The I2R-NWPU-NUS Text-to-Speech System for Blizzard Challenge 2018

Jinba Xiao[*], *Shan Yang*[*], *Mingyang Zhang*[†], *Berrak Sisman*[†],
Dongyan Huang[⅄], *Lei Xie*[*], *Minghui Dong*[⅄], *Haizhou Li*[†]

[*]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xian, China, 710129
[⅄]Human Language Technology Department,
Institute for Infocomm Research, A*STAR, Singapore, 138632
[†]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

jinba.xiao@gmail.com, syang@nwpu-aslp.org, mingyang.zhang@u.nus.edu,
berraksisman@u.nus.edu,
huang@i2r.a-star.edu.sg, lxie@nwpu.edu.cn, mhdong@i2r.a-star.edu.sg,
haizhou.li@nus.edu.sg

Abstract

This paper presents I2R-NWPU-NUS team's text-to-speech system to Blizzard Challenge 2018. Instead of using unit selection based concatenative speech sysnthesis previous years. we adopt the general deep neural network (DNN) statistical parametric method to synthesize the speech. The frame level acoustic parameters and phone duration are modelled using bidirectional long short-term memory (BLSTM) recurrent neural networks (RNNs). For duration model, 5 states of phone duration are used to predict the duration of each phoneme. Finally, the predicted acoustic parameters (MGC, LF0, V/UV) are taken as inputs to WORLD vocoder to generate the synthetic speech. The listening tests show improvement compared with the results of DNN baseline system.

Index Terms:Text-To-Speech, Blizzard Challenge 2018, LSTM-RNN, HMM, WORLD

1. Introduction

In Blizzard Challenge 2018, the EH1 task is to require the participating teams to deliver high quality synthesized speech from their text-to-speech systems given about 6.5 hours speech data from professionally-produced children's audiobooks, all data are from a single speaker. In addition, the organizers also provided cleaned-up text, sentence-level alignment between text and speech, segmented waveforms and associated extracted linguistic and acoustic features. Each participating team is then required to synthesize news items, new audiobooks, and semantically unpredictable sentences (SUS) with the test set, some of which are from previous years. The systematic listening tests are conducted by involving a variety of listeners to evaluate submitted synthesized speech.

For Blizzard Challenge 2017, we used a trajectory tiling method guided by a deep neural network to build our unit selection text-to-speech (TTS) system [1], which made several significant improvements. Firstly, phones are taken as the units instead of frames. The candidated phones are chosen and then concatenated together for synthesizing the target waveform. Secondly, the KLDs are used to measure divergence between spectrum HMMs of candidate and target phones for preselecting the phone candidates. Thirdly, in keeping with the unit change, the concatenation and target costs are also redefined.

For Blizzard Challenge 2018, we use a statistical parametric speech synthesis method. The BSLTM-RNNs [2] are used to model 5 states of phone duration and acoustic features including mel-generalized cepstral (MGC), BAP, LF0, V/UV flag and its corresponding dynamic features $\Delta \mathbf{c}_t$ and $\Delta^2 \mathbf{c}_t$. The final vocoder parameters are generated using the maximum likelihood parameter generation algorithm (MLPG) [3] by taking the predicted static and dynamic features into account to get a smooth trajectory. Finally the WORLD vocoder is used to synthesize the speech.

The organization of the paper is as follows. In Section 2 we present how to prepare and process the given data for developing the system. In section 3, the system is described in details. In Section 4 , the evaluation results are shown and discussed . Finally we conclude our work in Section 5.

2. Data Preparation

2.1. Transcription of Audiobook Dataset

Several audiobooks in the dataset are not transcribed. Thus we only use 54 stories of the dataset consists of 7119 sentences. Firstly, we convert the provided audio files to 16 bits wave files at sampling rate of 16k Hertz. Then these wave files are segmented into sentences and the corresponding text is split accordingly.

2.2. Full-Context Label and Alignment

The Festival frontend toolkit are used to extract the phoneme level full-context linguistic features including features on phoneme, syllable, word and syntactic phrase levels [4]. They are largely based on [5]. Then the state level alignment is done using the HTS [6] toolkit. For duration modeling, we predict 5 states of phone duration at state level, which can be modelled more accuracy than predictive duration at phone level. The HTS format full-context labels are converted by a question set, which consists of 416 questions, to binary features. Then the binary features are taken as inputs to duration model. For each input of the acoustic model, we append the forward and backward location information of current frame to the converted full-context label and time-aligned frame-by-frame with the acoustic features.

2.3. Acoustic Features

After we processed the provided original audio files, we use WORLD to extract the acoustic features. The acoustic features

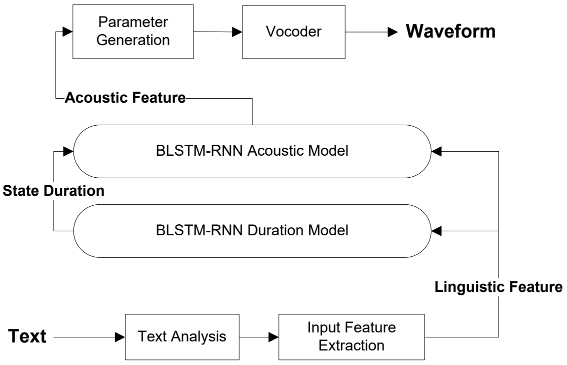

Figure 1: *System Architecture.*

$\mathbf{o}_t$ consist of static features $\mathbf{c}_t$, called vocoder parameters, and corresponding dynamic features $\Delta\mathbf{c}_t$ and $\Delta^2\mathbf{c}_t$, written as

$$\mathbf{o}_t = [\mathbf{c}_t^\top, \Delta\mathbf{c}_t^\top, \Delta^2\mathbf{c}_t^\top]^\top \tag{1}$$

The dynamic features are calculated from the sequence of static features and are used as a constraint to produce smooth parameter trajectories during generation. The extracted features includes mel-generalized cepstral (MGC), BAP, LF0 and V/UV flag.

3. Overview of the System

The overview of the proposed text-to-speech (TTS) system, is depicted in Figure 1. In the training processing, the acoustic parameters and phone duration are modelled with BLSTM-RNNs using the selected dataset. In the synthesis processing, 5 states of phone duration are predicted and the location information are appended to the linguistic feature as input of acoustic model, then MLPG is applied to the predicted acoustic parameters to get vocoder parameters. Finally we use the WORLD vocoder to synthesize the target waveform.

3.1. Acoustic Model

3.1.1. Deep Bidirectional LSTM Recurrent Neural Network

As powerful sequence neural networks, RNNs have come into sight again as a potential acoustic model for statistical parametric speech synthesis (SPSS). The LSTM architecture introduced in 1997 [7], is particularly attractive because it addresses the vanishing gradient problem in standard RNNs. Recent research studies have shown that LSTMs can attain much better performance than other deep neural networks on SPSS. Especially, the bidirectional RNN [8] can capture the sequence context information more effectively. Our acoustic model consists of 6 hidden layers of which the bottom 3 hidden layers are feed-forward structure with 512 nodes per layer, while the top 2 layers are bidirectional LSTM-RNNs structure (512 nodes) which followed by a fully connected output layer without activation.

Input features for duration model are generated from the full-context labels. The categorical features such as POS, phoneme IDs, and phrase types are transferred into binary features. The positions of words, syllables and phonemes are numerical features. The input vectors consist of totally 416 dimensions, where 373 dimensions are binary features for categorical linguistic contexts, and 43 dimensions including numerical sentence-level and linguistic contexts features.

The main ingredient input features for frame level acoustic modeling are the same as those used for duration modeling, 9 dimensions frame information are appended to the input features of duration model, which consist of forward and backward fraction through state, frame length of state, state index, frame length of phone, fraction of the phone made up by current state, forward and backward fraction through phone. The acoustic features are extracted by WORLD with 5 ms frame hop to form a vector of 199 dimensions which consists of MGCs (60), BAP (5), LF0 (1) and their corresponding Δ and Δ^2 features, besides a V/UV flag.

In all experiments, we use the Merlin package [9] and Tensorflow [10] to draw out the acoustic parameters and learn the deep neural networks respectively. We also investigated the mixture density network in BLSTM based TTS and end-to-end speech synthesis for EH1 task. The informal subjective listening tests show that the proposed system performs much better than these two systems.

3.1.2. Mixture Density Network

For mixture density network (MDN), the distributions of acoustic features given linguistic features can be a mixture of Gaussian distribution, because the same text can be spoken by human in many different manners. The outputs of a neural network trained by minimizing the squared loss function approximate the conditional mean of the outputs in the training data. This method may not be able to model distributions more complex than a unimodal Gaussian distribution. To address these issues, Zen et al. [11] have examined the usage of MDN as acoustic model for SPSS. The MDNs can be used to do multimodal regression and to predict variances as well. For MDN experiments, we use the same acoustic features and same architectures as BLSTM system, except with different outputs. The results showed that MDN performs worse than BLSTM because of over smoothed parameter trajectories and smaller energy.

3.1.3. End-to-End Speech Synthesis

A traditional statistical parametric TTS system consists of a text frontend extracting linguistic features, an acoustic model, a duration model, and a complex signal-processing-based vocoder. These components design require broad domain expertise. They are also trained independently, so errors from each component may compound. Based on sequence-to-sequence with attention mechanism, Y. Wang et al. [12] proposed an end-to-end TTS model named Tacotron, which attains a 3.82 subjective 5-scale mean opinion score on US English. The system outperforms all the existing parametric system in naturalness. It does not require phoneme-level or state-level alignment, we can simply train the model with original wave files and corresponding transcripts. We used the same acoustic features mentioned in section 2, and the same hyper-parameters and network architectures described in [12] except the reduction factor r is set to 10. Difference from BLSTM and MDN, the length of outputs of Tacotron is usually much longer than that of the inputs. This causes quick accumulation of prediction errors. The larger difference between input and output lengths, the more difficult it is to train the model. Hence, we set larger reduction factor r for 5ms frame shift length to reduce the length gap between inputs and outputs.

We found that the end-to-end system's performance is very sensitive to the training dataset. For blizzard challenge 2011's released dataset, Tacotron can get much better results than BLSTM, but can not even learn good alignment when using the

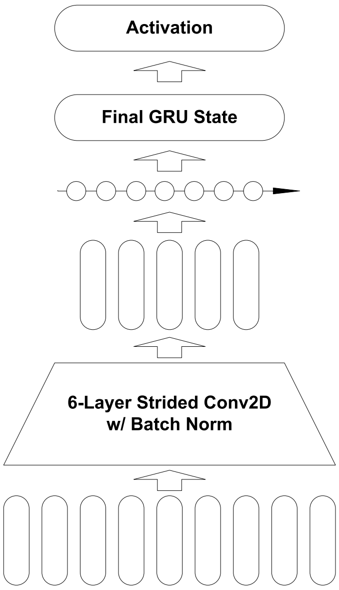

Figure 2: *The prosody reference encoder module.*

dataset of this year. The model barely predicts intelligible sentences in most cases. We supposed that attention mechanism is the most critical part of end-to-end system to learn the mapping between input transcripts and output acoustic features. At each decoder time step, a stateful recurrent layer produces the attention query to extract the corresponding context of input features, then the context vector guides the decoder to generate appropriate acoustic features. Incorrect attention context vector leads to the decoder predictions fit the targets poorly. Hence a more robust attention mechanism is required. Another reason is that the speaker will employ various timbre and rhythm to show the characteristics of the different characters in audiobook, such as age, gender and mood, speech segments from the dialogue or the aside differ widely in both acoustic and prosodic aspects. It is hard for Tacotron to learn an appropriate alignment between the input transcripts and output features for with the current attention alignment mechanism.

3.2. Reference Encoder

In order to produce more realistic speech, the model must take into account many external factors that are not given in a simple text. Such factors include the stress, intonation, and rhythm of the speech, are referred to prosody. In order to avoid labeling prosody manually, Skerry-Ryan et al. [13] proposed an encoder architecture named reference encoder, the encoder extracts a fixed-length representation of prosody from acoustic features. The fixed-length representation of prosody can be used to control the predicted prosody explicitly. As mentioned above, the speaker employs various timbre and rhythm to show the characteristics of the different characters in audiobook. The different types of prosodies can be extracted by the reference encoder automatically. In synthesis stage, we need to choose appropriate reference acoustic features to generate speech with similar prosody. For example, we can choose more expressive reference segments from the dialogue to synthesize speech.

We extend the BLSTM architecture by adding a reference encoder module which takes the reference signal as input, and computes a fixed-length embedding from it. Then prosody embedding is combined with the linguistic feature via a broadcast-concatenation. Figure 2 illustrates the structure of reference encoder module. For the reference encoder architecture, we use a 6-layer convolutional network, each layer is composed of 3×3 filters with 2×2 stride. SAME padding, RELU activation and batch normalization are applied to each convolutional layer. Take the outputs of the stacked convolutional layers as inputs of a recurrent network layer with 128-nodes Gated Recurrent Unit (GRU), the final 128-dimensional outputs of the GRU are projected to the desired dimensionality. For the same input sentence, there are significant differences between the two predicted pitch tracks under different reference signals. Figure 3 shows two predicted pitch tracks for the same utterance under different reference signals. We can see that the widely varies between the two pitch trajectories which demonstrate the prosody extraction ability of reference encoder module.

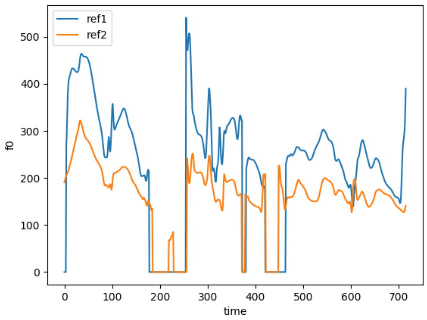

Figure 3: *The pitch trajectory of same sentence under different reference.*

3.3. Vocoder

WaveNet [14] is an autoregressive generative model for waveform synthesis. It operates at a very high temporal resolution of raw audios and produces human-level audio of quality. In comparison to the linguistic and acoustic features used in WaveNet, the vocoder features are a simpler and lower-level acoustic representation of audio signals. So we usually take WaveNet as a neural vocoder to generate high quality audio. In our experiments, we build a WaveNet with 30 dilated convolution layers, grouped into 3 dilation cycles, i.e. the dilation rate of layer $\mathbf{k}(k = 0...29)$ is $2^{k(mod10)}$. For local condition features, we use the same acoustic features consisting of MGC, BAP, LF0, and V/UV. We train WaveNet using ground truth features and then synthesize speech using features predicted by BLSTM system. Because of the large mismatch between groud truth and predicted features, the final synthesized speech by WaveNet contains too much noise and is not as good as synthesized speech by WORLD. We tried to retrain the WaveNet but get worse results. There may be errors in our implementation while training WaveNet vocoder, so finally we used WORLD as vocoder.

4. Subjective results

This section discusses the evaluation results of our system in Blizzard Challenge 2018 in details. Our system is identified as ”G”. Whereas system A is natural speech, B is the Festival unit-selection benchmark system, C is the HMM benchmark built

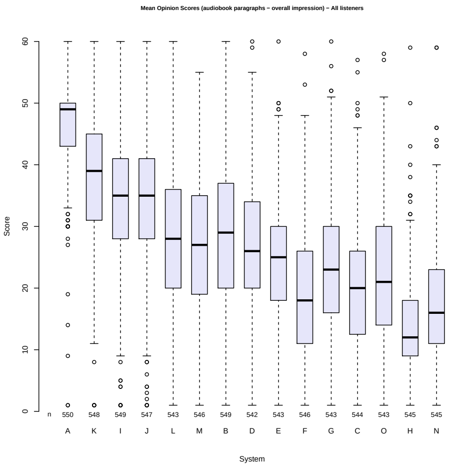

Figure 4: *Results of the overall impression MOS of audiobook paragraphs.*

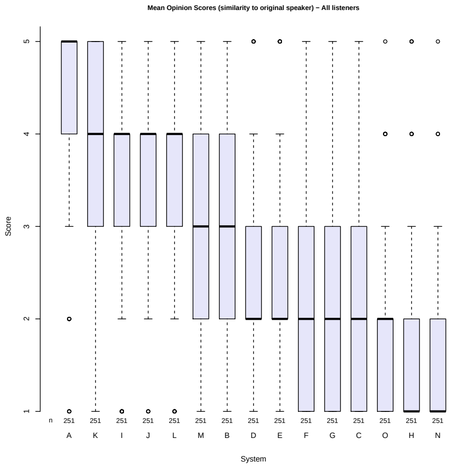

Figure 6: *Results of the similarity MOS.*

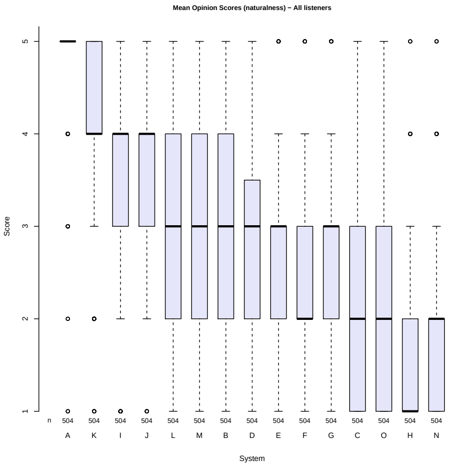

Figure 5: *Results of the naturalness MOS.*

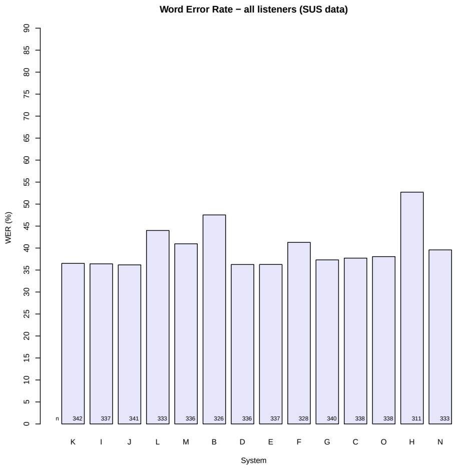

Figure 7: *WER in the SUS testing.*

using the HTS toolkit, D and E are DNN benchmark built using the HTS toolkit. Systems F to O are the 10 participating teams. The submitted synthesized audio files go through comprehensive listening tests which include four major parts.

- Part 1 is made up of of two multi-dimensional tests of the book paragraphs. The tested dimensions are overall impression, pleasantness, speech pauses, stress, intonation, emotion and listening effort.
- Part 2 contains two naturalness tests of the book sentences.
- Part 3 is a similarity test of the book sentences. The listeners are requested to judge the similarity between the synthesized speech and the provided speech.
- Part 4 composed of two intelligibility tests of the SUS speech. The sentences are semantically unpredictable, and it is difficult to make a guess of a word through the surrounding words. The listeners are requested to write down the words after listening of a sentence.

In the tests, three types of listeners involved: paid listeners, online volunteers and speech experts.

The test results include the overall impression, naturalness, similarity and intelligibility as evaluated by all types of listeners. All the figures show results of all the systems.

Figure 4 shows the results of the overall impression MOS of audiobook paragraphs given by all listeners for all the systems. Figure 5 shows the results of the naturalness MOS of all the systems given by all listeners. Figure 6 shows the results of the similarity MOS of all the systems given by all listeners. Figure 7 shows the word error rate of all the systems based on the evaluation of all listeners. The WER decreases compared with our result of last year. Compared to our results of last year, the unit-selection system can get higher quality speech than conventional BLSTM neural networks. Advanced TTS with WaveNet will be our focus in near future.

5. Conclusions

In this paper we introduce the development of our speech synthesis system to Blizzard Challenge 2018. This time we use the general BLSTM as speech synthesis system. Mixture density network and end-to-end architectures are investigated to build different TTS system, and reference encoder is applied to control the predicted prosody. The evaluation results indicate that the BLSTM system is a stable baseline system comparing with end-to-end system.

6. References

[1] Yanfeng Lu, Zhengchen Zhang, Chenyu Yang, Huaiping Ming, Xiaolian Zhu, Yuchao Zhang, Shan Yang, Dongyan Huang, Lei Xie, Minghui Dong, “The I2R-NWPU Text-to-Speech System for Blizzard Challenge 2017.”

[2] 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.

[3] 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.

[4] A. Black, P. Taylor, R. Caley, R. Clark, K. Richmond, S. King, V. Strom, and H. Zen, “The festival speech synthesis system, version 1.4.2,” *Unpublished document available via http://www.cstr.ed.ac.uk/projects/festival.html*, 2001.

[5] H. Zen, “An example of context-dependent label format for hmm-based speech synthesis in english,” *The HTS CMUARCTIC demo*, vol. 133, 2006.

[6] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, “The hmm-based speech synthesis system (hts) version 2.0.” in *SSW*. Citeseer, 2007, pp. 294–299.

[7] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[9] Z. Wu, O. Watts, and S. King, “Merlin: An open source neural network speech synthesis system,” *Proc. SSW, Sunnyvale, USA*, 2016.

[10] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “Tensorflow: a system for large-scale machine learning.” in *OSDI*, vol. 16, 2016, pp. 265–283.

[11] 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.

[12] 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*, 2017.

[13] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” *arXiv preprint arXiv:1803.09047*, 2018.

[14] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio.” in *SSW*, 2016, p. 125.