The NLPR Speech Synthesis entry for Blizzard Challenge 2017

Jianhua Tao[1,2,3], Ruibo Fu[1,3],Yibin Zheng[1,3], Zhengqi Wen[1], Ya Li[1], Biu Liu[1]

[1]National Laboratory of Pattern Recognition,
[2]CAS Center for Excellence in Brain Science and Intelligence Technology,
Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
[3]University of Chinese Academy of Sciences

{jhtao, ruibo.fu, yibin.zheng, zqwen, yli, liubin }@nlpr.ia.ac.cn

Abstract

The paper describes the CASIA speech synthesis system entry for Blizzard Challenge 2017. About 6.5 hours of speech data from professionally-produced children's audiobooks is adopted as the training data for the construction this year. Our synthesis system is built based on the BiLSTM guided unit selection and waveform concatenation approaches by using the provided corpus. Different from our previous system, some improvements about unit selection strategies were made to adapt to different types of the utterance. In this paper, the definitions of the acoustic and the contextual parameters, strategies of candidate unit selection, the calculation of costs based different contexts will be introduced and discussed. Finally, the results of the listening test will be presented.

Index Terms: speech synthesis, phone duration modeling, BiLSTM, unit selection, Blizzard Challenge 2017

1. Introduction

This paper describes details about our fifth entry speech synthesis for Blizzard Challenge. The task of this year is to build a speech synthesis system suitable for reading audiobooks to children based on the provided data. The articulation of the audiobooks is full of expressiveness and the speaker express several different types of characters.

Statistical parametric speech synthesis (SPSS) systems have flexible and robust advantages [1] compared to the unit selection [2] systems. However, the degradation of the speech quality and the naturalness is very significant during the process of extracting and modeling speech parameters, followed by re-synthesis. Therefore, these systems are sound consistently too much and less natural than unit selection system, as we can see in the results of many Blizzard Challenges [3, 4, 5, 6].

With the aim to improve the quality of speech and make the children audiobooks more readable, a bi-direction long short-term memory recurrent neural network (BiLSTM) guided unit selection synthesis system is built for the Blizzard Challenge 2017. This system is an improvement version of our last year' system [7] in mainly three aspects, (i) the calculation of the concatenation cost's will be adjusted to the context of the utterance and the calculation will combine more the acoustic and lexical features than just weighted linear combination. (ii) more delicated data selection works have been done to delete some over-expressive utterances and each utterances auto labeled with the degree of expression (iii) target cost calculation approach, as BiLSTM guided synthesis system doesn't have the concept of "state". Some target cost computation approaches are investigated and compared in our system.

The rest of the paper is organized as follows. Section 2 gives an overview of our methods used for system construction. Section 3 introduces the works about data selection and auto-labeling. Section 4 introduces the unit selection module, including the pre-selection of units, calculation of target and concatenation cost. In section 5, the evaluation results of our system in Blizzard Challenge 2017 are shown and discussed. The conclusions are presented in section 6.

2. System Overview

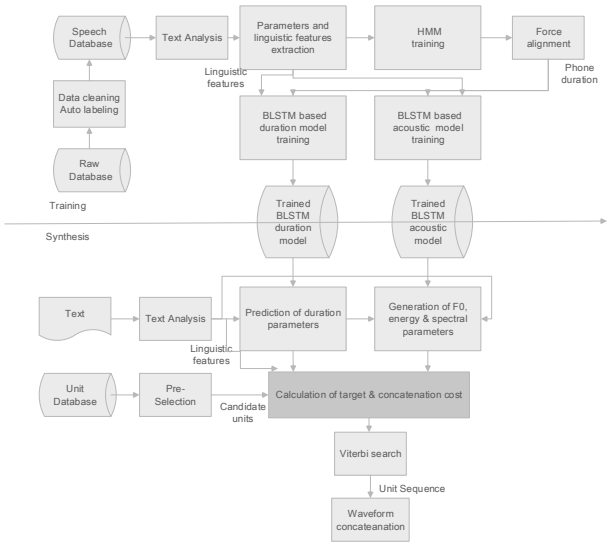

Figure 1: *An overview of our system.*

A DNN-hybrid synthesis system is composed of the BiLSTM statistical models to generate speech parameter trajectories and unit-selection based on the generated acoustic parameter and the lexical parameter preprocessed by the text analysis front-end [8, 9, 10] . We adopt this framework as our synthesis entry for Blizzard Challenge 2017. HMM is the preferred statistical model in hybrid system's target cost function in previous years. However, as recent but compelling evidence that BiLSTM is superior to the regression tree employed in HMM systems [11, 12, 13], a hybrid synthesis system based on BiLSTM is employed to synthesis the voices for Blizzard Challenge 2017. The flowchart of the BiLSTM guided unit selection speech synthesis system for Blizzard Challenge 2017 is shown in Figure 1. It consists of two main stages: the training stage and

the synthesis stage, to build the BiLSTM guided unit selection speech synthesis system.

In the training stage, the data will be preprocessed by an utterance expression recognition to elimate the over expressive data. And the output of the model also give guidance about the degree of the consistency between of the two concating units. After preprocessing, BiLSTM based acoustic and duration model is trained to guide the unit selection. Before the training of BiLSTM based acoustic model, a HMM based force alignment is performed first. In the HMM based force alignment part, acoustic parameters are extracted from the speech waveforms. The complete feature vector for each frame consists of static, delta and acceleration components of the spectral parameters and the logarithmized F0. With the segmental and lingustic features data from text analysis module (which is done by festival toolkit [14]), the spectral part is modeled by continuous probability HMM and F0 part is modeled by multi-space probability HMM (MSD-HMM). Then the phone boundaries of the training utterances are determined by Viterbi alignment using the trained HMM model. Then the linguistic features, together with the phone duration from the force alignment part is made up of the input for BiLSTM training. As for the output of BiLSTM training part, the complete feature vector for each frame only consists of static components of the spectral parameters and the logarithmized F0, together with a flag of unvoice/voice (U/V). The BiLSTM based acouctic model is used to calculate the target cost. The linguistic features and the phone duration also made up of the training corpus for the BiLSTM based duration model. The BiLSTM based duration model is used to predict the phone duration, target and concatenation cost in the open test.

In the synthesis stage, firstly, the contextual information of the text to be synthesized is analyzed and extracted by text analyzer (festival toolkit). Secondly, the pre-selection procedure is conducted according to the contextual information (including the type of utteraces). Then the phone duration is predicted using the trained BiLSTM duration model. Then the phone duration model, together with the linguistic features are fed into the BiLSTM to predict the target acoustic parameters. Next, the target cost of candidate unit and the concatenation costs between each pair of adjacent candidate units can be calculated. The optimal candidate units are selected by Viterbi search. Finally, the waveform fragments of optimal units are concatenated, and the silence sections are inserted between some adjacent words based on the value predicted by silence model.

3. Data Preprocessing

3.1. Data selection

The provided data contain some utterances that is too loud or too small. Some may even sounds a little exaggerate. So we train a model by acoustic features (energy,spectral) to recognize these over expressive utterances. Experiments show that the system will be more stable after deleting some outlier utterances. In oreder to make sure that we really delete those over expressive utterances, we manually listen those deleted data and correct some mistakes. And then uses these manually checked data that need to be deleted as the negative samples to re-train the recognition model.

3.2. Data auto-labeling

We find that there are two types of utterances (statements and dialogue). Some dialogue are more expressive than the some dialogue because they are in different contexts. However, when we build the database for unit-selection, we find that it will cause inconsistency due to two degrees of expressive units concatenated together. To solve this problem, we also train a model to label the degree of expressiveness of each utterances. We use self-trainning method to train our model.

4. Unit Selection Method

4.1. Pre-selection module

In a corpus based speech synthesis system, there are too many candidate for each target unit. Conducting unit selection procedure on such a large database is very time-consuming. To decrease the number of candidate units and thus improve the running speed, a contextual information difference (CID) based pre-selection is conducted. The *CID* is defined in Equation (1) as below:

$$CID = \sum_{i=1}^{N} w_i * D_i \quad (1)$$

,where N is the number of contextual information category, D_i is the difference of the *i-th* contextual information between current candidate unit and the target unit and w_i is the weight of the *i-th* contextual information.

The *CID* depicts the difference of contextual information between the candidate unit and the target unit to be synthesized. The contextual information used here includes the location of the current speech unit in word, phrase and sentence, the name of the phone, the length of word, phrase and sentences, the boundary types before and after the current unit, etc.

After the pre-selection, a small number of candidate units which have the smallest *CID* will be kept for the later processing.

4.2. Target cost calculation approaches

Target cost is defined as the difference between the predicted parameters and the parameters of candidate unit. In our work, the parametrers used for target cost include F0, duration, energy and spectral parameters). The context embeddings derived from a neural network, or alternatively the actual speech parameters predicted at the output of the network, can be thought of as a non-linear projection of the input lingutic features. The projection is learned in a supervised manner, according to whatever optimization criterion is used to train the network. We suppose that these BiLSTM-derived features are more powerful than the purely linguistic features or HMM-derived features. The motivation for using a BiLSTM – that, crucially, has been trained to perform the state-of–the-art performance in SPSS in recent research.

The system we built for the Blizzard Challenge 2017 operates on the phone units. Different from HMM model, the BiLSTM based model doesn't have the concept of "State". Therefore, it leaves an important problem to calculate the target cost effectively. For this year's Challenge, we iprove our target cost calculation method that make the cost show more the similuarty abot the core articulation of each unit. And four target cost calculation methods are tested and compared for the output of speech parameters from BiLSTM based

acoustic model (including the F0, erengy, spectra). And for the output of duration parameters from BiLSTM based duration model, only the Euclidean distance is used.

4.2.1. The Kullback Leibler divergence (KLD)

We divided each phone into 4 sections. The features being used for the target cost (output of speech parameters from BiLSTM based acoustic model) are gathered toghteher across all frames within each of these 4 regions, from which we compute the mean and variance per section. The variance is floored at 1% of the gloval variance per feature (the floor value was chosen via informal listening). This is done for both the candidate and the target units.

The Kullback Leibler divergence (KLD) is computed for each of the 4 sub-phone regions individually.

The KLD between distribution *f* of the features computed for the frames corresponding to a given section in the test sentence, and distribution *g*, is:

$$D_{KL}(f \parallel g) = \frac{1}{2}[\log\frac{\left|\sum_g\right|}{\left|\sum_f\right|} + Tr[\sum_g^{-1}\sum_f] - \mathrm{d} + (\mu_f - \mu_g)^T\sum_g^{-1}(\mu_f - \mu_g)] \quad (2)$$

where μ and $\sum$ are mean and covariance and *d* is the dimensionality of the feature vector. The KLD for each of the 4 sections comprising a phone is summed toghther to give the final divergence score. The average of $D_{KL}(f \parallel g)$ and $D_{KL}(g \parallel f)$ was used in order to make the measure symmetrical.

4.2.2. Maximum likelihood criterion (LL)

The same as the approach mentioned in section 4.2.1, we also divided each phone into 4 sections. The mean and variance calculation approaches is also the same as section 4.2.1. The only difference is that maximum likelihood criterion is employed for the target cost.

The maximum likelihood criterion (LL) is computed for each of the 4 sub-phone regions individually.

The LL between distribution *f* of the features computed for the frames corresponding to a given section in the test sentence, and distribution *g*, is:

$$LL(f \parallel g) = \frac{D_f}{D_g}\sum_{i=1}^{D_g}(x_{fi} - \mu_g)^T\sum_g^{-1}(x_{fi} - \mu_g) \quad (3)$$

where the likelihood of acoustic model is normalized by the candidate sub-phone duration D_g and predicted sub-phone duration D_f, and the x_{fi} is the speech parameters in *i-th* frame in the sub-section of the candidate phone.

4.2.3. Relative position based Euclidean distance (ED)

As the BiLSTM based acoustic model doesn't possess the concept of "State", and we suppose that relative position of the acoustic parameters can capture the trajectory of the acoustic parameters well. As we choose the same number of relative position for the candidates and target units, then they would have the same length. As a result, Euclidean distance can easily be employed in such situation. Therefore, a relative positon based Euclidean distance (ED) is tested in our system.

The relative position based Eucidean distance (ED) is computed for each phone regions individually (we don't need to divide each phone into 4 sub-phone in this situation). The ED between candidate features X_f and target features X_g, is:

$$ED(\mathrm{X}_f \parallel \mathrm{X}_g) = \sum_{i=1}^{N}(x_{fi} - x_{gi})^2 \quad (4)$$

where x_{fi} is the speech parameters in *i-th* relative frame in the candidate phone, x_{gi} is the speech parameters in *i-th* relative frame in the target phone, and *N* is the number of the relative position.

4.3. Concatenation cost

The concatenation cost which includes acoustic (spectra, energy and F0) and contextual cost is trying to make spectra and prosody smoothing for the synthesized speech. The final concatenation cost will be the sum of the acoustic and contextual concatenation cost. And the calculation will consider the desired generating utterance's contextual information and the degree of the expression. For concatenation cost, we simply used deviation between two speech units:

$$Concatenation\_cost = R_{DC} * \left(w_{F0} * D_{F0} + w_{energy} * D_{energy} + w_{spec} * D_{spec} + w_{\mathrm{context}} * D_{\mathrm{context}} + w_{\mathrm{env}} * D_{\mathrm{env}}\right) \quad (5)$$

where D_{F0}, D_{energy}, D_{spec}, $D_{context}$ and D_{env} are the deviation of F0, energy ,spectral, context information and the environment information from auto-labeling model between two speech units, and w_{F0}, w_{energy}, w_{spec}, w_{context} and w_{env} are their corresponding weight value. R_{DC} is a coefficient that describe the degree of connection between two tagets units.

4.4. Best unit series selection

All in all, our cost denifition is comprised by two parts: the concatenation cost and the target cost. The formula is as follows:

$$Cost = w_{target} * Target\_cost + w_{cat} * Concatenation\_cost \quad (6)$$

The weights are not assigned equally. For instance, the weightes related to prosody parameters like F0 are normally higher than others. Based on the cost definition in Equation (6), a Viterbi search algorithm will be used to find the best path with the minimum cost. The final unit selection results will be found from this path.

5. System Building for Blizzard 2017

5.1. Speech database

The speech database is the British English Speech Corpus for the Blizzard Challenge 2017, which is produced by Usborne Publishing. It contains about 6.5 hours of speech data from professionally-produced children's audiobooks, which is recorded by a single female talker. This includes the approx. 4 hours of pilot data from last year's Blizzard Challenge. A sentence-level alignment between text and speech for some of the data is provided by Toshiba's Cambeidge Research Laboratory.

The task (**Main task 2017-EH1: UK English Children's Audiobooks**) is to build a voice from this provided data that is suitable for reading audiobooks to children.

5.2. Building system

The speech corpus consists of high quality, clean speech data under controlled recording condition. Speech signal is down sampled at 16 kHz frequency, windowed by 25-ms Blackman window for each frame with 5-ms shift, then 40 th order Linear Spectral Pair (LSP) coefficients and fundamental frequency F0 in log scale are extracted as static features. The delta and acceleration components are appended to the static features to form the observation vector for conventional HMM training. Multi-space Probability Distribution HMM (MSD-HMM) of 5 states, left-to-right with no skip topology are used to represent basic speech units. Single Gaussian with diagonal covariance matrix is used in each HMM state. Speech waves are forced aligned with its text transcription by HTS tool HSMMAlign [15]. The case 1 algorithm in [16] is used throughout our experiments for its simplicity.

Concerning the textual features used for training the BiLSTM based phone duration and acoustic model, a varity of linguistic features are used, such as the phone identity, POS and etc. All features in full label is encoded to numeric values and normalized, to be exact, nominal feature such as phone identity is encoded with one hot method, and numeric feature is divided by its maximum value. All encoded values are then concatenated as predictive veators of 341 dimensions to train BiLSTM based phone duration model. These predictive vectors of 341 dimensions, together with the duration position vector of 2 dimensions, is consisted of the input vectors to train the BiLSTM based acoustic model.

For both BiLSTM-based (including duration and acoustic model) systems, a 3-layer neural network consisting a single non-recurrent layer, followed by 2 stacks of bidirectional layers (each with 256*2 LSTM hidden units) is used. All networks are trained with a momentum of 0.9, an initial learning of 0.0005 for the first 5 epoch, and then decreases by 20% after each epoch.

5.3. Internal evaluation

5.3.1. Concatenation cost calculation approaches

We also conducted a small scale listening test to compare different concatenation cost calculation approaches and their combinations. Therefore, 4 systems were compared:

1) Unit selection based on only 3 types of acoustic parameters (f0, energy,lsp)
2) Unit selection based on only 3 types of acoustic parameters (f0, energy,mfcc)
3) Unit selection based on only 4 types of acoustic parameters (f0, energy,lsp,mfcc)
4) Unit selection based on combination of acoustic parameters (f0, energy,lsp,mfcc) and contextual information

Five listeners, all of them are majored in speech related field, took part in the test. For each system, 20 sentences were played to each listener. The listeners were asked to give a 5-point mean poinion score (MOS) for each sentence they had heard. The results are shown in Figure 2. It appears that the system 4, which uses the combination of coustic parameters (f0,energy,lsp,mfcc) and contextual information, outperformed the other three systems. Therefore, system 4 is employed to generate our final submission voices.

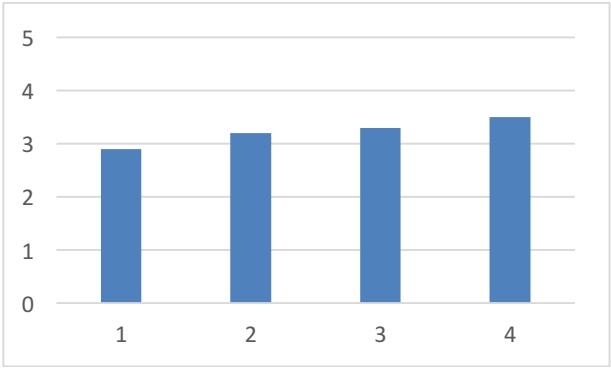

Figure 2: *MOS of the four system using different* concatenation *cost calculation approaches.*

5.4. Evaluation results

13 participants attend the evaluation for **Single task 2016-EH1.** The naturalness (MOS), similarity (MOS) and intelligibility (word error rate (WER)) were calculated. The identifier of our team is F. The results are shown in Figure 3-Figure 6.

5.4.1. Discussion of the results

From the evaluation result, there is still a great gap between our system to the top one. There are many reasons leading this results. And the mainly one is that there still exist some inconsistency between two concatenation units, which would effect the listeners' impression. And the text analysis module is only based on the festival toolkit, which may not quite accurate as we checked some of sentences. The unaccurate text analysis would have an undesirable consequences to the HMM training, force alignment, BiLSTM based acoustic model training and BiLSTM based duration model training. These results reminder us there is still many works need to be done, especially on improving the accuracy of the text analysis.

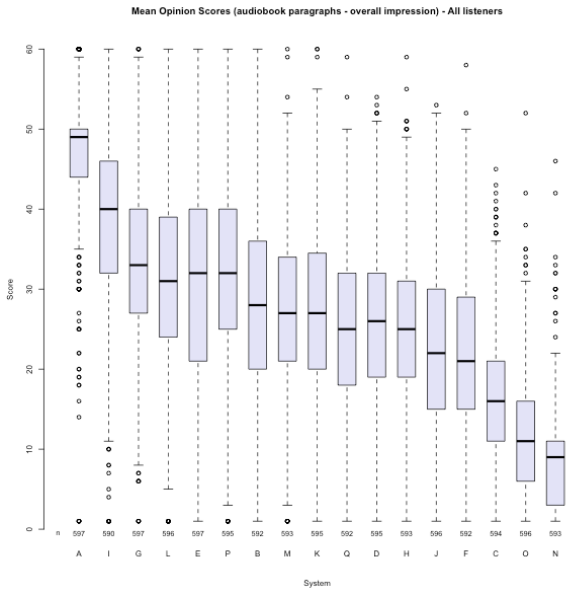

Figure 3: *Boxplot of MOS on overall impression*.

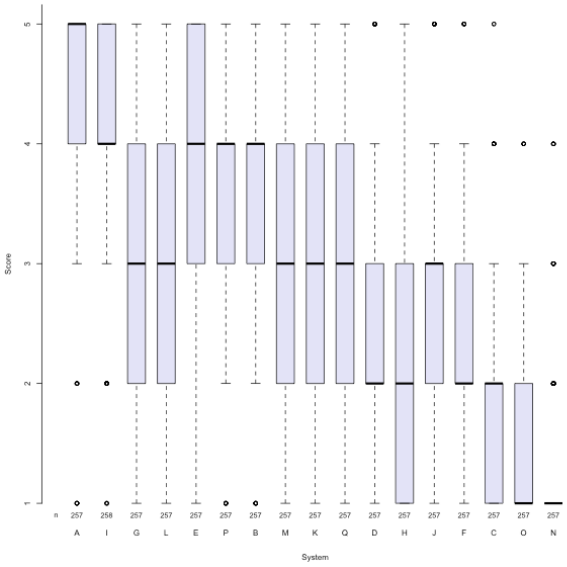

Figure 4: *Boxplot of MOS on similarity evaluation*.

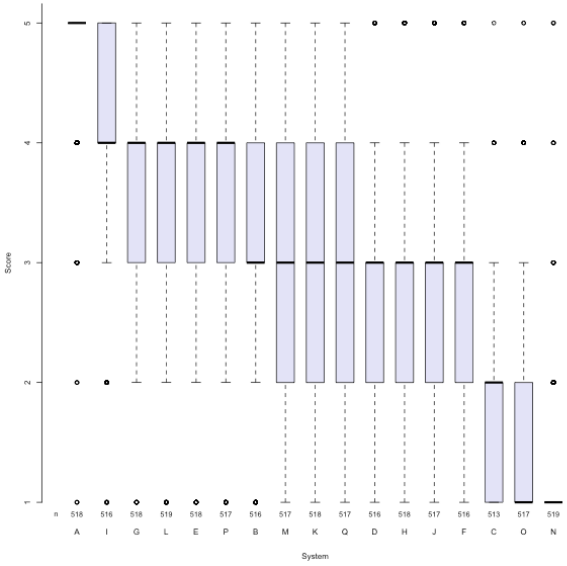

Figure 5: *Boxplot of MOS on naturalness evaluation*.

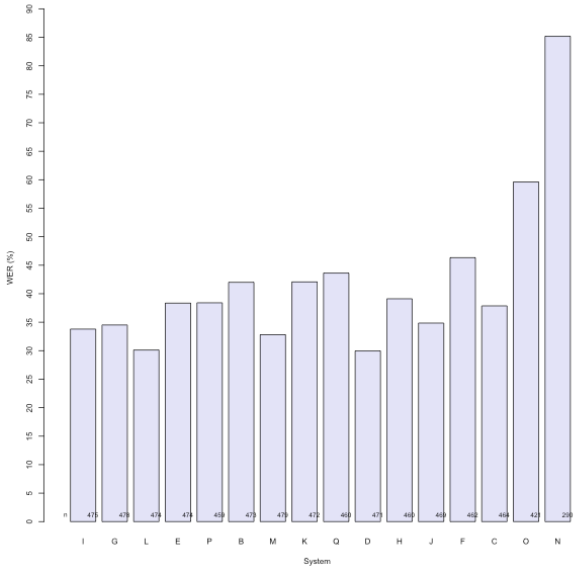

Figure 6: *Word error rates of all participants*

6. Conclusion

In this paper, the BiLSTM based unit selection speech synthesis system built for Blizzard Challenge 2017 by CASIA is introduced. There are three differences from our previous Challenge system.The first one is the improvement in the the calculation of the concatenation cost . The second one is the use of self-training for data selection and categories. The final one is the new target cost calculation approaches. The internal evaluation results show that the effectiveness of these three techniques. Also, the evaluation resuls from the Blizzard Challenge committee shows that, the naturalness, similarity and intelligibility of our system are of average level. Many works need to be done, especially on improving the consistency between two concatenation units .

7. Acknowledgements

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No. 2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61305003, No.61425017，No.61403386), the Strategic Priority Research Program of the CAS (Grant XDB02080006) and partially supported by the Major Program for the National Social Science Fund of China (13&ZD189) and the National Key Research & Development Plan of China (No. 2016YFB1001404).

8. References

[1] Heiga Zen, Keiichi Tokuda, and Alan W Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] Andrew J Hunt and Alan W Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP. IEEE, 1996*, vol. 1, pp. 373–376.

[3] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2011," in *Proc. Blizzard Challenge*, 2011.

[4] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2012," in *Proc. Blizzard Challenge*, 2012.

[5] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2013," in *Proc. Blizzard Challenge*, 2013.

[6] Simon King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, 2014.

[7] J.H. Tao, Y.B. Zheng, etc, “A BLSTM Guided Unit Selection Synthesis System for Blizzard Challenge 2016,” in Proc. Blizzard Challenge, 2016.

[8] Zhen-Hua Ling, Heng Lu, Guo-Ping Hu, Li-Rong Dai, and Ren-Hua Wang, "The USTC system for Blizzard Challenge 2008," in *Proc. Blizzard Challenge*, 2008.

[9] Zhi-Jie Yan, Yao Qian, and Frank K Soong, "Rich-context unit selection (RUS) approach to high quality TTS," in *Proc. ICASSP*, 2010, pp. 4798‐4801.

[10] Yao Qian, Frank K Soong, and Zhi-Jie Yan, "A unified trajectory tiling approach to high quality speech rendering," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 21, no. 2, pp. 280‐290, 2013.

[11] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*, 2015.

[12] Heiga Zen, "Acoustic Modeling in Statistical Parametric Speech Synthesis - From HMM to LSTM-RNN," in *Proc. MLSLP*, 2015.

[13] Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen M Meng, and Li 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.

[14] Festival [online]. Available: http://www.cstr.ed.ac.uk/projects/festival/

[15] HTS [Online]. Available: http://hts.sp.nitech.ac.jp/

[16] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Speech parameter generation algorithms for HMM based speech synthesis," in *ICASSP*, 2000, pp. 1315-1318.