adopted as vocoder to reconstruct waveforms [19, 20]. The original Tactoron2 system was firstly proposed in [18], which achieved satisfactory performance directly from simple character- or phoneme-level text representation for English language. But for Chinese, G2P, word boundary and prosodic boundary directly influences the intelligibility and naturalness of the synthesized speech [21]. Hence we still use a text analyzer in our system. Besides, we adopt a GMM-based attention mechanism [22] to stabilize the generating process for long sentences and such a system can correctly generate speech from a sequence of hundreds of Chinese characters. Finally, to further improve the perceptual quality of the generated speech, we apply a bandwidth extension method to up-sample the reconstructed waveforms.

2. System Description

2.1. Data processing

The data provided by the organizer are 480 audio files at 24 kHz sampling rate and the corresponding texts. Note that the audio and the text do not exactly match: there are many colloquial words in the audios, but not in the texts. So we correct all the texts based on the audios. We use Sogou TTS front-end to predict the needed information: normalized text, phones and tones, word boundaries and prosodic boundaries. For the English content occasionally embedded in the text, we obtain the English phonemes according to the CMU dictionary. At last, in order to facilitate subsequent training, we divide the long sentences into several short segments. We also extract 80-dimensions Mel spectrogram at 24 kHz using a 50ms frame length, 12.5ms frame hop and a Hann window function, which is used to train the acoustic model and as local condition for the WaveNet model.

2.2. Text analyser

In our system, the front-end module mainly contains text normalization, Chinese word segmentation, part-of-speech (POS) tagging, polyphone prediction and prosodic boundary prediction.

For text normalization, regulation rules are applied to match the given context. According to the matched rules, we convert all symbolic chars into Chinese characters. As for word segmentation, we train a Bi-GRU model to predict the word boundaries from the normalized texts. Besides, an extra user-defined dictionary is adopted to handle some special words like named entities. Specifically, we substitute the continuous chars from the word dictionary with one tag, and then feed these tags and other left chars into the model for prediction. For polyphone prediction or G2P, both a phonetic dictionary with thousands of words and a maximum entropy model are used. We directly use the phonetic information from the dictionary if the word is included in the dictionary, in which way we can easily enroll new words if needed. Otherwise we use a maximum entropy model to do polyphone prediction. We train a maximum entropy model for each polyphone character, resulting in about one hundred models in total. Finally for the prosodic boundary prediction, we predict two-level boundaries using a conditional random fields (CRF) model. The inputs of the CRF model include the current word context, POS tags and the numeric features of the word context. Thousands of annotated sentences are used to train the model.

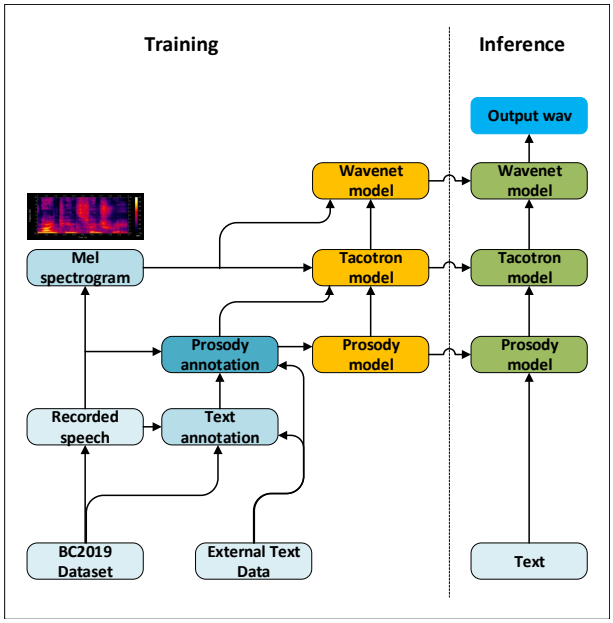

Figure 1: *The architecture of our system.*

2.3. Seq2seq acoustic model

In our system, we adopt the attention-based seq2seq framework as acoustic model to predict Mel spectrogram from the text representations. Similar to the common seq2seq models, our system contains a text encoder, a GMM attention module and an auto-regressive decoder. Figure 1 shows the building blocks of our system.

For the encoder part, we adopt the CBHG network as described in Tacotron [14], which is also proved to be effective in traditional statistical parametric speech synthesis [23]. Since the attention module is the key factor which directly affects the stability of the end-to-end system, we investigate different attention mechanisms, including location-sensitive [24], guided-attention [25] and GMM attention. Finally we choose the GMM-based attention [22, 26] in our system as it shows robustness in generating particularly long sentences. In practice, we use a GMM with 10 mixtures.

As for the decoder part, we follow the architecture of Tacotron2 [18] to predict Mel spectrogram from the encoded input sequence one frame at a time. The auto-regressive decoder contains 2 fully connected layers of 256 hidden ReLU units as pre-net, which is essential for learning attention [18]. And a sub-network of 2 uni-directional LSTM layers with 1024 units is adopted to fuse the information from the pre-net output and the attention context vector. Then the LSTM output is concatenated with the attention context to predict Mel spectrogram. Besides, an extra post-net is also applied to predict a residual to the prediction of foregoing sub-networks.

2.4. Wavenet-based vocoder

We train a Wavenet model to reconstruct waveform from the predicted Mel spectrum. Wavenet is a fully probabilistic and auto-regressive generative model that can generate audio samples directly:

$$p(x|h) = \prod_{t=1}^{T} p(x_t|x_1, x_2, ..., x_{t-1}, h), \quad (1)$$

where $x = x_1, ..., x_T$ is a given audio sample sequence and each audio sample x_t is conditioned on the samples at all previous timesteps. And h is the conditional inputs. Here we use predicted Mel spectrogram as local conditions.

The architecture of the Wavenet model is very similar to our system in Blizzard Challenge 2018 [27]. The entire model consists of 40 layers, grouped into 4 dilated residual block stacks of 10 layers. In every stack, the dilation rate increases by a factor of 2 in every layer, and no dilation for the first layer. The predicted Mel spectrograms are passed through a stack of 2 bidirectional QRNN layers with 256 units. And the local conditions are up-sampled to match the frequency of waveform. To overcome the distribution mismatch between the ground-truth spectrogram and the predicted spectrogram, we generate all the training data using teacher-force mode. And the generated spectrogram is subsequently used to fine-tune the pre-trained Wavenet vocoder.

2.5. Frequency band extension

We find that the real sampling rate of the audio is 21kHz, although the audio files are given in 24KHz sampling rate. To further improve the quality of the synthesized speech, we expand the sample rate of the synthesized speech to 32KHz through a bandwidth extension module.

Figure 2 illustrates the procedure of bandwidth extension. We first generate 32 KHz random noise signals from a normal Gaussian distribution. And then we filter the noise signals with a high pass filter, leaving the signal of which the frequency energy below 11KHz is zero. After predicting Mel spectrum using the acoustic model, we convert the predicted Mel spectrum to linear spectrum and adjust the amplitude of the filtered signals according to the average energy between 10kHz and 10.5 kHz of the linear spectrum. Finally we convert the sample rate of the synthesized speech to 32 kHz from 24 kHz and add the adjusted signals into it.

3. Results

In this year's challenge, there are 26 systems in total including natural speech (system A) and Merlin baseline (B). Our submitted system is annotated as C. There are four criteria in the evaluation: Mean Opinion Score (MOS), speaker similarity, Pinyin Error Rate (PER) and Pinyin+Tone Error Rate (PTER). We will discuss the details as follows.

3.1. MOS evaluation

The MOS test results are based on all the listeners' responses, including paid listeners, volunteers and experts. Figure 3 shows the MOS results of all systems.

From the naturalness MOS results, as expected, the original natural speech achieves the highest score of 4.7. System M achieves the highest score of 4.5 among all the submitted systems, which is very close to the natural speech. The MOS of our system is 3.8. Analyzing the results, we guess that the main influence factors are the performance of the vocoder and the prosody of generated speech.

3.2. Similarity evaluation

Similarity tests are carried out by the organizer, in which listeners are asked to judge whether the generated speech is similar to the target speaker. The speaker similarity MOS results collected from all listeners are shown in Figure 4.

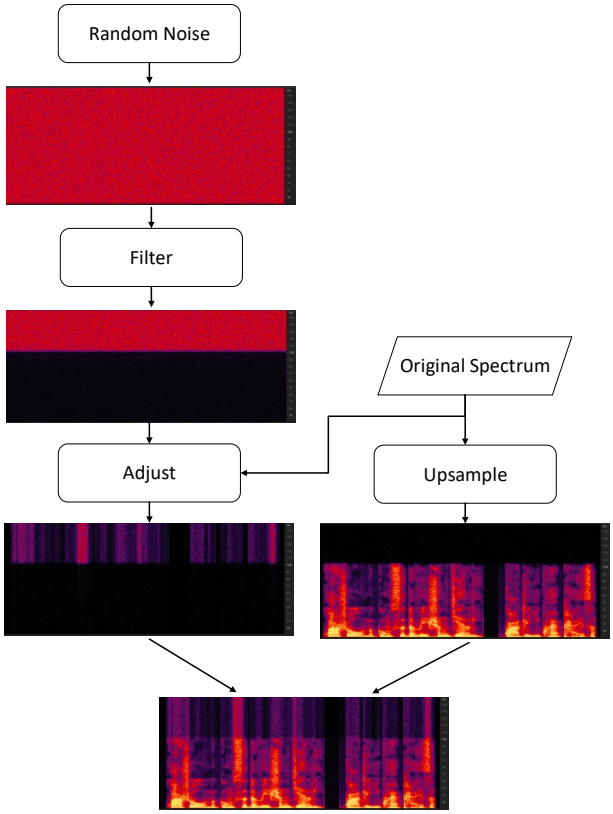

Figure 2: *The architecture of band extension.*

From the similarity results, we can find our system achieves a good speaker similarity. As mentioned earlier, this year's challenge only provides a single speaker dataset. To ensure speaker similarity, we only use the provided data to train the acoustic model and the Wavenet vocoder. We don't use any external speech data to train the acoustic model. Hence we believe that the vocoder performance directly affects the speaker similarity. Meanwhile, the generated prosody from the acoustic model may also have a clear influence on the perceived speaker similarity by the listeners. We also find that System I and W achieve very good MOS in terms of naturalness, but their similarity scores are not so good. This is probably because some external data are used to help to build the speech synthesis system for the target speaker, which affects the speaker similarity.

3.3. PER evaluation

Besides the MOS evaluation of naturalness and similarity, listeners are also asked to transcribe the generated speech, i.e., intelligibility evaluation. Figure 5 shows the pinyin error rate (PER) of each system.

In this evaluation test, the PER of our system is 12.3%. Since we use a front-end module, the performance of our G2P module directly affects this result. Besides, there also may exist some pronunciation errors caused by the seq2seq acoustic model. We can improve the intelligibility of generated speech from these two aspects in future.

3.4. PTER evaluation

Different from English, the pronunciations of Chinese characters are also affected by the tones, so the Pinyin+Tone Error Rate (PTER) is calculated from the listeners' transcripts. The results are shown in Figure 6.

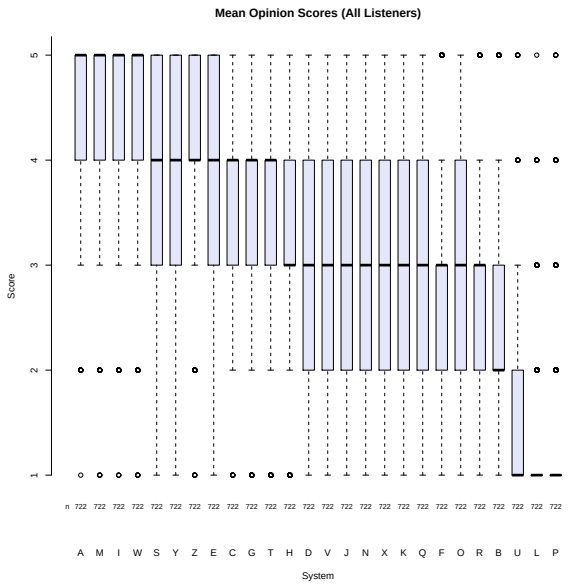

Figure 3: *Naturalness mean opinion score of each system.*

The results indicate that PER and PTER of each system have very similar trend, where PTER is a little bit higher than PER. We guess that all systems suffer from the tone errors. It may be caused by the phonetic annotation process, such as the polyphone disambiguation module. Besides, we also find that the acoustic model may generate a few wrong tones. Through internal experiments, we find that using separate representation of tones can alleviate this problem especially for limited training data.

4. Conclusions and future work

This paper presents the details of our submitted system and summarizes the results in Blizzard Challenge 2019. In our system, attention-based seq2seq model and neural vocoder are used in order to achieve natural and high-fidelity speech. From the results, we believe that there is still substantial space for performance improvement in building speech synthesis systems from stylistic wild data.

5. References

[1] 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.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1315–1318.

[3] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1229.

[4] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and*

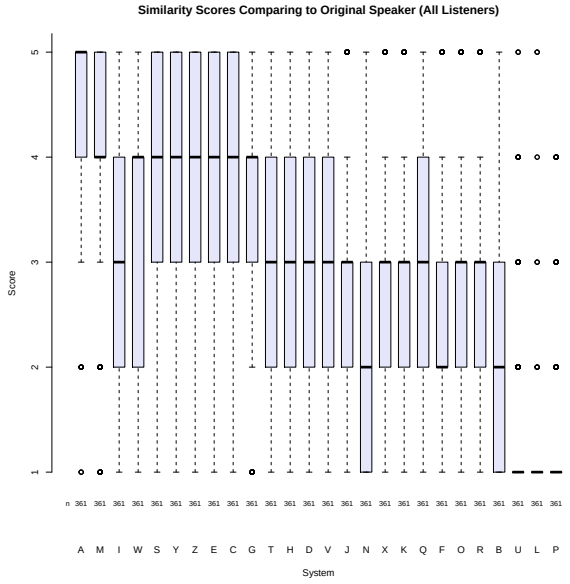

Figure 4: *Mean opinion score for speaker similarity of each system.*

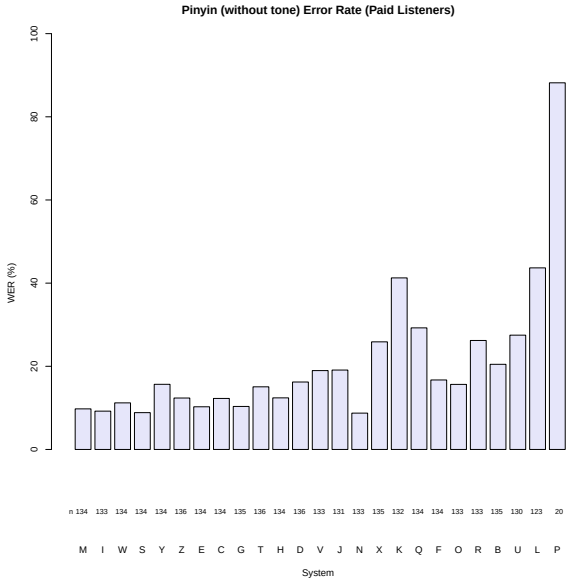

Figure 5: *Pinyin Error Rate of each system.*

Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 7962–7966.

[5] 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.

[6] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, "From HMMs to DNNs: where do the improvements come from?" in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5505–5509.

[7] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3844–3848.

[8] 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.

[9] M. Li, Z. Wu, and L. Xie, "On the impact of phoneme alignment in dnn-based speech synthesis," in *9th ISCA Speech Synthesis Workshop, Sunnyvale (USA)*, 2016, pp. 196–201.

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014, pp. 3104–3112.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] W. Wang, S. Xu, and B. Xu, "First Step Towards End-to-End Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention," in *Proc. INTERSPEECH*, 2016, pp. 2243–2247.

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

[14] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, 2017, pp. 4006–4010.

[15] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *arXiv preprint arXiv:1710.07654*, 2017.

[16] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality tts with transformer," *arXiv preprint arXiv:1809.08895*, 2018.

[17] S. Yang, H. Lu, S. Kang, L. Xie, and D. Yu, "Enhancing hybrid self-attention structure with relative-position-aware bias for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6910–6914.

[18] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *arXiv preprint arXiv:1712.05884*, 2017.

[19] 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.

[20] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. INTERSPEECH*, vol. 2017, 2017, pp. 1118–1122.

[21] C. Lu, P. Zhang, and Y. Yan, "Self-attention based prosodic boundary prediction for chinese speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7035–7039.

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

[23] H. Li, Y. Kang, and Z. Wang, "Emphasis: An emotional phoneme-based acoustic model for speech synthesis system," *arXiv preprint arXiv:1806.09276*, 2018.

[24] 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.

[25] X. Zhu, Y. Zhang, S. Yang, L. Xue, and L. Xie, "Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis," *IEEE Access*, vol. 7, pp. 65 955–65 964, 2019.

[26] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," *arXiv preprint arXiv:1707.06588*, 2017.

[27] K. Liu, F. Meng, Y. Song, B. Fan, W. Duan, and W. Chen, "The sogou speech synthesis system for blizzard challenge 2018."

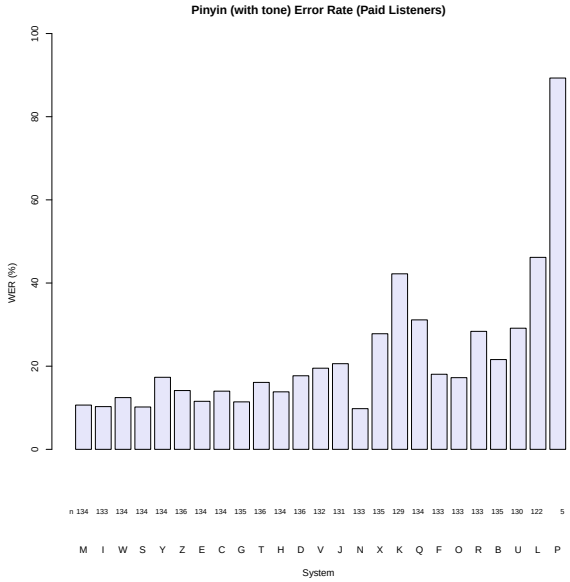

Figure 6: *Pinyin+Tone Error Rate of each system.*