I^{2}R Text-to-Speech System for Blizzard Challenge 2010

Minghui Dong, Paul Chan, Ling Cen, Bin Ma, Haizhou Li

Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore 138632
{mhdong, ychan, lcen, mabin, hli}@i2r.a-star.edu.sg

Abstract

This paper describes I^{2}R's submission to the Blizzard Challenge 2010 speech synthesis evaluation. This is our third participation in the challenge. In this paper, we will describe our main approaches to building the required voices. We will introduce the procedure of database processing, the definitions of the acoustic, prosodic and linguistic parameters, the components of cost functions, etc. Finally, we will look at the listening test results. The evaluation results show that our Mandarin system performed well in the evaluation.

Index Terms: speech synthesis, unit selection, cost function, and Mandarin text-to-speech.

1. Introduction

Blizzard Challenge [1-3] provides speech synthesis researchers a good opportunity to evaluate the corpus-based speech synthesis technology developed by different teams using the same database. Same as the previous year, this year's evaluation is still focused on English and Mandarin speech synthesis. However, some tasks are different from previous years.

For English, the organizer provided two British English databases, "rjs" and "roger". The "rjs" database, which was supplied by Phonetic Arts, contains 5 hours (4000 utterances) of speech recordings of a male professional speaker. It is available in 16kHz and 48kHz sampling rates. Also the standard Festival labels are provided as well. The "roger" database contains 1 hour (1000 utterances) of speech recordings of a male speaker. The hand corrected labels, which was supplied by iFlyTek, and the standard festival labels are supplied. For Mandarin, the organizer provided a 9-hour (6000 utterances) database, which was supplied by the Chinese Academy of Sciences.

Participants may choose to enter the evaluation of one or both languages. For each language, there are hub tasks and spoke tasks. The hub tasks require all participants to build synthetic voices using a big database (6 hours speech data for English and 9 hours speech data for Mandarin) and a smaller database (about 1 hour speech). The optional spoke tasks test the following: (1) Synthesizing speech with 100 utterances of training data. (2) Synthesizing speech that is suitable for noisy environment. (3) Synthesizing speech with higher sampling rate (48 kHz instead of 16 kHz for other tasks).

2. Overview of Our Approach

The unit selection approach [4-7] to speech synthesis has been shown to be one of the best approaches currently used. I^{2}R's blizzard 2010 system adopted the unit selection based approach. Both of our English and Mandarin system is based on the same engine. The methods that we use follow last year's work [8] with some improvements.

The first step in unit selection is database labeling. In our work, we use the automatic forced alignment method employing speech recognition technology. We also use other automatic methods to exclude some possibly defect units.

Prosody parameters, which include pitch, duration and energy information, are usually used to maintain the naturalness of the synthetic speech. However, the spectral suitability of a speech unit is also very important towards the quality of synthesized speech. Therefore, in our system, we defined a set of acoustic parameters that is designed to cover spectral information in our unit features. The cost function is designed to include these parameters as well.

For the Mandarin speech synthesis, instead of using the usual initial-final definitions for each speech unit, we decided to use a smaller phone-sized unit. This allows our system to handle missing syllables easily and makes it possible to generate speech with very small TTS databases.

In the following sections, we will first introduce database processing, prosody model, unit selection process, then we look at the evaluation result, and finally the conclusion of the paper is given.

3. Speech Database Processing

In this part, we explain how we process the speech database.

3.1. Forced Alignment

HTK is used for automatic alignment of pronunciation and the speed data. We defined phone-sized speech segments as our basic unit. 39 dimensional MFCC feature is used for the training of the phone models. The frame size is 25ms and the frame shift is 10ms. Three states are defined for each context-independent HMM model for each phone. Optional silence models are inserted into the phone sequences so that long silences in the utterances are taken care of. The phone models are first trained with the speech corpus. Unit boundaries are then obtained by the forced alignment of speech with its phonetic sequence.

3.2. F0 Calculation

F0 feature is one of the most important features of prosody of speech. We used the Praat software [9] to calculate the F0 of speech utterance. The F0 values of every 0.01 second interval are calculated. To avoid 0 values for unvoiced part of speech data, interpolation is done to give none-zero F0 values for each unvoiced segment. Then we apply a simple smoothing process to this F0 sequence. The smoothing is done with moving average represented with the following formula:

$$p'_i = (p_{i-1} + p_i + p_{i+1})/3 \qquad (1)$$

where p_i is the F0 value of the i-th frame.

4. Prosody Model

We use the same prosody models as last year. In this part, we describe how the prosody model of the speech synthesis system was built.

4.1. The Acoustic Parameters

We first calculated a set of parameters that describe spectral and prosodic features of each HMM state, and boundary frame. These parameters are supposed to include all the possible parameters in our consideration. The main values that we capture include the statistical values of each individual HMM state as well as the values of boundary (start and end) frames of the unit. The initial parameter set that we used consists of the following values:

- Spectral features: MFCC mean for the 3 HMM states, MFCC for boundary frames.
- Pitch features: Mean, maximum, minimum, and range of pitch values and pitch derivative values for 3 HMM states, and boundary frames.
- Duration features: Durations of the 3 states, duration of the unit.
- Energy features: Mean energy of frames in the 3 HMM states, and boundary frames.

The defined parameter set forms a long vector (with a dimension of 308), which contains a lot of redundancy. Therefore, we use the principal component analysis approach to reduce the dimension. The dimension reduced vector is considered a compact form of representation of the prosodic and spectral features of the unit. Finally, we have a 40-dimensional vector for both English and Mandarin.

4.2. The Prosodic Parameters

The acoustic parameters define both spectral and prosodic information. However, because there are more parameters conveying spectral information than those conveying prosodic information that are being defined in the long vector, prosodic information is actually less prominent in the acoustic vector. Nevertheless, we still need a set of prosodic parameters to emphasize the prosodic properties in speech. The prosodic parameters for each unit consist of the following:

- Pitch mean of the unit
- Duration of the unit
- Energy mean of the unit
- Pitch range of the unit.

4.3. Linguistic Features

Linguistic features are derived from input text. They are used for predicting the acoustic parameters. Due to differences in the languages and available resources for the each of them, we have defined different linguistic features for English and Mandarin.

The English corpus comes with the utterance structure for each speech file. We have defined the features for it similar to those that are used in the HTS system [10]. We have derived the following linguistic features from the utterance files (the number of parameters are given in brackets):

- Context units: phone identities of the previous 2 and next 2 units. (4)
- Syllable information: Stress, accent, length of the previous, current and next syllables. (9)
- Syllable position information: syllable position in word and phrase, stressed syllable position in phrase, accented syllable position in phrase, distance from the stressed syllable, distance from the accented syllable, and name of the vowel in the syllable. (13)
- Word information: length and part-of-speech of the previous word, current word and next word, position of the word in phrase. (12).
- Phrase information: Lengths (in number of words and syllables) of previous phrase, current phrase and next phrase, position of the current phrase in major phrase, boundary tone of the current phase. (8)
- Utterance information: Lengths in number of syllables, words and phrases. (3)

Putting all the features together, we form an input linguistic feature vector of 53 elements for English.

For the Mandarin corpus, we have defined less linguistic features. The features we used include:

- Context units: phone identities of the previous 2 and next 2 units. (4)
- Tone information: The tones of the current, previous two and next two syllables. (5)
- Phone location in syllable: Number of phones in the syllable, position of the phone counting from left boundary, position of the phone counting from right boundary. (3)
- Word information: length and part-of-speech of the previous word, current word and next word, position of the syllable in word. (8).
- Prosodic phrase information: Lengths of prosodic phrases of different levels, syllable locations of prosodic phrases of different levels. (12)

Altogether, we have a linguistic feature vector of 32 elements for Mandarin.

4.4. Parameter Prediction

The acoustic parameter prediction process calculates the parameters from the linguistic features. The prediction can be represented with the following formula:

$$y_i = F_i(X) \qquad (2)$$

where y_i is the i-th parameter for the unit and X is the linguistic feature vector for the unit.

In our system, the linguistic features are the predictors and the acoustic and prosodic parameters are the responses. We build our models using the CART [11] approach. Each individual parameter is predicted separately with a CART tree.

5. Unit Selection

Unit selection method is used in all the voices that we have built. In this part, we describe how we define the cost function.

The unit selection process is based on the cost function that consists of two parts (1) a target cost to measure the difference between the target unit and the candidate unit. (2) a join cost to measure the acoustic smoothness between the concatenated units.

Our target cost further consists of three parts (1) the cost of acoustic parameters, (2) the cost of prosodic parameters, and (3) the cost of context linguistics features. The target cost c_t is defined as the following:

$$c_t = w_{ta}c_{ta} + w_{tp}c_{tp} + w_{tl}c_{tl} \quad (3)$$

where, c_{ta} , c_{tp} and c_{tl} are the cost of acoustic parameters, prosodic parameters and linguistic features respectively, and w_{ta} , w_{tp} and w_{tl} represent their corresponding weights.

The reason why we use three cost components here is that each of them alone is not sufficient to describe the target cost. The cost of the linguistic feature is to ensure the general spectral and prosodic accuracy of the candidate unit. However, due to variations in speech, using this cost on its own may easily lead to extreme cases (abnormal spectrum and prosody). The use of cost of acoustic parameters can avoid the selection of the extreme cases, because statistical models favor average values. The use of prosodic cost is to emphasize the importance of prosodic features.

The total cost *c* is calculated with the following function.

$$c = w_t \sum_{i=0}^{n} c_t(i) + w_j \sum_{i=1}^{n} c_j(i) \quad (4)$$

where *n* is number of units in the sequence, $c_t(i)$ is the target cost of unit *i*, $c_j(i)$ is the join cost between unit *i-1* and unit *i,* and w_t and w_j are weights for target cost and join cost respectively.

The best unit sequence is determined by searching for a best path among the candidate unit lattice to minimize the total cost of the selected sequence. Viterbi algorithm is used to find the best sequence. The weights in the cost function are manually tuned.

6. Building Voices

In this year's evaluation, we have participated in the following tasks: EH1 ("rjs" voice, full data set), EH2 ("roger" voice, arctic data set), ES2 (voice for noisy environment) ES3 (48 hHz "rjs" voice), MH1 (full data set), MH2 (small data set: 1000 utterances), and MS1 (very small data set: 100 utterances), and MS2 (voice for noisy environment). All the voices are built using unit selection methods, including MS1 task, where only 100 utterances are used. For the voices for noisy environment, we submitted the same voice as the hub tasks.

6.1. English Voices

English databases consist of three data sets, i.e. 16kHz "rjs" data, 16kHz "roger" data, and 48 kHz "rjs" data. The rjs data consists of 6 hours speech recording, while the "roger" data only consists of 1 hour data for the valuation tasks.

The first thing we need to do is the labeling of phonetic units in the database. As the data comes with Festival labels, we extract the phone sequence of each utterance from the file. We performed force alignment for the 16 kHz "rjs" data and the 16kHz "roger" data. Then we directly use the labels generated from the 16kHz data set as the labels for the 48kHz data set.

For both "rjs" and "roger" data, the Festival utterance files were provided. We directly use the linguistic features derived from the utterance file for our training and synthesis. Though the roger data set also provided manually corrected data, we did not use it as its format is incompatible with our format. We have built voices for EH1, EH2, ES2, and ES3. The voice for ES3 task is actually the same as that for EH1.

6.2. Mandarin Voices

The mandarin database consists of about 9 hours speech recording. Machine generated pronunciation, word sequence, prosodic words, and part-of-speech (POS) information was provided. We made use of the pronunciation to do forced alignment, and also generated linguistic features from the provided annotations.

Mandarin is a syllable based language, in which each Chinese character is pronounced as a mono-syllable. There are about 408 base syllables in Mandarin. Each base syllable can be decomposed into an Initial-Final structure similar to the Consonant-Vowel relations in other languages. Each base syllable consists of either an Initial followed by a Final or a single Final. The Initial is the initial consonant part of a syllable and the Final is the vowel part including an optional medial or a nasal ending. In Mandarin Chinese, there are 22 different initials (including a null-initial) and 38 different finals [12].

Table 1. *Initial and Finals of Mandarin*

22 Initials	b c ch d f g h j k l m n p q r s sh t x z zh null-initial
38 Finals	a ai an ang ao e ei en eng er i ia ian iang iao ie in ing iong iu iz izh ong ou u ua uai uan uang ueng ui un uo v van ve vn

In our system, we further divided the finals into 1-4 phonemes, similar to the phone set used for English speech recognition. Hence, we defined 43 phones as shown in Table 2. The advantage of using the smaller unit is that we are able to handle missing syllables easily.

Table 2. *Mandarin phone set*

18 vowels	a aa ah e ea ee een eeng eh er i iz izh o oh oo u v
25 consonants	b c ch d f g h j k l m n ng p q r s sh t vh wh x yh z zh

Because we used a smaller unit size, there are more unit candidates available despite the small data collection. In task MS1, we calculated the number of units in the first 100 utterances. Totally, there are 4515 units in the small data set, which means an average of 100 units for each phone. Therefore, we use the data set as a unit selection database.

7. Results and Discussions

The organizers of the Blizzard Challenge 2010 has conducted listening evaluation and released its results. This helps us to better understand the performance of the method we used in the system.

This is the third time that we participated in the evaluation. Although the different databases were used in this year's and last year's evaluations, the MOS scores may give us an approximate comparison. Comparing the evaluation results of this year to the last, we have noticed that the median of naturalness MOS for Mandarin voice (MH1) for the all listener category remains at 4. However, the median of naturalness for English voice (EH1) has increased from 2 to 3. The median of similarity MOS scores of the English (EH1) and Mandarin (MH1) voices remain the same as those of last year's (3 for English, 4 for Mandarin).

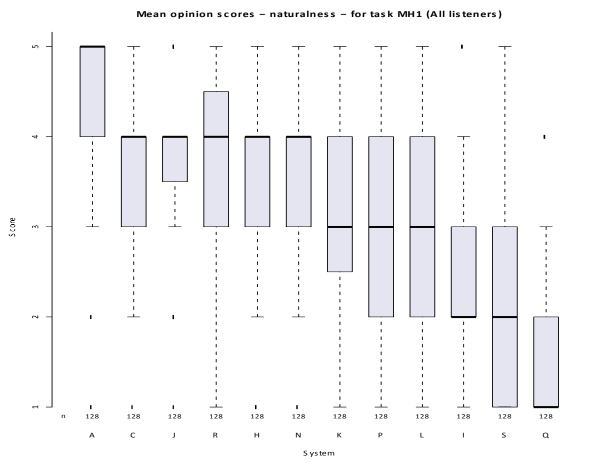

Figure 1. MOS score for Mandarin voice MH1 (All listeners)

The major difference between this year's system and last year's is that this year we have tuned the cost functions. Also we have made a lot of corrections in the code.

Since we have done notably well in Mandarin voices, we will move on to examine the results of Mandarin voices.

7.1. Mandarin Voice MH1

For the Mandarin voice MH1 task, we have achieved a mean natural score of 3.5 and a mean similarity score of 3.4. Figure 1 shows the statistics of naturalness score for voice MH1 from all listeners' feedback. Our system is H in the Figure. From the figure, we can see that our system has achieved a median score of 4. This shows that our method has the potential to achieve high naturalness in synthesizing Mandarin voices.

Figure 2 shows the statistics of similarity score for Mandarin voice from all listeners' feedback. From the figure, we can see that our system achieved a median score of 4 for similarity to the original speaker. This shows that our method is able to retain the speaker's characteristics very successfully.

To examine the intelligibility, we will look at the PTER (Pinyin plus Tone Error Rate). Figure 3 shows the PTER of the MH1 task. From the figure, we know that we achieved error rate of 23% for pronunciation and tone for all listeners. Compared with other systems, this value is in the middle range.

When comparing with the benchmark system C, which is HTS-2005 [13], we noticed that there are no statistically significant differences between our system and system C for both naturalness and similarity. However, system C has higher intelligibility.

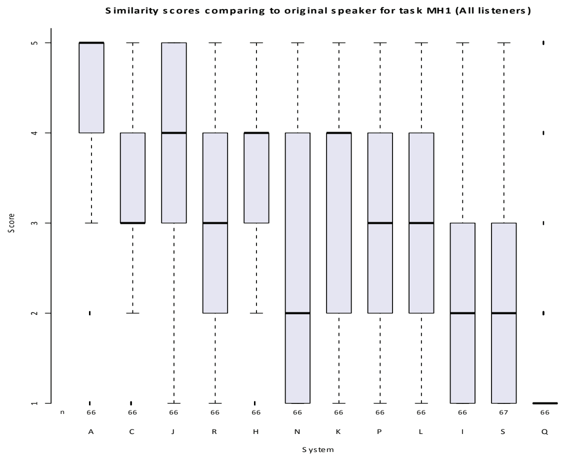

Figure 2. Similarity score for Mandarin voice MH1 (All listeners)

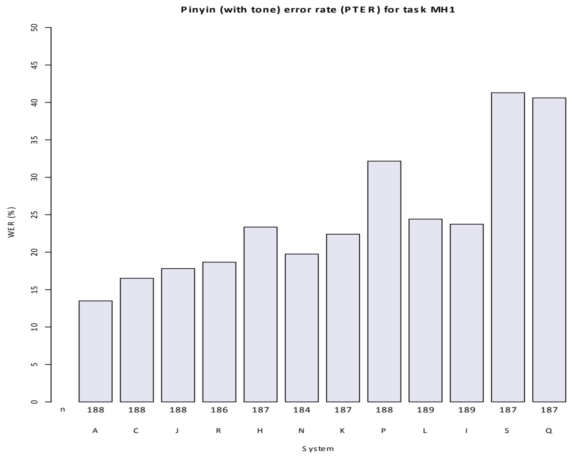

Figure 3. Intelligibility score for Mandarin voice MH1 (All listeners)

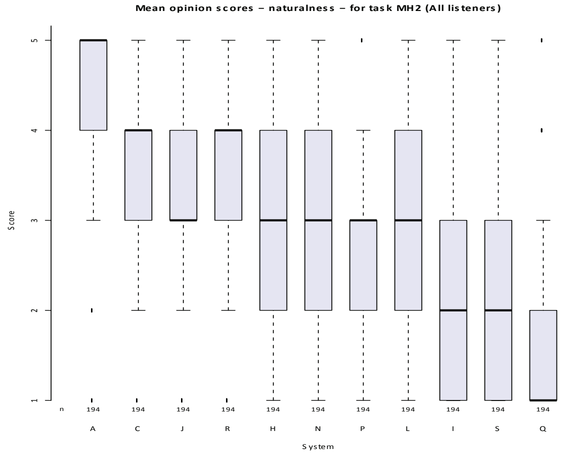

Figure 4. MOS score for Mandarin voice MH2 (All listeners)

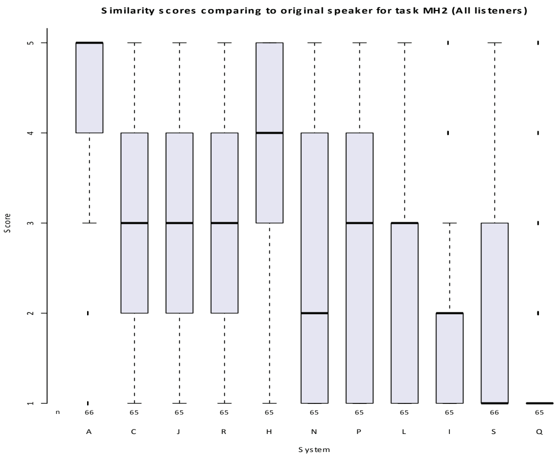

Figure 5. Similarity score for Mandarin voice MH2 (All listeners)

7.2. Mandarin Voice MH2

For the Mandarin voice MH2 task, we have achieved a mean natural score of 3.2 and a mean similarity score of 3.6. Figure 4 shows the statistics of naturalness score for voice MH2 from all listeners' feedback. From the figure, we can see that our system H has achieved a median score of 4. This shows that our method has the potential to achieve high naturalness in synthesizing Mandarin voices for a smaller database (800 utterances).

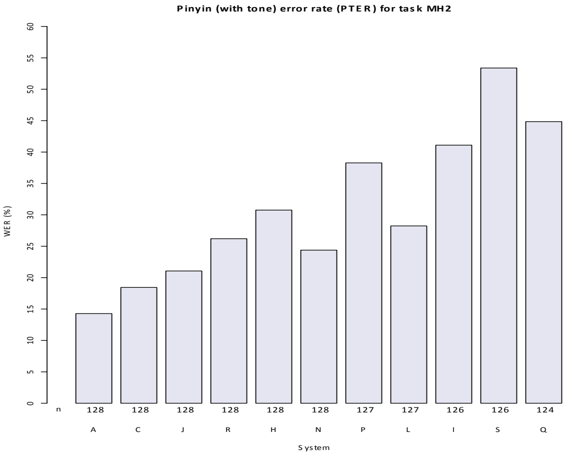

Figure 6. Intelligibility score for Mandarin voice MH2 (All listeners)

Figure 5 shows the statistics of similarity score for Mandarin voice MH2 from all listeners' feedback. From the figure, we can see that our system achieved a median score of 4 for similarity to the original speaker. This is higher than other systems. The statistics results also show that there is no significant difference between our system H and human voice A. This shows that our method is able to retain the speaker's characteristics very successfully in a small size database.

Figure 6 shows the intelligibility scores. The figure shows that the PTER of our system is 31%. This is higher than the result using the full set database.

When comparing with the benchmark system C, we noticed that there are no statistically significant differences between our system and system C for naturalness. However, our system has higher similarity score, and system C has higher intelligibility.

7.3. Mandarin Voice MS1

When building voice MS1, where there are 100 utterances available, we have tried the unit selection based synthesis method. It is remarkable that the results show that our result is comparable to those of other systems (HTS systems).

Figure 7 shows the statistics of the naturalness score for voice MS1 from all listeners' feedback. From the figure, we can see that our system H has achieved a median score of 3. This shows that our method is able to achieve high naturalness in synthesizing Mandarin voices with a very small database.

Figure 8 shows the statistics of similarity score for Mandarin voice from all listeners' feedback. From the figure, we can see that our system has achieved a median score of 3 for similarity to original speaker. This shows that our method has been very successful in retaining the speaker's characteristics when using a very small database.

Figure 9 shows the intelligibility of Mandarin voice MS1. From the figure, we can see that our system is able to achieve comparable intelligibility with other systems.

The success of synthesizing MS1 voice with the unit selection method suggests that, with careful design of speech database, we are able to generate high quality Mandarin speech with a very small data set with unit selection methods.

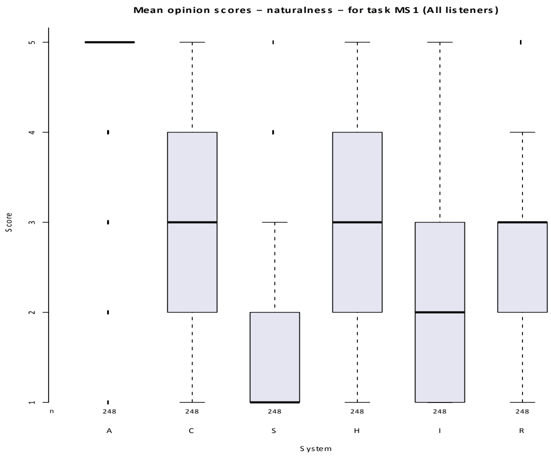

Figure 7. MOS score for Mandarin voice MS1 (All listeners)

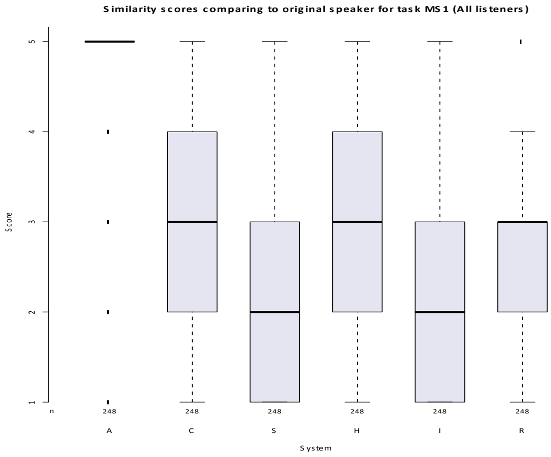

Figure 9. Similarity score for Mandarin voice MS1 (All listeners)

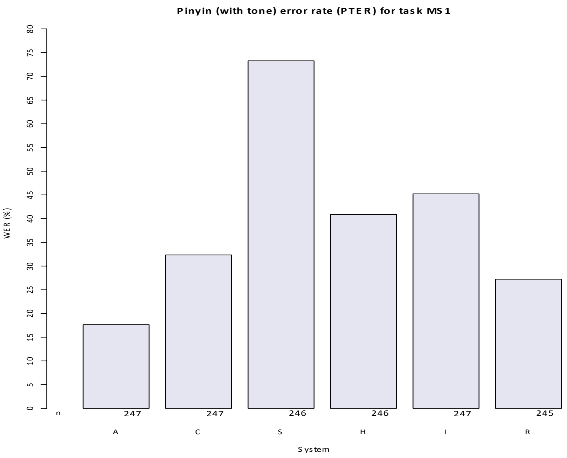

Figure 9. Intelligibility score for Mandarin voice MS1 (All listeners)

7.4. Mandarin Voice MS2

Our Mandarin voice MS2 is the same as MH1. Figure 10 shows the intelligibility of Mandarin voice MS2. From the figure, we can see that our system is able to achieve comparable intelligibility with other systems.

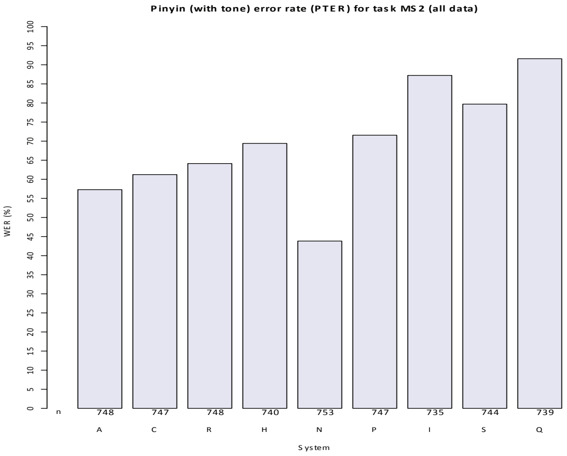

Figure 10. Intelligibility score for Mandarin voice MS2 (All listeners)

8. Discussion and Future Work

Though both English system and Mandarin system use the same speech synthesis engine, we have noticed that Mandarin system works relatively better. It is difficult to compare due to different languages and different database sizes. We will try to figure out the possible reasons and improve the English system.

From the evaluation, we have also noticed that the intelligibility of our English voices is relatively low. More effort should be put to improve it. The statistical parametric approaches have a lot of advantages in synthesize high quality speech. We will try to apply these techniques in our future system.

9. Conclusion

This paper has described our speech synthesis approach for the Blizzard Challenge 2010. We have used the unit selection based approach for all the voices. The evaluation results show that our Mandarin voice is good in both naturalness and similarity. We have also managed to use unit selection for the small database of 100 Mandarin utterances. The evaluation results show the method works well for generating Mandarin speech.

10. References

[1] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, S. King, Statistical Analysis of the Blizzard Challenge 2007 Listening Test Results, Proc. Blizzard Challenge Workshop, 2007, Bonn, Germany.

[2] S. King, V. Karaiskos, “The blizzard Challenge 2009.” Blizzard Challenge Workshop, Sept. 2009.

[3] A. W. Black and K. Tokuda, “The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets,” in Proc Interspeech 2005, Lisbon, 2005.

[4] A. W. Black, P. Taylor, “Automatically Clustering Similar Units for Unit Selection in Speech Synthesis,” in Proc. Eurospeech 97, vol 2 pp 601-604, Thodes, Greece.

[5] R. Clark, K. Richmond, V. Strom, S. King, “Multisyn voice for the Blizzard Challenge 2006,” Blizzard Workshop 2006.

[6] M. Schroder, A. Hunecke, S. Krstulovic, “OpenMary – Open Source Unit Selction as the Basic for Research on Expressive Synthesis,” Blizzard Workshop 2006.

[7] M. Chu, H. Peng, Y. Zhao, Z. Niu and E. Chang, “Microsoft Mulan – a Bilingual TTS System”, Proc. of ICASSP 2003, Hong Kong, 2003.

[8] M. Dong, L Cen, P. Chan, D. Huang, B. Ma, H. Li, “I2R Text-to-Speech System for Blizzard Challenge 2009”, Blizzard Challenge Workshop, Sept. 2009.

[9] Boersma, Paul, “Praat, a System for Doing Phonetics by Computer.” Glot International 5:9/10, 341-345, 2001.

[10] K. Tokuda, H. Zen, A.W. Black, “An HMM-based Speech Synthesis System Applied to English,” in Proc. of 2002 IEEE SSW, Sept. 2002.

[11] L. Breiman, , J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and Regression Trees”. Monterey, Calif., U.S.A.: Wadsworth, Inc., 1984.

[12] H. W. Hon, et al. “Towards large vocabulary Mandarin speech recognition,” in Proceedings of ICASSP 1994. pp:545-548.

[13] H. Zen and T. Toda, “An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005,” in Proc. Blizzard Workshop, 2005.