The I2R-NWPU-NTU Text-to-Speech System at Blizzard Challenge 2016

Zhengchen Zhang[*], *Mei Li*[†], *Yuchao Zhang*[†], *Weini Zhang*[†], *Yang Liu*[†], *Shan Yang*[†], *Yanfeng Lu*[*], *Van Tung Pham*[$], *Lei Xie*[†], *Minghui Dong*[*]

[*]Human Language Technology Department,
Institute for Infocomm Research, A*STAR, Singapore, 138632
[†]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xian, China, 710129
[$]School of Computer Science and Engineering,
Nanyang Technological University, 50 Nanyang Avenue Singapore, 639798

Email: lxie@nwpu.edu.cn, mhdong@i2r.a-star.edu.sg

Abstract

In this paper, we introduce a trajectory tiling method guided by deep neural networks (DNNs) for text-to-speech (TTS), which is the entry to Blizzard Challenge 2016 by I2R-NWPU-NTU team. We build a deep bidirectional LSTM (DBLSTM) based network to predict the phoneme level duration and frame level acoustic parameters. After the acoustic parameters are predicted, the best units are selected from the database using a trajectory tiling method. Experiments demonstrate that, under the DBLSTM framework, the context information of a phoneme extracted in text processing will help the duration prediction, while not help the acoustic modeling. The results of subjective evaluation are also discussed.

Index Terms: Blizzard Challenge 2016, Text-To-Speech, DBLSTM-RNN, Trajectory Tiling

1. Introduction

This paper introduces the system submitted to Blizzard Challenge 2016 by Institute for Infocomm Research (I2R), A*STAR, Singapore, Northwestern Polytechnical University (NWPU), Xi'an, China, and Nanyang Technological University (NTU), Singapore.

The task of Blizzard Challenge 2016 is to build a voice that is able to read audiobooks for children. About 5 hours of speech corpus, which is the voice of a same female speaker, were released as the training data. The data is selected from children's audiobooks produced by professionals. A testing set of sentences are also provided. All participates are asked to generate the speech files on the testing set, and submit the files to the organizer. A large scale subjective evaluation will be conducted. The participates are allowed to use external data except some exclusions given in the rules.

Recently, motivated by the success of Deep Neural Networks (DNNs) in speech recognition, many DNN research attempts have been tried in the speech synthesis field [1, 2]. DNN offers a powerful modeling paradigm for the complex relationship between input features and output features. Recurrent Neural Networks (RNNs), especially with bidirectional Long Short Term Memory (LSTM) cells [3, 4, 5, 6, 7], in principle can capture information from anywhere in the feature sequence. The deep bidirectional LSTM (DBLSTM) architecture, which is the integration of deep bidirectional RNN and LSTM, by taking advantages of DNN and LSTM, can model the deep representation of long-span features more precisely. The system proposed in this work is a trajectory tiling method guided by DNN. A DBLSTM architecture is built to do prosody and acoustic modeling. After the acoustic parameters are predicted, a trajectory tiling method is employed to select the best units from the database. No external training data was used in our system.

2. Data Processing

2.1. Preprocessing and Alignment

The audiobook data provided by the organizing committee contain two parts: the transcribed part and the untranscribed part. The transcription provided consists of time indices accompanied with pure texts without punctuation. Only scripts in PDF format are provided for the untranscribed part. For the first part we use the provided time indices directly. But we try to find the original texts from the book scripts provided based on the pure texts. The original texts contain punctuation marks. These are helpful for generating more accurate full-context labels. For the second part we make our own transcription based on the scripts in PDF format. All the audio books are converted and segmented into sentence level wave files with their corresponding original texts. Then we put the combined dataset through our Automatic Speech Recognition (ASR) engine to do phoneme level alignment. The ASR engine takes pronunciation variations into account. The phoneme sequence for a certain word is the most likely one based on the acoustic model pre-trained with a large database. This phoneme level alignment is the basis of full-context label generation and acoustic modelling.

2.2. Text Analysis

The lexical features used to generate full-context label can be classified into four categories: phoneme, syllable, word, and syntactic phrase. Phonemes and syllables of words are obtained from our dictionary. The Part-Of-Speech (POS) is obtained using a simple statistical tagger which achieved 91% accuracy on the Brown cd, ce, cf subsets in the Penn Treebank-3 corpus [1]. The context of phonemes, syllables and POS of words in an utterance are considered. Position information is an important feature for TTS. The forward and backward position of a phoneme in the syllable, the position of a syllable in a word, and the position of a word in a phrase are all used as features in our system. The syntactic phrase information is added to improve the prosody modeling. We use ZPar [8] to obtain the syntactic tree of a sentence. The details of the features are shown in Table 1.

3. Speech Synthesizing

In the synthesis stage, the full-context labels of the test sentences are first converted to input vectors of our phoneme level duration model. Then we use the trained duration model to predict the frame number of each phoneme of the sentences. With the frame numbers and the phoneme level features, a sequence of frame-level linguistic features is generated and fed into the acoustic model to predict acoustic parameters. The generated parameters are then used for trajectory tiling.

3.1. Duration Prediction and Acoustic Modeling

A hybrid of DNN and BLSTM-RNN is built for our duration prediction and acoustic modeling. There are 4 hidden layers with 256 nodes per layer, where the bottom 2 hidden layers are feed-forward structure with sigmoid activation functions, while the top 2 hidden layers are Bidirectional RNN structure with LSTM (128 forward nodes and 128 backwards nodes).

Input feature vectors for duration prediction are generated from the full-context label we get from Section 2 and time-aligned frame-by-frame with the output features. The categorical features like phonemes, POS types, and phrase types are transferred to binary features. The positions of phonemes, syllables and words are numerical features. There are a total of 348 dimensions in the input vectors, where 295 dimensions are binary features for categorical linguistic contexts and 53 dimensions are numerical linguistic contexts. As the context information has been naturally considered in the LSTM network, we do another experiment, in which we remove the context information in the full-context label. All the information like previous and next phonemes, POS types of previous and next words are removed. A new feature vector is generated, in which only the properties of the current phoneme are used. This vector contains 106 dimensions, where 81 dimensions are binary features for categorical linguistic contexts and 25 dimensions of numerical features.

The output of the duration prediction is a vector with five dimensions, which are the frame number of previous two phonemes, the current phoneme and the next two phonemes.

The input feature vector for frame level acoustic modeling is almost the same with those for duration prediction, except that 3 dimensions containing frame information are added: the frame numbers in the current phoneme as well as the forward position and the backward position of the current frame in the phoneme. This results in the input feature vector which has context information containing 351 dimensions and the input feature vector which does not have context information containing 109 dimensions. The output of the acoustic modeling network is a vector with 51 dimensions, which are parameters including 41-dimensional LSPs and linearly interpolated F_0 in log-scale with their previous 4 and next 4 F_0, plus a voicing/unvoicing (V/UV) flag. Both of the input and output features are normalized by mean-variance normalization (MVN).

3.2. Trajectory Tiling

In general, the longer the units are used for trajectory tiling, the less concatenation points are needed and the higher the voice quality will be obtained. Nevertheless, considering the capacity of the corpus, frame (5ms) segments are used for tiling in this work.

3.2.1. Unit Pre-selection and Lattice Construction

Target cost, the distance between generated parameter trajectories and units in database, is used to pre-select the candidate units for lattice construction. The distances of F_0, gain and LSP features are defined respectively by [9]

$$d_{F0} = |log(F0_t) - log(F0_c)| \tag{1}$$

$$d_G = |log(G_t) - log(G_c)| \tag{2}$$

$$d_\omega = \sqrt{\frac{1}{I}\sum_{i=1}^{I} w_i(\omega_{t,i} - \omega_{c,i})^2} \tag{3}$$

$$w_i = \frac{1}{\omega_{t,i} - \omega_{t,i-1}} + \frac{1}{\omega_{t,i+1} - \omega_{t,i}} \tag{4}$$

where absolute value in log domain is used to evaluate F_0 and gain distance, while LSP distortion is measured

[1] https://catalog.ldc.upenn.edu/LDC99T42

Category	Feature detail
Phoneme	Pre-previous, previous, current, next, and next-next phoneme
	Position of current phoneme in the current syllable (forward and backward)
Syllable	Number of phonemes in current, previous, and next syllables
	Vowel of current syllable
	Number of syllables in the utterance
	Position of current syllable in the word (forward and backward)
	Syllable number of current, previous, and next words
	Are previous, current, and next syllable accented or not
	Number of syllables from the previous accent syllable to the current syllable
	Number of syllables from the current syllable to the next accented syllable
Words	Part-Of-Speech (POS) of previous, current, and next words
	Number of words in the utterance
Syntactic Phrase	Phrase type of father phrases (FPs) of current, previous, and next words
	Phrase type of grandfather phrases (GPs) of current, previous, and next words
	Level of FPs and GPs of current, previous, and next words
	Forward and backward indices of current word in FP

Table 1: Lexical features used in this work.

by a weighted root mean square. Owing to the intrinsic property of LSP, in Eq. (4), the inverse harmonic mean weighting (IHMW) function is used to evaluate weights w_i [10]. In view of the fact that frame units are used, no alignment is needed for distance calculation. Finally, the distances of these three features are first weighted, and then added together:

$$d(u_t, u_c) = w_{F0}\bar{d}_{F0} + w_G d_G + w_\omega d_\omega. \quad (5)$$

To avoid the weight tuning, we normalize the distances of all features to a standard normal distribution with zero mean and unit variance and the resultant normalized distance is

$$d(u_t, u_c) = N(\bar{d}_{F0}) + N(\bar{d}_G) + N(\bar{d}_\omega). \quad (6)$$

3.2.2. Dynamic Programming-based Search and Concatenation

With Eq. (6) as target cost for dynamic programming-based search, we use Normalized Cross-Correlation (NC-C) as the objective measure of concatenation smoothness for searching the optimal unit path, namely join cost. Given two time series $x(t)$ and $y(t)$, the NCC $r(d)$ between them can be calculated by

$$r(d) = \frac{\sum_t [(x(t) - \mu_x) \cdot (y(t) - \mu_y)]}{\sqrt{\sum_t (x(t) - \mu_x)^2} \cdot \sqrt{\sum_t (y(t) - \mu_y)^2}}. \quad (7)$$

In order to calculate the NCC, overlap between adjacent waveform units is needed. However, to avoid length shrinking, we extend waveform units with waveform contexts. As for join cost, we shift adjacent waveform units to maximize the NCC, recording the corresponding offset. Then dynamic programming-based Viterbi search is used to find the optimal unit path which minimizes the accumulated weighted target and join cost. Finally, adjacent waveform units along the optimal path are shifted by the recorded offset and concatenated with triangular cross-fading.

4. Experimental Results

A total of 5,425 sentences were used in our experiments. We randomly selected 5,000 sentences for model training, 225 for development, and 200 for testing. We conducted two groups of experiments to investigate the influence of context information of the input linguistic features both on acoustic modeling and duration modeling.

4.1. Duration Prediction

We use root mean squared errors (RMSE) as the evaluation criterion. The results obtained on the testing set are listed in Table 2.

Table 2: Results of phoneme level duration prediction.

Input	Duration RMSE (ms)
No context	26.04
With context	25.61

From the results, we can see that the input vector which does contain context information has the lower RMSE of predicted phoneme duration. This suggests that the properties of the neighboring phoneme, word and syllable may provide useful information for the prediction of current phoneme durations. Hence, we use the one which contains context information to train our final duration model.

4.2. Acoustic Modelling

The results of acoustic modeling are listed in Table 3. Log-spectral distance (LSD), root mean squared errors (RMSEs) of F_0, and voiced/unvoiced error rates (V/UV) are used to evaluate the model.

Table 3: Results of acoustic modelling.

Input	LSD (dB)	F_0 RMSE (Hz)	V/UV (%)
No context	0.175	49.774	6.431
With context	0.177	48.431	6.559

From the results, we can see that including context information in the input feature vector can reduce the root mean square error of F_0. However, the input feature vector which does not contain context information has the lower log spectral distance and voiced/unvoiced error rates. This may suggest that BLSTM can powerfully learn the long context dependencies. Including the neighboring phoneme, word, syllable and phrase properties may somehow weaken the properties of the current one, and result in the poor prediction especially for spectral information and UV. Hence, we use the one does not contain context information to train our final acoustic model.

4.3. Subjective Evaluation Results

In Blizzard Challenge, all participates will generate the audio files for a set of testing sentences, and submit them for evaluation. Three types of listeners are invited to evaluate the submitted systems this year: paid participants, volunteers, and speech experts, respectively. Here we only select the results given by the paid participants. There are a total of 1251 sentences in the 2016 testing set. To evaluate the systems' capability of utilizing linguistic context, 4 types of speech files need to be generated, namely books, chapters, pages, and lines. In the book speech file, a whole book is read. In the lines speech file, only one sentence is spoken. There are also 200 Semantically Unpredictable Sentences (SUS) and 200 news sentences provided, which are the same with those of 2012 and 2013 [11]. The evaluation results of all systems are shown in Figures 1, 2, 3, and 4. Our team is marked as O. System A is natural speech, System B is an unit selection based system submitted by CSTR in Blizzard Challenge 2007 [12], System C is the HTS benchmark, System D is a DNN benchmark and Systems E to Q are different participants.

Figure 1: Mean Opinion Scores for audiobook paragraphs given by paid listeners.

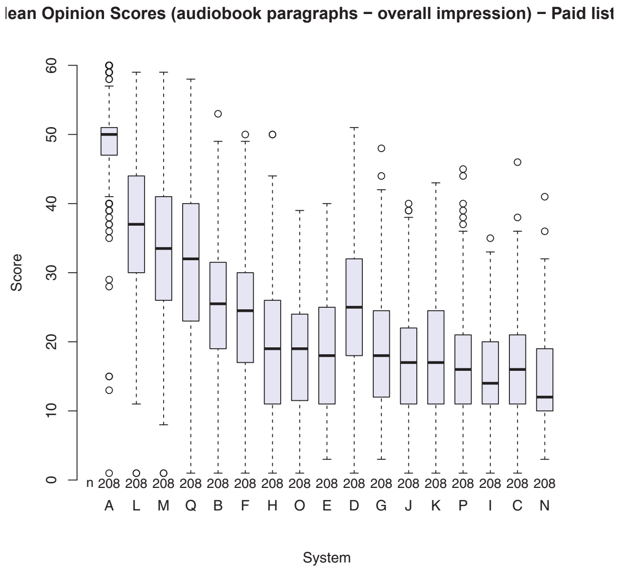


Figure 2: Mean Opinion Scores for naturalness given by paid listeners.

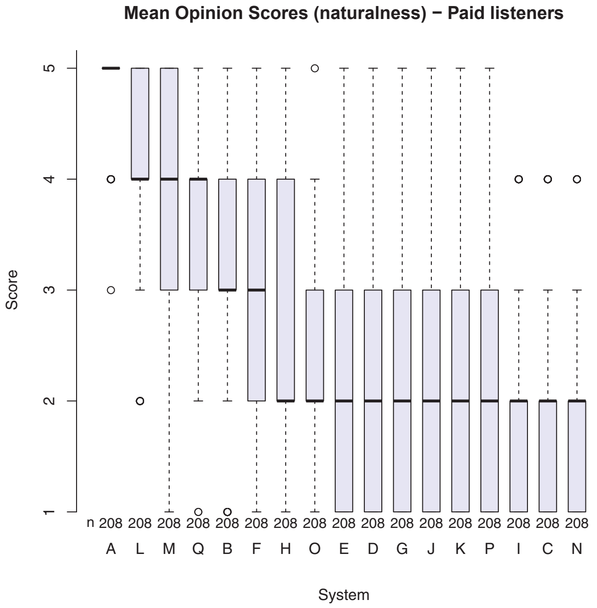


Figure 1 illustrates the Mean Opinion Scores (MOS) for audiobook paragraphs. Our system obtained an average MOS of 19, which is lower than System B (26) and System D (25), and is higher than System C (17). Figures 2 shows the MOS of naturalness given by the paid listeners. System O got a mean score of 2.4, which is higher than System D (2.1) and System C (1.8), and is lower than System B (3.2). Figure 3 shows the MOS of similarity to the original speakers. System O obtained an an average score of 2.1, which is higher than System C (1.5) and System D (1.9), and is lower than System B (3.5). Figure 4 shows the Word Error Rate (WER) of all systems in the SUS testing. System O got a mean error of 0.26, which is higher than System B (0.19), C (0.18) and D (0.16). The results demonstrate that our system performs worse than System B, although both of them are unit-selection methods. Our system outperforms System D in terms of naturalness and similarity, while got higher WER than it. That may be

Figure 3: Mean Opinion Scores for similarity to the original speaker given by paid listeners.

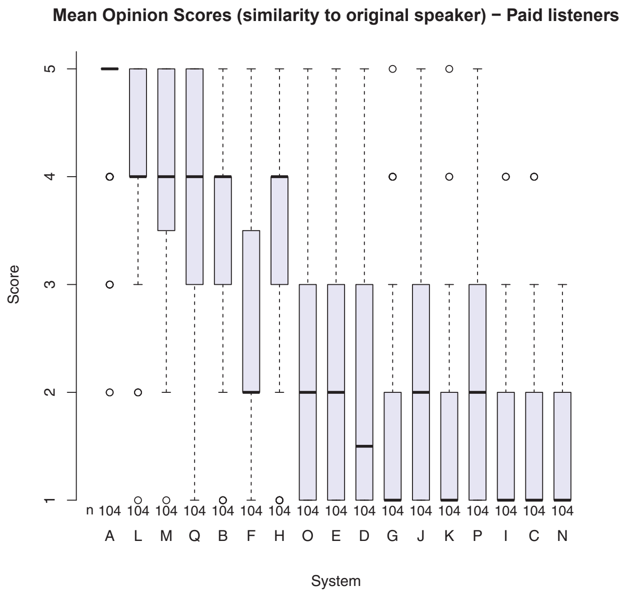
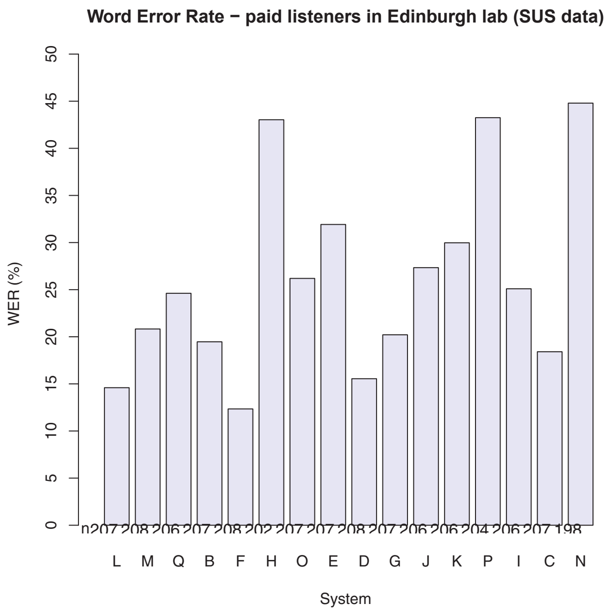


Figure 4: Word Error Rate (WER) in the SUS testing.



the reason that we got lower MOS than D for audiobook paragraphs, because the listeners sometimes cannot hear the speech clearly.

5. Conclusion

In this paper, a speech synthesis system has been built to read audiobooks for children. A hybrid method has been employed, which is a trajectory tiling method guided by a DNN network. The experiments demonstrated that in the duration prediction, the context information in the input vectors will improve the system performance even though the LSTM network has considered the context information. However, the context information did not help much in acoustic modelling. According to the subjective evaluation results, there is still much room for improvement of our method.

6. Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (Grant No. 61571363).

7. References

[1] Heiga Zen, Andrew Senior, and Mike Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (I-CASSP)*. IEEE, 2013, pp. 7962–7966.

[2] Yao Qian, Yuchen Fan, Wenping Hu, and Frank K Soong, “On the training aspects of deep neural network (dnn) for parametric tts synthesis,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3829–3833.

[3] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, “Speech recognition with deep recurrent neural networks,” in *2013 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2013, pp. 6645–6649.

[4] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed, “Hybrid speech recognition with deep bidirectional lstm,” in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (AS-RU)*. IEEE, 2013, pp. 273–278.

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

[6] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber, “Learning precise timing with lstm recurrent networks,” *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.

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

[8] Yue Zhang and Stephen Clark, “Syntactic processing using the generalized perceptron and beam search,” *Computational linguistics*, vol. 37, no. 1, pp. 105–151, 2011.

[9] Zhi Jie Yan, Yao Qian, and F. K. Soong, "Rich-context unit selection (rus) approach to high quality tts," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4798–4801.

[10] R. Laroia, N. Phamdo, and N. Farvardin, "Robust and efficient quantization of speech lsp parameters using structured vector quantizers," in *1991 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1991, pp. 641–644.

[11] Kishore Prahallad, Anandaswarup Vadapalli, Naresh Elluru, G Mantena, B Pulugundla, P Bhaskararao, HA Murthy, S King, V Karaiskos, and AW Black, "The blizzard challenge 2013–indian language task," in *Blizzard Challenge Workshop 2013*, 2013.

[12] Mark Fraser and Simon King, "The blizzard challenge 2007," in *Blizzard Challenge Workshop 2007 (in Proc. of the 6th ISCA Workshop on Speech Synthesis)*, 2007.