The Alibaba-iDST Entry to Blizzard Challenge 2017

Heng Lu, Ming Lei, Zeyu Meng, Yuping Wang, Miaomiao Wang

Alibaba-iDST

{h.lu,lm86501,mandy.mzy,miaomiao.wmm}@alibaba-inc.com,yuping.wyp@taobao.com

Abstract

This paper describes the the Text-to-Speech system by Alibaba-iDST in the Blizzard Challenge 2017. This is the first time Alibaba-iDST joined the Blizzard Challenge, but the Mandarin version of submitted system has been under development within the company for years. Our submission is a hybrid system composed of both acoustic modeling module, and the unit-selection module to conduct the final unit selection and waveform rendering. Multi-task DNN-BLSTM model is applied for pitch and U/V identification modeling, while the other 5 target and concatenation statistical models are implemented using MGE-HMM. Since data selection is of crucial importance to unit-selection, we include multiple data pre-processing steps in our voice building process. Speaker verification model based likelihood score is employed to prune low similarity units. And energy based waveform normalization is also conducted to brought the volume of periods of speech to the same level. Meanwhile, various cross-utterance tags were also designed in our system to increase the expressiveness of the synthesis speech.

Index Terms: speech synthesis, hybrid TTS system, Blizzard Challenge 2017, BLSTM, HMM, MGE training, deep neural network, recurrent neural network.

1. Introduction

Blizzard Challenge has been held every year to evaluate Text-to-Speech(TTS) methods since 2005[1, 2, 3]. In Blizzard Challenge 2017 hub task EH1, about 6.5 hours of British English speech data from a single female talker is release as training data set. All the speech data is from professional story books produced by Usborne Publishing[4] for children and youth of various ages. The speaking style and degree of expressiveness also differs according to different groups of audiences, which makes building expressive TTS system upon the data a challenging task.

This is the first year Alibaba-iDST attend the Blizzard Challenge. Our system is an statistical model driven unit-selection and concatenation TTS system – the hybrid TTS system. Motivated by the many successful applications of Deep Neural Networks(DNN) models in Speech Recognition[5, 6, 7] and TTS[8, 9, 10], and of course also inspired by many Blizzard Challenge entries from previous years[11, 12, 13], a DNN combined bi-direction long short-term memory(BLSTM) recurrent neural network(RNN) model is trained on pitch then used for guiding the unit selection process along with several Minimum Generation Error(MGE)[14] criterion trained Hidden Markov Models(HMMs). The DNN-BLSTM was trained in the multi-task style, and all the LSFs, log f0 and Voiced/Unvoiced identification bit are trained together, but only generated pitch contour were used in the final unit selection.

The Data pre-process is necessary step to make training database more consistent[13]. A simple speaker verification models, built only using source training data, were employed here to prune away the utterances that does not sound like source speaker. Volume normalization was also conducted here to balance the volume in each group of sentences , since we noticed some very emotional expression may have considerably low volumes, which may stop them from being fully modeled.

Because the tone, sentence accent and break information plays a crucial role in emotions expressing in synthesis speech[11], ToBI[15] tags including 'H' and 'L' phrase tone, sentence accent and break information were trained and predicted by BLSTM models during the training and synthesis phase respectively. Stanford parser for Part-of-Speech(POS)[16] and Name Entity Recognition(NER)[17] were successfully deployed in the Front-End context features label generation steps to make the semantic information richer. Several new cross-utterance context features were designed to capture the longer term emotion trends. On the contrary, multiple position related context features were excluded from context labels to reduce data sparsity problem.

After the speech wave was generate from unit-selection and concatenation, a noise reduction method[18] was applied to the generated waveform as post processing.

The remainder of this paper is organized as follows. Section 2 introduces the data cleaning for the training database. Section 3 describes the proposed hybrid TTS system in details. Section 4 shows Mean Opinion Score(MOS) evaluation results for various tests under different criteria. Discussion and conclusion are given in Section 5.

2. Data preparation

2.1. Data labeling

We manually checked the training text, and labeled them with detailed tags. The tags include more accurate times at phone boundaries, current sentence is emotional or not(this is largely overlapping with the quotation mark), and whether current utterance is valid for training or not(does it sound like speech? Or more like a laughter, scream, animal barking). According to the experience of previous Blizzard Challenge entries, ToBI tones for phrase 'H' and 'L', utterance accent(whether current word is accented) and break information(word boundary, weak phrase boundary, phrase boundary, sentence boundary) were also labeled for ToBI tag modeling and later prediction in synthesis stage. To better model the sentence tone and emotion, sentence type was automatically tagged according to exclamation mark, question mark or quotation mark. Text data shared by other teams from last year Blizzard Challenge was used here for comparison and double check.

2.2. Data pruning

A simple GMM-UBM speaker verification model is built to prune the utterances that not sound like this speaker, such as extremely emotional cries not suitable for modeling. Training data of this model include 100 positive utterances (sound like this speaker) and 100 negative utterances (not sound like this speaker). Both positive and negative utterances are randomly selected and manually labeled. The similarity score to this speaker was calculated by this model for each utterance throughout the whole training corpus, and threshold was set to prune the ones with the lowest similarity score.

Along with speaker verification model, standard HMM model was also trained on whole training corpus. The average log likelihood of each utterance can be obtained from the HMM embedded training process. It is observed the utterances with lower log likelihood tend to have lower text-speech consistency. Therefore, utterances with the lowest log likelihood score were also pruned to prevent having negative impact on system building.

Finally, about 1,000 utterances were pruned from the whole training corpus.

2.3. Volume normalization

During system building process, we found the average volume of the waveform from different stories varies in a large range, and thus the synthesized speech by Statistical Parametric Speech Synthesis(SPSS) sounded quite unstable. Several volume normalization methods were experimented and compared, finally a simple method was selected due to preliminary experiments: the mean-square-energy of each utterance were normalized to a fixed value - to make sure the volume is not too low and no clipping issue produced. This makes the synthesized waveform sounds more stable.

3. System description

3.1. Front-End

3.1.1. Festival OALD

For the Front-End in our system, we used Festival[21] with Oxford Advanced Learners' Dictionary(OALD) for British English lexicon. But we found some inconsistency in the Festival OALD dictionary afterwards – almost all "y" at the end the words are labeled as the long "ii", e.g. "only" ((ou n) 1) ((l ii) 0), "amazingly" (((@ m) 1) ((ei z) 1) ((i ng) 0) ((l ii) 0)), but the word "key" is also labelled as "key" (((k ii) 1)). It does not have any discrimination between the long "ii" and short "i". This harmed the performance of our TTS system to some extent.

After parsing with the Fesitval Front-End, generated parsing files were then transformed to the HTS[22][23] style labels for MGE-HMM training, and at the same time, also converted to the one hot and numerical style for feeding into neural networks for multi-task pitch modeling.

3.1.2. ToBI

Given the labelled 6.5 hour database with "ToBI phrase tones" for 'H' and 'L', "utterance accent" and "break information", we built a multi-task BLSTM model with 2 512-nodes layer for ToBI label prediction. Input feature included 200-order "word embedding of current word", "Part-of-Speech of current word", "punctuation after current word", "whether current word is in quotation marks", "emotion tag of current utterance" and "the case style of current word". The word embedding was obtained from additional huge text data collected from webpages. This multi-task model have 2 tasks, i.e. 2 different output softmax layers: first task for ToBI break prediction, and second task for ToBI tone prediction.

Table 1: *Newly added context labels for acoustic modeling*

Context Tag Name	Description
ELT	English Learner Tag
EMO	Emotional Tag
Sentence_Type	Current sentence is narrative, interrogative or exclamatory
P_ELT	ELT of previous sentence
P_EMO	EMO of previous sentence
P_Sentence_Type	Sentence_Type of previous sentence
F_ELT	ELT of the following sentence
F_EMO	EMO of the following sentence
F_Sentence_Type	Sentence_Type of the following sentence
SPOS	Part-of-Speech from Stanford POS Parser
Name_Entity	Name Entity class from Stanford Name Entity Recognizer

To evaluate this ToBI break and tone model, about one tenth utterances of the whole training corpus were held out as validation set, while all other data were used for training. The final average F-score on break prediction is 0.92, while on tone is 0.535. We believe the reason for high F-score of our ToBI model may due to the fact that phrase boundary always comes along with a comma. Another reason for the high F-score for break may also because of the average length for utterances in the story training corpus is short.Of course, since we have manually checked the labels and utterance alignments for training corpus, the training data for ToBI tag prediction is more consistent.

Another BLSTM model with the same structure as ToBI break and tone model was employed here for the utterance accent prediction. The input feature of accent prediction model is same as ToBI break and tone prediction. The output for the top softmax layer is for binary class indicating whether current word is accented or not. The final F-score on predicting ToBI utterance accent is 0.565.

3.1.3. Other context labels

Besides basic context labels obtained from Festival and the ToBI tags, several new tags were designed for better modeling the prosody. Stanford Part-of-Speech(POS) parser were employed here to extract more accurate POS tag. Because the output POS classes number by Stanford Parser is around 30 and that is too much for modeling, we then design a mapping table to map them to fewer commonly used classes. Stanford Name Entity Recognizer is also used here to analyze whether current word is a person name, place name, organization name, or none of above.

As described in previous section, we have manually labelled the emotion tag for the training corpus, but later find it largely overlapping with whether there is quotation mark. And in the mean while, In order to better model long range emotional changes, cross utterance level emotional tags are also included in our training label. Table 1 describes all the additional tags added to our system.

As shown in table 1, the English Learning Tag(ELT) was designed to capture the age information of the target objects for each of these stories, which is given along with the database. The ELT has 5 levels from baby(18 months+), Elementary(4+), Lower Intermediate(4+), Upper Intermediate(5+), to Advanced(6+). We designed the ELT tag for the reason that

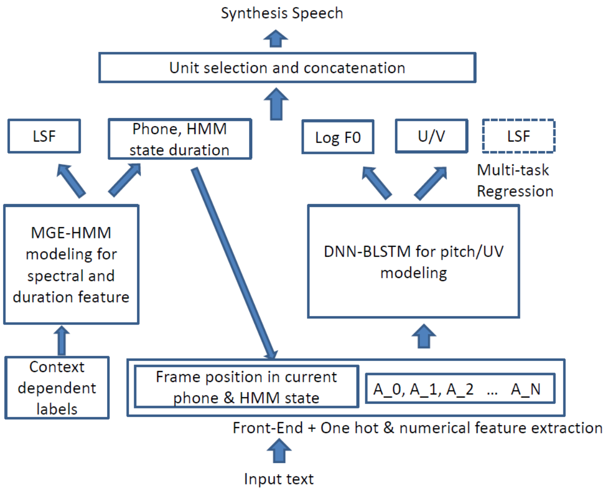
Synthesis Speech
Unit selection and concatenation
LSF
Phone, HMM state duration
Log F0
U/V
LSF
Multi-task Regression
MGE-HMM modeling for spectral and duration feature
DNN-BLSTM for pitch/UV modeling
Context dependent labels
Frame position in current phone & HMM state
A_0, A_1, A_2 ... A_N
Front-End + One hot & numerical feature extraction
Input text

Figure 1: *back-end acoustic modeling architecture.*

stories books for lower age English Learners were observed to be read slower with huge emotion, while on the contrary, facing with more Advanced readers, stories books tend to read more quickly with much more descriptive paragraphs. In the synthesis stage, all the ELT tags were set to Lower Intermediate(4+), which made the synthesis speech slower but clearer and with more emotion.

3.2. Acoustic models

The back-end acoustic modeling architecture of the Alibaba-iDST system is shown in Fig. 1. The statistical acoustic parameter modeling part is composed of an HMM model training module for the spectrum and phone & HMM state duration modeling, as well as a DNN-BLSTM training module for pitch modeling. Line Spectral Frequencies(LSF) based HMM models were trained by the MGE criterion for spectrum feature modeling. 25 dimensional LSF features are extracted along with pitch and Voice/Unvoice information. DNN-BLSTM network with a bottom layer DNN with 1024 nodes and 3 upper layers of BLSTMs with 512 nodes were implemented for pitch modeling and prediction. DNN structure is designed here because it worked better for the bottom feature extraction. We also found adding one DNN layer at the bottom layer of the structure leads to faster overall convergence. Before modeling pitch with DNN-BSLTM, continuous log f0 is firstly generated from linear interpolation between voiced part though out the training corpus. In the training stage, multi-task training was conducted to predict all the continuous log f0, LSF and Voice/Unvoice indicator streams simultaneously with each task weight set to 1.0. Delta and acceleration log f0 were also included in the DNN-BLSTM output target acoustic vector. One hot & numerical style converted context linguistic features introduced in previous subsection were used as input feature in the neural network training process. Both the ”current frame positions in HMM state” and ”current frame positions in phone” were included in the DNN-BLSTM training input feature vector to increase the model accuracy. All input and output feature were normalized to 0-mean and unit-variance. Meanwhile, Some context labels representing absolute frame positions in phrase or utterance were removed from back-end acoustic modeling input vector to increase the robustness of the model. The identity of the phone before previous phone and the identity of the phone after the following phone were also removed after experiments and comparison, here we use only tri-phone as our phone identity context.

Instead of replacing the whole HMM framework to neural networks, in the proposed system, only pitch was modeled and predicted with DNN-BLSTM structure. Because in our system construction process, we found DNN-BLSTM contribute most to the prosody contour generation instead of spectral features, especially in the hybrid TTS systems. During some informal experiments, we compared the system proposed and system without DNN-BLSTM pitch coutour as guidance, and found that the proposed system has better prosody especially with emotional tag. Compared with HMM-based SPSS system, the pitch contour generated by DNN-BLSTM model was observed to be more smooth and more expressive, carrying more emotional information to audience.

Besides the target acoustic models for the LSF, pitch and HMM state duration, concatenation acoustic models were also trained for the 3 streams using the HMM modeling framework. Given the acoustic feature differences at the phone or HMM state boundaries, 3 HMM based concatenation models were trained respectively to facilitate the unit selection process along with target models in the synthesis phase.

3.3. Unit-selection and concatenation

In unit selection step, Kullback Leibler divergence(KLD) calculated out from HMM decision trees nodes’ distribution were used for pre-selection. Candidates number were reduced to 200 at each phone position(phone is our basic unit). Then the dynamic programming process – the Viterbi algorithm was implemented to generate the optimal unit sequence. In the Viterbi process, all target costs, concatenation costs and KLD distances are summed together with it’s own steam weight to produce the final sequences score. As shown in equation 1, both HMM likelihood score and the minus error between candidate pitch contour and DNN-BLSTM generated target pitch contour were added together to make the final cost.

$$
\begin{aligned}
U^* =& \underset{U}{\text{argmax}}\{\sum_{m=1}^{M}\omega_m[\log P(X(U,m)|C,\lambda_m)] \\
&- \omega_{pitch} Distance(CT_{candidate}|CT_{gen}) \\
&- \omega_{KLD} D(KLD)\}
\end{aligned}
\tag{1}
$$

where λ_m with $m = 1, 2, ..., M$ indicates the HMM acoustic models for target LSF, target state duration, concatenation LSF, concatenation pitch, concatenation state duration feature streams respectively. Here $M = 5$. $\log P(X(U,m)|C,\lambda_m)$ is the log likelihood for the acoustic features of candidate unit given target context dependent HMM model. For target cost of pitch, we calculated the Euclidean distance $Distance(CT_{candidate}|CT_{gen})$ of the scaled candidate unit pitch contour $CT_{candidate}$ given the DNN-BLSTM generated pitch contour. Candidate unit pitch contour was normalized to the same length as DNN-BLSTM generated pitch contour before calculating Euclidean distance. Minus KLD distance $-D(KLD)$ was also incorporated in the total cost calculation. ω_{KLD}, ω_{pitch} and ω_m, and are the weights for KLD, the pitch stream and other target and concatenation steams respectively.

We would like to mention that there was a bit inconsistency in our HMM model training and the synthesis phase, since we were using MGE criterion for HMM model training to minimize the generation error between generated acoustic feature

Table 2: *7 MOS tests*

MOS sections	Description
section 1:	Mean opinion scores - book paragraphs
section 2:	Mean opinion scores - book paragraphs
section 3:	Mean opinion scores (naturalness) - book sentences
section 4:	Mean opinion scores (naturalness) - book sentences
section 5:	Similarity with original speaker
section 6:	Semantically unpredictable sentences
section 7:	Semantically unpredictable sentences

and the real target acoustic feature in training corpus, but in the unit-selection process, we calculate the log likelihood instead of generation error from trained HMMs for conducting the unit selection process. However, we have done experiments to compare these two systems previously showing MGE-HMM renders better quality in SPSS, and even slightly better performance in hybrid concatenation in our system framework.

3.4. Waveform renderer and post-processing

After the final optimal unit sequence was chosen, the Waveform Similarity based Overlap-Add technique(WSOLA)[20] is performed to render the synthesis speech waveform. In the end, because we noticed there were sometimes small recording noises in the raw recordings as well as in synthesis speech, a noise reduction process[18] is applied to all the synthesis speech to make the output waveform more clean.

4. Results

During the evaluation session, 17 systems were evaluated in 7 sub MOS tests. Details of the tests can be found in table 2. Test section 1,2 were focusing on various criteria on synthesis book paragraphs. These criteria include ”Pleasantness”, ”Stress”, ”Emotion”, ”Speech Pauses”, ”Intonation”, ”Listening effort” and ”Overall impression”. Test section 3,4 on book sentence, 5 on similarity with original speaker, and 6 on speech intelligibility when synthesizing semantically unpredictable sentences.

Amongst all 17 systems, system A is natural speech by human being. System B is the Festival benchmark – an unit selection benchmark system by CSTR. System C and D are HMM benchmark system by HTS and DNN benchmark system by Merlin[24] separately. The ID for Alibaba-iDST system is P. And all other systems are the entries from other participants.

Hundreds of listeners, including volunteers, speech experts, paid listeners, native speakers, non-natives speakers take part in the listening tests. Since paid listeners are native British English speakers and more likely to finish the whole tests with headphones, we focus on results from paid listeners here to get a general impression of the results.

Figure 2 shows the overall impression of synthesized book paragraph. In this test, we ranked the 3rd, lowers than system I, G, and higher than the other systems. For the 3 benchmark systems, the Festival benchmark outperforms the DNN and HMM benchmarks and ranked the 5th, and this result shows unit-selection system is still preferred by listeners for its recording speech quality. However, the gap between SPSS system and hybrid unit selection is becoming fast smaller, as the new neural network structure and learning methods emerges fast in the recent 2 years. The WAVNET[25] system is a good example. Our system did pretty good in the emotion expression and intonation modeling as one can see in Fig. 2 and 3. We believe this is due to the DNN-BLSTM model structure which nicely depicts the pitch contour given various emotional conditions. Cross-utterance emotional tags, the ToBI accent tags, also the newly added POS tag and Name Entity tag, added more to the prosody modeling and prediction in proposed system.

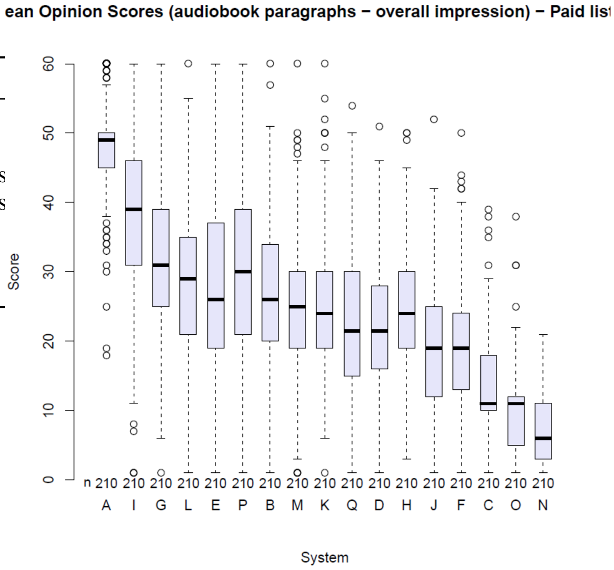

Figure 2: *MOS score for synthesis book paragraphs – Overall Impression.*

Speaker similarity scores are presented in Fig. 4. Comparing SPSS systems, unit selection and concatenation systems has advantage in the similarity test, because they directly use real speech units for concatenation and speech generation. In this test, our system ranked 2nd, at the same level as E and B, and lower than system I. Fig. 5 presents the naturalness score for synthesized book sentences. Our system P ranked 2nd, the same with systm G, L and E, and worse than system I. As in Fig. 5, the word error rate for the listening and writing semantically unpredictable sentences test are described. SPSS systems performs best in this field because they can always generate the most ”correct” but over averaged acoustic feature contours. The DNN benchmark performs the best in this test.

5. Discussion and Conclusions

This paper presents the system building and MOS test results for the Alibaba-iDST entry for Blizzard Challenge 2017. Overall speaking, we did a good job in our system building. Since this is the first year we take part in the worldwide Blizzard Challenge speech synthesis evaluation, we have referred to, and also experimented with many methods proposed by the previous Blizzard Challenge entries. However, we also tested a lot with the new speaker verification based data cleaning method. Designed new cross-utterance and emotional tags for better emotion modeling. And we also tried unit-selection using MGE-HMM with the LSF acoustic feature. In the back-end acoustic modeling phase, both MGE-HMM and DNN-BLSTM were trained. The DNN-BLSTM based pitch modeling was a success for generating the more expressive pitch contour. Finally, the speech enhancement techniques was employed to make the synthesized speech sounds more clean. One obvious defect in

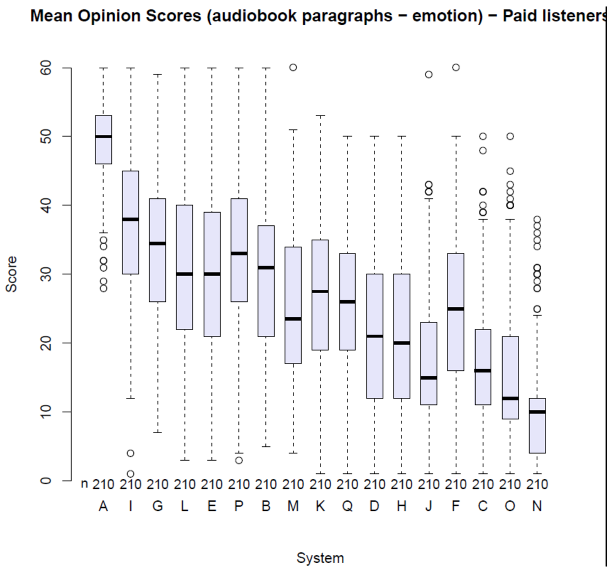

Figure 3: *MOS score for synthesis book paragraphs – Emotion.*

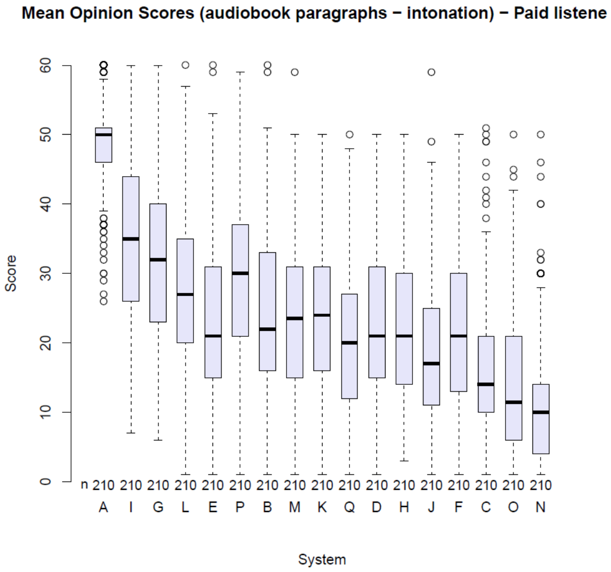

Figure 4: *MOS score for synthesis book paragraphs – Intonation.*

the data preparing process is for the Front-End dictionary. Our system may suffer from the inaccuracies in the Festival OALD lexicon, and this harms the performance of our system.

Generally speaking, we believe unit-selection and concatenation system has some advantage in Blizzard Challenge for the high fidelity speech sound. However, building nice quality unit selection systems requires extremely accurate labeling, which may need plenty of checking work, even checking manually. On the other hand, as the techniques for neural network developing really fast, there emerges more and more new and powerful recipes for feature modeling and learning, like the WAVNET, Generative Adversarial Network(GAN)[26] and End-to-end[27] TTS. The gap between SPSS and unit-selection hybrid system is inevitably becoming rapidly smaller. We are planning to head towards the neural network based SPSS direction, and aims to improve our current system by new neural network structures and learning methods.

6. Acknowledgements

We would like to thank Biao Tian from Alibaba-iDST for providing his noise reduction tools and very professional advises. Thank Yuan Liu from Alibaba-iDST for performing speaker verification model training and inference for the similarity check for our training data. We would also like to thank Zhijie Yan from Alibaba-iDST for his expert suggestions and support from all aspects.

7. References

[1] A. W. Black and K. Tokuda, "The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets," *Proc Interspeech 2005*, Lisbon, 2005.

[2] S. King and V. Karaiskos, “The Blizzard Challenge 2016,” *in Blizzard Challenge Workshop*, 2016.

[3] ”The Blizzard Challenge website,” *http://www.synsig.org/index.php/Blizzard Challenge.*

[4] ”http://www.usborne.com”

[5] G. Hinton, L. Deng, D. Yu, et al. ”Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, 2012, 29(6): 82-97.

[6] R. Collobert, J. Weston. ”A unified architecture for natural language processing: Deep neural networks with multitask learning” *Proceedings of the 25th international conference on Machine learning. ACM*, 2008: 160-167.

[7] G. E. Dahl, D. Yu, L. Deng, et al. ”Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition”. *IEEE Transactions on audio, speech, and language processing*, 2012, 20(1): 30-42.

[8] H. Zen, A. Senior, M. Schuster. ”Statistical parametric speech synthesis using deep neural networks”, *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE*, 2013: 7962-7966.

[9] H. Lu, S. King, O. Watts, ”Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis”, *SSW. 2013*: 261-265.

[10] Z. Wu, C. Valentini-Botinhao, O. Watts, et al. ”Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis”, *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE*, 2015: 4460-4464.

[11] L. Chen, Y. Jiang, M. Zhou, Z. Ling, L. Dai ”The USTC system for blizzard challenge 2016,” *Blizzard Challenge Workshop*, 2016.

[12] J. Tao, Y. Zheng, Z. Wen, Y. Li, B. Liu ”A BLSTM Guided Unit Selection Synthesis System for Blizzard Challenge 2016” *Blizzard Challenge Workshop*, 2016.

[13] T. Merritt, S. Ronanki, Z. Wu, O. Watts, ”The CSTR entry to the Blizzard Challenge 2016” *Blizzard Challenge Workshop*, 2016.

[14] Y. Wu, and R. Wang, ”Minimum generation error training for HMM-based speech synthesis”, *Acoustics, Speech and Signal Processing (ICASSP), 2006 IEEE International Conference on. IEEE*, pp. 89-92, 2006

[15] M. E. Beckman ”Guidelines for ToBI Labelling”

[16] K. Toutanova, D. Klein, C. Manning, and Y. Singer. ”Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network.” *Proceedings of HLT-NAACL* 2003, pp. 252-259.

[17] J. R. Finkel, T. Grenager, and C. Manning. ”Incorporating Non-local Information into Information Extraction Systems by Gibbs

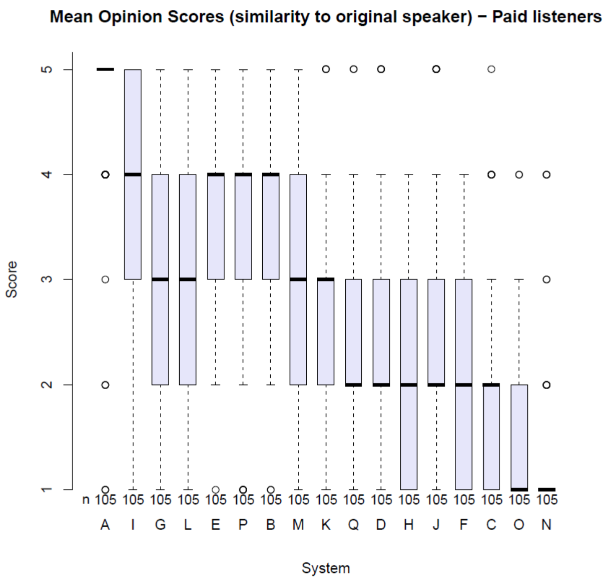

Figure 5: *MOS score for synthesis book sentences – Speaker Similarity.*

Sampling.” *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)* , pp. 363-370. http://nlp.stanford.edu/ manning/papers/gibbscrf3.pdf

[18] I. Cohen, ”ptimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator” *IEEE Signal Processing Lett.*, vol. 9, no. 4, pp. 113116, Apr. 2002.

[19] H. Kawahara,”STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds,” *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 349353, 2006.

[20] W. Verhelst and M. Roelands, ”An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech,” *Acoustics, Speech and Signal Processing (ICASSP), 1993 IEEE International Conference on. IEEE*, Apr. 1993, pp. 55457.

[21] ”http://www.cstr.ed.ac.uk/projects/festival/”

[22] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, ”The HMM-based speech synthesis system version 2.0” , *Proc. of ISCA SSW6*, Bonn, Germany, Aug. 2007

[23] ”http://hts.sp.nitech.ac.jp/?Home”

[24] Z. Wu, O. Watts, S. King, ”Merlin: An Open Source Neural Network Speech Synthesis System” *Proc. 9th ISCA Speech Synthesis Workshop (SSW9)*, September 2016, Sunnyvale, CA, USA

[25] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. (2016a). ”Wavenet: A generative model for raw audio.” *arXiv preprint arXiv:1609.03499.*

[26] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio ”Generative Adversarial Networks” *arXiv:1406.2661 [stat.ML]*

[27] Y. Wang, et al. ”Tacotron: Towards End-to-End Speech Synthesis” *arXiv:1703.10135 [cs.CL]*

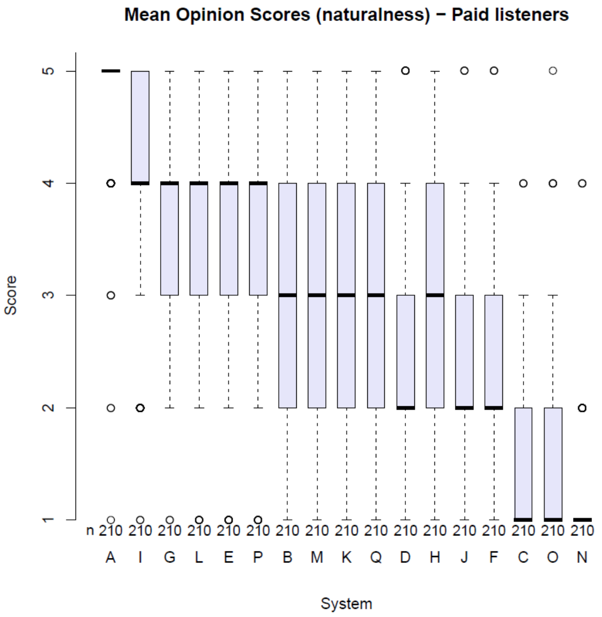

Figure 6: *MOS score for synthesis book sentences – Naturalness.*

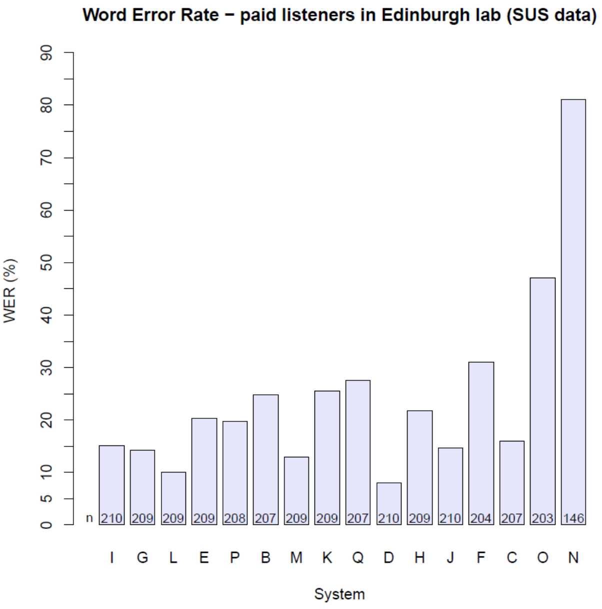

Figure 7: *Word error rate for semantically unpredictable sentences.*