NTTS participation in the Blizzard Challenge 2008

Feng Ding, Jari Alhonen

Nokia Research Center, Beijing, China
Feng.F.Ding@nokia.com

Abstract

This paper describes NTTS participation in the Blizzard Challenge 2008. The Blizzard Challenge 2008 extended the evaluation languages to Mandarin. In this year, the basic NTTS system was updated with a new Mandarin phrase prediction module. According to the listening evaluation results, all three aspects of English voice, similarity, MOS and word error rate, have been improved slightly. However, the performance of Mandarin voice was not as good as we expected. Some result analyses on Mandarin voice are presented in this paper.

Index Terms: speech synthesis, unit selection, Blizzard Challenge

1. Introduction

Blizzard Challenge [1] has been held several times to evaluate different Text-To-Speech systems based on common databases since 2005. Many research institutes participated in this system evaluation. In the last several years some systems [3,4,5,6] have achieved quite high in naturalness and intelligence, according to statistical analysis of the listening test results[2].

The quality of the whole text-to-speech system depends on many different aspects. As a whole system evaluation, the results reflect many years' accumulation of a certain research institute. It is more challenging for a developing system. Despite this disadvantage, the new developing system can benefit from idea exchanges; understanding more about the current technology trend. The Blizzard Challenge also provides a valuable opportunity to carry out extensive listening tests with the benchmark from other participants' systems.

In 2007, the Nokia Research Center Beijing participated in the Blizzard Challenge for the first time with a developing system called NTTS. Based on the listening test results, we can identify the weakness of our NTTS system. At the same time, NTTS is formally tested because there are usually not enough native English speakers available.

The organizer of Blizzard Challenge 2008 decided to extend the evaluation languages to Mandarin Chinese. Testing two languages on the same system framework allows for checking and verification of the system's multilingual capability.

In order to achieve better naturalness, the basic NTTS was updated with a new Mandarin phrase prediction module.

This paper is organized as follows: Section 2 gives the overview of the NTTS system where main modules are described. Section 3 shows changes made in this year. Section 4 presents the voice building procedure for the Blizzard Challenge 2008, including English and Mandarin. In section 5, listening test results are provided with additional attention paid to the Mandarin results' analysis. Finally section 6 is the summary.

2. Overview of NTTS system

As described in [8], Nokia TTS system (NTTS) is a wave concatenation unit selection system. Currently NTTS is still under development. It consists of three main modules including text processing, unit selection and waveform generation. Currently there is no explicit prosody module in our system. The whole text to speech procedure is shown in figure 1, taking "hello" as an example.

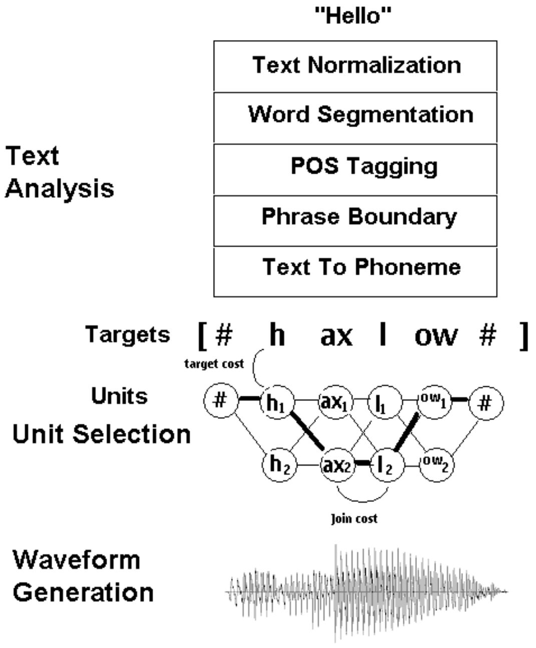

Figure 1: NTTS speech synthesis procedure

Three main modules will be described in following:

2.1. Text processing

The text processing module is composed of text normalization, word segmentation (if applicable.), POS (Part-Of-Speech) Tagging, phrase boundary prediction, and TTP (Text-to-phoneme) modules.

The text normalization module converts the inputted sentence into specific standard form. Encoding conversion, abbreviation expansion and digit-to-text string transformation are carried out in this module. The dictionary of abbreviation needs to be maintained from domain specifically. In order to cover as many phenomena as possible, the digit processing function is dynamically revised from time to time.

Word segmentation is necessary for languages whose writing system doesn't mark word boundaries. Chinese is a typical language for that. But there is no general standard to clearly define the word boundaries. Even in the definition of “word” itself, no word set exists that everyone would find acceptable. In practice, both lexical words (grammar word) and prosodic word are widely used. Lexical words are targeted in grammar analysis, e.g., POS tagging. The prosodic words are targeted in prosody analysis and rhythm study. Different word segmentation methods have been explored. But no perfect one exists for all applications. After the word boundary is given, character to pinyin conversion can be for the most part clarified.

2.2. Unit selection

During the unit selection phase, a non-uniform selection method is implemented through a search strategy. The whole search procedure consists of searching in different layers. Three layers, including syllable, word and phrase, are used. All units of a certain layer are taken as trunks. In order to maximize the integrity of the fundamental unit, the decoding is done from the bottom to up. The unit selection procedure is shown in figure2.

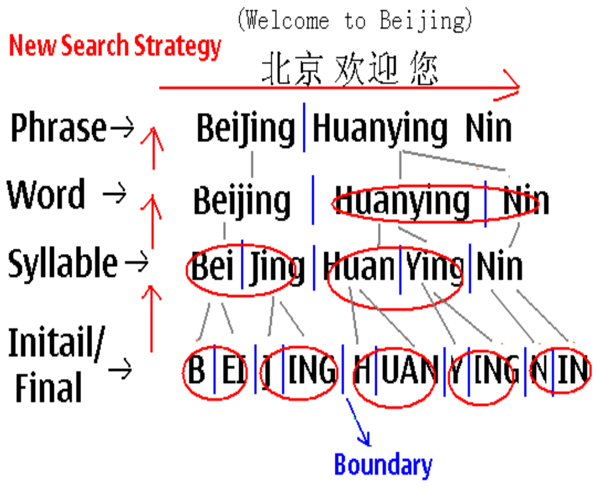

Figure 2: Unit selection procedure

In order to maximize the integrity of the fundamental unit, the new search strategy decodes the chunks layer after layer, from bottom to top. Given the output from text analysis, all units in different layers can be seen as chunks. From bottom up, they are phonemes, syllables, words and phrases. We take a phoneme as the smallest unit in this paper.

During the search procedure, in the first round, all units at the syllable layer can be decoded separately using Viterbi. These instances in the voice database should have a low join cost. If these syllables exist in the voice database, the value of join cost is zero. If the number of instances is high, pruning will take place, and only those instances which have the most similar or exactly the same context will survive. If the number of instances is less than the N-best requirement, some phoneme sequences will be generated using phonemes from different syllables. This scheme provides solutions for unseen words, or just a new context environment which was not seen in source speech.

After the first round, all top N-bests form the candidate list for the corresponding syllables. The second round for the word layer can be searched out as a syllable round. Candidates for each chunk are from the result of the previous round. Again, the output of each chunk forms a candidate for the next layer phrase.

From the candidates of words, it is easy to search at the phrase level. The results can be seen as the result for the whole sentence.

Using the above procedure, for a target sequence, if an instance of this sequence exists in the source database, this instance will be selected. Basically the maximum length of the target sequence will be selected, as the non-uniform unit selection usually does.

During the multilayer search, it is possible to skip some middle layers, e.g., the syllable layer. The whole layer framework depends on the concrete language, or what kind of granularity is wanted.

The new search strategy focuses more on the integrity or quality of the fundamental unit. Inside the chunk, e.g., a word, the accuracy of phoneme boundary annotation is not as sensitive as in the usual search method. The corresponding unit sequence will be selected. Of course, the boundary accuracy does effect wav concatenation. It is possible some part of the wav, no matter if it is in time domain, frequency domain or another parameter domain, will be missed. But it will not affect the selection procedure. In the case where a long unit sequence is selected, the boundary errors will not accumulate. Only the boundary error of the beginning and ending unit will be revealed. Furthermore, in the units which the beginning and ending unit abut may be silence or pause, their relative boundary tolerance is wider than other units’. Concatenations costs at points of low power or pauses are relatively low.

During selection procedure, prosody information is not specifically considered. It is assumed that prosody is contained in the context implicitly. The prosody structure information, such as prosody phrase boundary, can be considered as a different layer for decoding.

2.3. Waveform generation

The original speech corpus is analyzed and converted into other parameter domains, e.g., lsp or mgc, etc. Depending on the application case, these parameterized waveforms can be encoded into low bit stream to save valuable storage space. Data compression is very important for embedded devices. NTTS has such a module, but the detail is not covered here because the footprint is not the crucial topic for Blizzard Challenge.

After unit selection, these selected candidates need to be reconstructed into time domain waveforms. In order to smooth the boundary of joint point, sometimes signal modification technologies are introduced. In NTTS, no signal modification function is used.

3. Changes to system 2007

Basically NTTS 2008 is very similar to NTTS 2007. Compared with NTTS 2007, changes are made to the text processing module to support Mandarin phrase prediction. The unit selection module is adjusted accordingly.

- Adding Mandarin phrase boundary prediction module to NTTS 2007. In NTTS 2007, there is no phrase prediction module for Mandarin. The unit selection for Mandarin mainly depends on word boundary, character position in word, position in sentences, and phonetic context and acoustic features.

- Emphasizing phrase boundaries. This year Mandarin is covered in Blizzard Challenge evaluation, and we want to make the prosodic phrase clear. The weight for phrase boundary is set much higher than the weight for other features.

Above changes should have no impact for English voice database. For our own Mandarin voice database, the training text corpus and test text corpus are manually annotated with phrase boundary information. Quality improvement can be observed after adding Mandarin phrase boundary support.

4. Building voices for the Blizzard Challenge 2008

4.1. English voice

Speech segmentation and voice build are done offline.

4.1.1. Speech corpus

The English speech corpus for the blizzard 2008 is composed of 9509 utterances. The speaker is a male from UK. Total time length of English corpus is around 15 hours. The wav files are in 16k sampling rate. Beside speech data, the orthographic transcription of the whole database is available.

The organizer kindly released the unilex1.3 for interested participants. All festival Utterance structures for whole database are also produced from this lexicon and made available. Due to our lack of familiarity of the UK accent, we fully depend on these festival utterance structure files. No manually check of these files has been carried out.

This database is more prosodically varied and slightly less 'newsreader' style[11].

4.1.2. Speech segmentation

From the utterance structure files the phoneme sequence for each utterance was extracted. HTK toolkit [10] was used to align the phoneme sequences with waveform files. As a typical HTK force alignment procedure, the whole labeling tool chain was trained from flat start, moving from mono-phone to tri-phone, then to question-based clustering. Finally speaker dependent tri-phone HMMs are trained.

No further manual segmentation corrections were undertaken to evaluate the quality of the automatic segmentation.

4.1.3. Voice database

After speech segmentation, no manual correction was carried out to refine the phoneme boundaries. Once the segmentation information is ready, the voice creation is straightforward. Pitch mark detection is another important step. Several pitch extraction tools are used to cross check the pitch information to protect from strange values.

All phoneme boundary information and context information are collected to build voice database. The building procedure is done automatically. Only voice A was built. The runtime voice database includes parameterized speech segment database, unit inventory with phonetic context, and acoustic features at phoneme boundaries.

High-level prosodic features such as phrase breaks are based on a syntactic analysis of the input text.

4.1.4. Speech synthesis

The test set in Blizzard Challenge 2007 was also used as test sentences. New English test set includes 520 sentences from five different categories. Totally, around 900 sentences are generated.

These English test sentences are not synthesized from Text. Again the festival Utterance structures files are used as input source to unit selection modules. Many features are extracted from utterance structure files for target cost calculation. These features include phonetic context, stress, phoneme position in hierarchy of utterance, phrase, word and syllable, etc.

Unit selection procedure follows the description in section 2.2. Using target cost and join cost as measurement, an optimal unit sequence can be selected and streamed into the waveform generation module.

4.2. Mandarin

4.2.1. Speech corpus

The Mandarin speech corpus for the blizzard 2008 is composed of 4500 utterances spoken by a female. The total time length of English corpus is around 6.5 hours. The wav files are in 16k sampling rate. The content of each wav file was given in the Chinese character string. Beside the transcription file, a labeling file is available for each sentence. One layer in the labeling file is a sequence of PINYIN with tone. Another layer in the labeling file includes the Initial/Final sequence.

The Mandarin text corpus was gone through quickly. It seems many sentences are just part of usual full sentences. After reading them, it is hard to grasp the meaning of sentences. The style of the text corpus is quite different from our own system.

4.2.2. Speech segmentation

HTK toolkit was used to segment the speech corpus, as we did for English. The basic unit for Mandarin could be syllable, Intial/Final or phoneme. Here Initial and Final with tone are used as basic units for the voice of Blizzard Challenge 2008. No tone or neutral tone is seen as tone number 5. For an example, PINYIN “Bei3 Jing1” can be converted into the pronunciation sequence “b ei3 j ing1”. “b”, “ei3”, “j” and “ing1” are from basic units. There are a total of 224 units, including pause.

4.2.3. Voice database

The transcription of Mandarin speech corpus is analyzed by the text processing module of NTTS. We used the simple maximum match method to segment the Chinese sentences into words. HMM models are used to model POS. The tagger is trained from the corpus of People’s Daily. Word boundary and POS attribute are used as features for the phrase prediction module.

When converting text to pinyin sequence, tone Sandhi has to be considered to match real pronunciation.

The structure of the Mandarin voice database for Blizzard Challenge is the same as the one for English.

4.2.4. Speech synthesis

The test set for Mandarin includes 647 news sentences and 50 semantically unpredictable sentences. The similar label files as training speech corpus are provided. Again we need to generate the phone sequence using front-end module.

5. Results of the Blizzard Challenge 2008

We only worked on the full set database for English. Two voices were submitted to Blizzard 2008, one for English, another for Mandarin.

5.1. English results

The English listening evaluation was organized almost the same as in last year. The detail design can be found in [2]. Totally there are five sections. However, more listener types have been introduced. There are eight listener types defined in 2008. Several listener types only have few participants.

Three aspects of listening results are analyzed: similarity to original speaker, mean opinion score, and word error rates for semantically unpredictable sentences (SUS). The results in 2007 are included to show the trend.

For each aspect, interested breakdown data is also presented. Three listener types in 2008 have their counterpart in 2007.

Table 1. *Listener types*

	Listener type identifier	
	2007	2008
Paid UK students	K	EUL
Volunteers	R	ER
Speech experts	S	ES

Another two systems, festival from CSTR[7] (participant letter B in 2007 and 2008) and HTS (participant letter N in 2007 and C in 2008), are used as reference systems. HTS-2007[9] used speaker independent approach with speaker adaptation.

Only data of voice A is included.

5.1.1. *Similarity test*

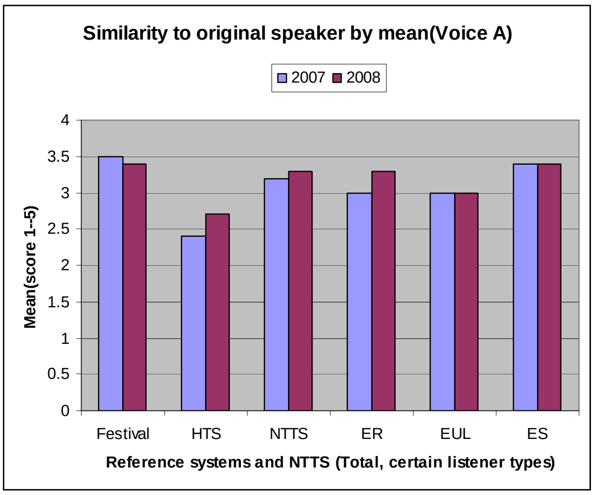

Figure 3: Similarity to original speaker by mean (English voice)

From Figure 3 we can see that the similarity of the original speaker in 2008 is close to the one in 2007. The difference is not statistically significant. The volunteer listeners gave different responses for 2007 and 2008 voices. The mean score is around 3.3.

5.1.2. *Mean Opinion Scores*

The MOS scores of NTTS system are 3.0 in 2007 and 3.2 in 2008. It seems a slight improvement can be observed as figure 4.

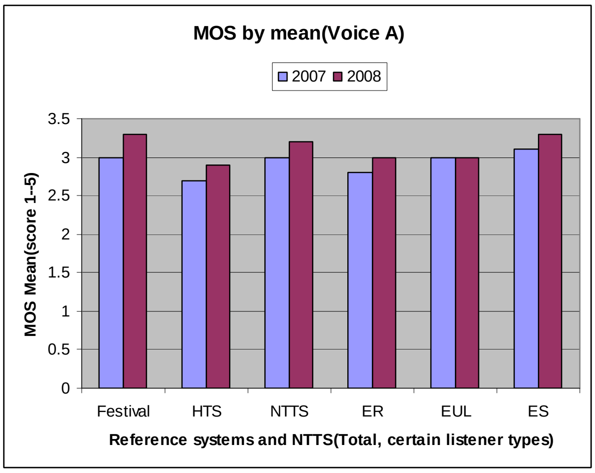

Figure 4: MOS score in 2007 and 2008 (English voice)

5.1.3. *Word error rates for SUS test*

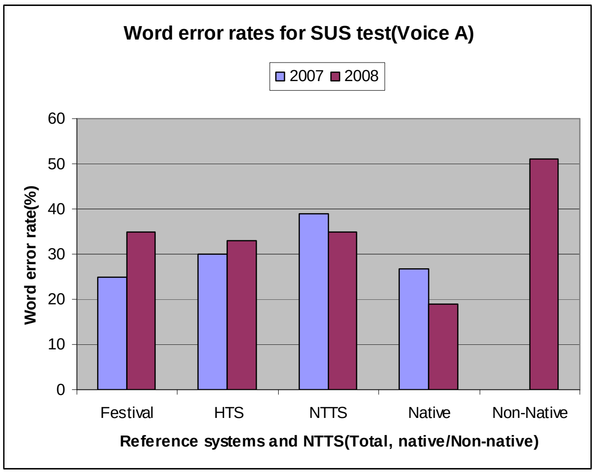

Figure 5: Word error rates for SUS test (English Voice)

The SUS test is the most challenging one among five evaluation sections, especially for non-native listeners. Many listeners didn’t complete this section.

Native word error rate for voice A in 2007 couldn’t be found. So for the last category in figure 5, there is only one column available.

In general, NTTS 2008 achieved better WER than 2007. From figure 5, it can be seen that our system has a worse word error rate than festival and HTS. The word error rate from native speakers is about 18%. However, the word error rate from non-native listeners is extremely high, above 50%. The results from native listeners and non-native listeners show a huge difference.

5.1.4. *Discussion on English voice*

In general our Blizzard voice 2008 performed a little better than our voice 2007 on all three evaluation aspects. For SUS test, it is noticeable that different listener type biased the results a lot.

The performance improvement of voice 2008 may come from following points:

- Better database preparation
- Using festival Utterance structure information from organizer. Our own generated information is inferior to those provided data.
- The size of speech corpus for Blizzard Challenge 2008 is bigger than the one of 2007.

5.2. Mandarin Results

This is the first time to include Mandarin into Blizzard Challenge. The listening test design follows the same principle as English.

There are four listener types: MC - paid participants in China (native speakers of Mandarin), ME - paid participants in Edinburgh (native speakers of Mandarin), MR – volunteers, MS - speech experts.

The listening test results for similarity, MOS and word error rates will be presented in following parts. The HTS (system “C”) is taken as reference.

5.2.1. *Similarity test*

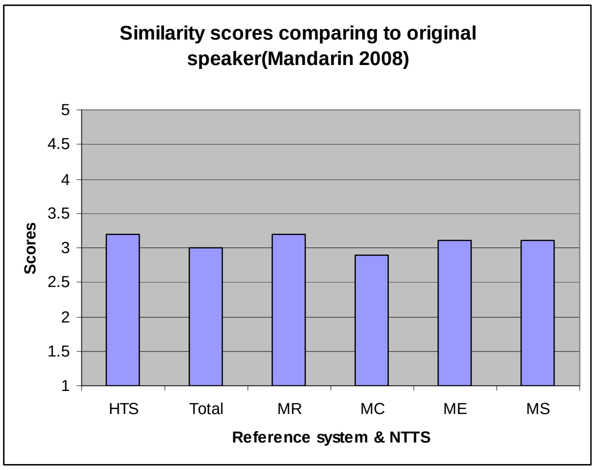

Figure 6: Similarity to original speaker by mean (Mandarin voice)

From figure 6, HTS system show higher similarity than NTTS. This is different from English case. The similarity score of NTTS is around 3.0.

5.2.2. *Mean Opinion Scores*

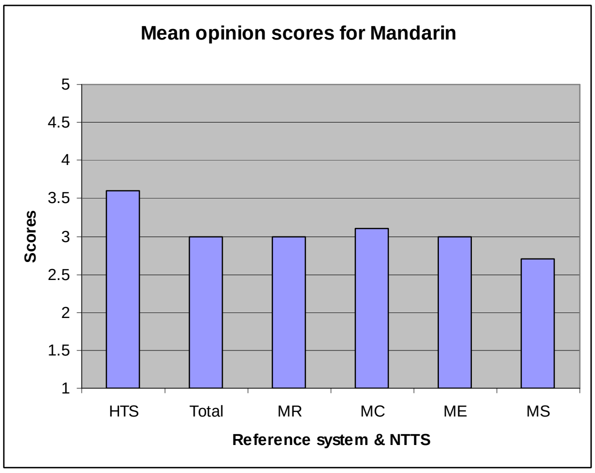

Figure 7: MOS score and 2008 (English voice)

From figure 7, the HTS system performed a little higher than NTTS. The MOS score of NTTS is around 3.0

5.2.3. *Word error rates for SUS test*

The evaluation of Mandarin voice is more complicated than English. First, any traditional Chinese characters are converted to simplified Chinese characters.

Three sub-errors are defined:

- CER: Character Error Rate (CER) is calculated using a similar procedure to WER, treating each character as a word. No spelling correction was used.
- PTER: Pinyin plus Tone Error Rate, which is choosing the pinyin plus tone path through the lattice that gives the lowest. All simplified Chinese characters are converted into pinyin plus tone.
- PER, Pinyin Error Rate, strip the tones leaving only pinyin, choosing the pinyin path through the lattice that gives the lowest PER

Above points are basic definitions. The procedure for calculation of error rates will be presented in detail at the workshop by the organizer.

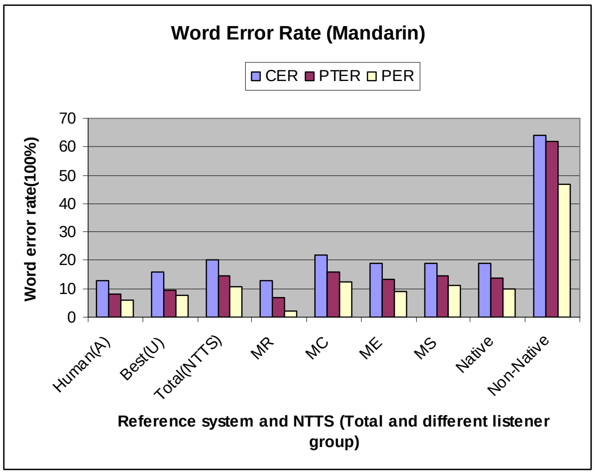

Figure 8: Word error rates for SUS test (Mandarin Voice)

System U achieved the best performance in word error rate test. Beside NTTS, the best system is also presented in figure 8. The result is interesting. The human voice received a character error rate 13%. Without considering tone, the PER of human voice is about 5.8%. For NTTS, the CER, PTER and PER are 20%, 14.7%, and 10.7% separately.

5.2.4. *Discussion on Mandarin voice*

This year we added Mandarin phrase prediction module to NTTS. The informal listening evaluation showed this new module help on the naturalness of samples generated. After generating the blizzard test set, we checked some samples. We found that several samples had prosody problem. The problem was traced back to phrase prediction with the blizzard Mandarin text corpus. The new added phrase prediction module has low accuracy on the text corpus.

The possible reason is the style of blizzard test corpus is different from our training corpus. Our own database is annotated with phrase boundary. In principle the workload to annotate the text corpus manually is affordable. However, for Blizzard Challenge database, we have no time to do that.

6. Conclusions

We have described the updated NTTS with which we participated in the Blizzard Challenge 2008 for both English and Mandarin. To build English voice database and generate the English test set, we directly use these festival utterance structure files provided by organizer. Comparing the English listening test results with the one in last year, slight performance improvement can be observed on all three evaluation aspect: similarity to original speaker, mean opinion score, and word error rate for semantically unpredictable sentences.

The updated NTTS has a new phrase prediction module for Mandarin. However, due to the inconsistence between module training text corpus and blizzard mandarin text corpus, this new phrase prediction module didn’t perform well on blizzard text corpus. It brought negative impacts to the Mandarin blizzard voice database.

In the future the listening test result will be further analyzed. More attention will be paid to front-end text analysis, and voice database preparation.

7. References

[1] A. Black and K. Tokuda, The Blizzard Challenge 2005: Evaluation corpus-based speech synthesis on common database, in Proceedings of Interspeech 2005, Lisbon, Portugal, 2005.

[2] R.A.J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, Statistical análisis of the Blizzard Challenge 2007 listening test results. Proc. Blizzard Workshop (in proc. SSW6), August 2007. Bonn, Germany.

[3] M. Kaszczuk and L. Osowski, The IVO Software Blizzard 2007 Entry: improving Ivona Speech Synthesis System. Proc. Blizzard Workshop (in proc. SSW6), August 2007. Bonn, Germany.

[4] Zhen-Hua ling, Long Qin, etc, The USTC and iFlyTek speech synthesis systems for Blizzard 2007. Proc. Blizzard Workshop (in proc. SSW6), August 2007. Bonn, Germany.

[5] Won-Suk Jun, Deok-Su Na, etc, The VoiceText Text-To-Speech system for the Blizzard challenge 2007. Proc. Blizzard Workshop (in proc. SSW6), August 2007. Bonn, Germany.

[6] Johan Wouters, SVOX participation in Blizzard 2007. Proc. Blizzard Workshop (in proc. SSW6), August 2007. Bonn, Germany.

[7] K. Richmond, V. Strom, R. Clark, J. Yamagishi and S. Fitt, Festival Mulstisyn voices for the 2007 Blizzard Challenge. Proc. Blizzard Workshop (in proc. SSW6), August 2007. Bonn, Germany.

[8] Feng Ding, Jari Alhonen, Non-uniform unit selection through search strategy for Blizzard Challenge 2007. Proc. Blizzard Workshop (in proc. SSW6), August 2007. Bonn, Germany.

[9] J. Yamagishi, Heiga Zen, T. Toda, K. Tokuda, Speaker-independent HMM-based speech synthesis system –HTS-2007 system for the Blizzard Challenge 2007. Proc. Blizzard Workshop (in proc. SSW6), August 2007. Bonn, Germany.

[10] HTK toolkit, http://htk.eng.cam.ac.uk/

[11] Volker Strom, Ani Nenkova, Robert Clark, Yolanda Vazquez-Alvarez, Jason Brenier, Simon King, and Dan Jurafsky. Modelling Prominence and Emphasis Improves Unit-Selection Synthesis. Interspeech 2007.