The NTUT Blizzard Challenge 2010 Entry

Yuan-Fu Liao, Ming-Long Wu and Shao-He Lyu

Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan
yfliao@ntut.edu.tw

Abstract

This paper describes our HMM-based speech synthesis system (HTS) submitted to Blizzard Challenge 2010. Three Mandarin Chinese voices were built for two hub (MH1and MH2) and one spoke (MS1) tasks this year (the voice for MS2 is the same as MH1's one). According to the evaluation results, our system got in average 2 points for both mean opinion scores (MOS) and similarity tests for MH1, MH2 and MS1. Beside, for MH1, about 22% and 24% pinyin error rates (without (PER) and with tone (PTER), respectively) and 28% character error rate (CER) were achieved for intelligibility test. However, for speech in noise task, MH2, the performance of our system is not satisfied, especially in low signal-to-noise (SNR) case. In conclusion, these results indicate there is still a lot of room for improvement, especially for dealing with different speaking style (comparing with last year's data) and noise interference.

Index Terms: speech synthesis, HMM, HTS

1 Introduction

This paper describes our HMM-based speech synthesis system (HTS) submitted to Blizzard Challenge 2010 [1], the open evaluation that compares the performance of different TTS systems with a common speech database.

In this year, NTUT Speech Processing Laboratory [2] has built three Mandarin Chinese voices for two hub (MH1and MH2) and one spoke (MS1) tasks. However, no noise interference was considered, so the voice for MS2 task is the same as MH1's one.

Basically, our system faithfully follows the framework of HTS [3-4] and the procedure of speaker dependent training procedure with minor modifications (mainly for solving the pop noise problem due to global variance (GV) [5] function. Beside, HTS version 2.1.1 was adopted to take the advantage of the new feature, i.e., context-dependent global variance (CDGV) [6].

Our main goal for participating Blizzard Challenge is to establish a reasonable TTS baseline for developing our own discriminative training algorithm (against noise) and prosody model in the future.

The organization of this paper is as follows. First, we describe our HTS-based Mandarin Chinese TTS system, especially with focus on the different points comparing with our previous system. Section 3 describes the tasks and evaluation settings. In Section 4, the evaluation results are presented. Finally some conclusions are drawn.

2 HTS-based Mandarin Voice Building

In this section, the voice building approach, especially the introduction of HTS version 2.1.1, is described in detail.

2.1 Synthesis Units and Models

59 Sub-syllable units were chosen as the basic synthesis units including e 21 initials and 38 finals. Besides, a short pause and a silence model were used.

Table 1 shows the hierarchical phonetic structure of Mandarin Chinese syllable. Here a final is made up of a nucleus vowel and an optional media and nasal.

For all synthesis units in our system, HMMs with 5 states, left-to-right transition and diagonal covariance matrix are adopted

Tone			
Initial	Final		
(Consonant)	(Media)	Nucleus vowel	(Nasal)

Table 1: The hierarchical phonetic structure of Mandarin Chinese syllable (consonant, media and nasal are optional).

2.2 Question Sets

Taking the advantage of new CDGV feature, two different question sets, for sub-syllable and CDGV models clustering, respectively, were used in the decision tree-based algorithm.

First, the question for clustering the sub-syllable models is composed of 6 layers and listed in Table 1. On the other hand, for clustering CDGV models, 3 utterance-level cues were explored as showed in in Table 2.

Moreover, prosody structure information is also utilized in the first question set, including (1) prosodic word, (2) intonation phrase and (3) utterance boundaries.

Layer	Question
Sub-syllable	the name and type of current and surrounding sub-syllables
Syllable	the tone type of current and surrounding syllables; the number and forward and backward position of syllables in a word
Word	the part-of-speech (POS) of current and surrounding words; the number and forward and backward position of words in a phrase
Phrase	the number and forward and backward position of phrases in an clause
Clause	the number and forward and backward position of clauses in an utterance
Utterance	the number of syllables, words, phrases and clauses in an utterance

Table 2: Hierarchical structure of question set for decision tree-based sub-syllable model clustering.

Layer	Question
Syllable	the number of syllables in a utterance
Word	the number of words in a utterance
Phrase	the number of phrases in a utterance

Table 3: Hierarchical structure of question set for decision tree-based CDGV model clustering.

2.3 Speech Parameters

24-order mel-generalized cepstrum (MGC) [7] and fundamental frequency, F0 [8], were extracted as the spectral and excitation parameters. Beside, their first and second order derivative features were also generated to form a 75-dimentional feature vector for each speech frame (with 5ms frame shift).

2.4 Training Procedures

The voice building steps are showed in Fig. 3. Comparing with our 2009 system, those blue blocks were added after adopting HTS version 2.1.1 for CDGV estimation and clustering.

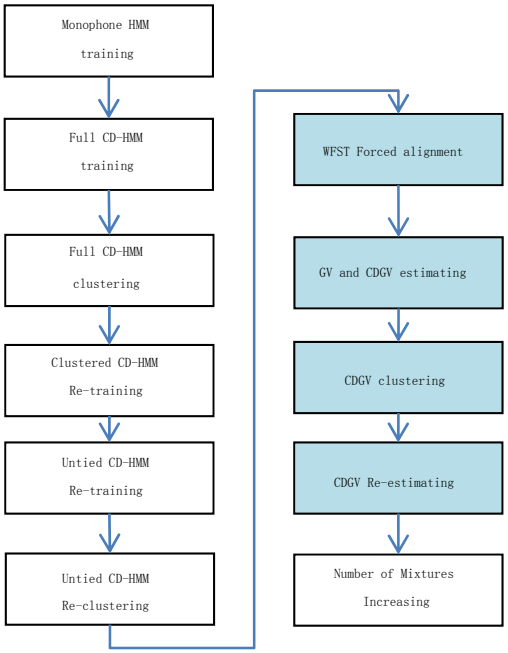


Figure 1: The block diagram of the voice building procedure using HTS version 2.1.1.

2.5 Speech Synthesis

For speech parameter generation, both GV and CDGV were considered. Besides, Mel Log Spectrum Approximation (MLSA) [9] filter and post-filtering were applied for speech synthesis.

Finally, to alleviate the pop noise problem due to the nonlinear characteristics of GV and CDGV, the amplitude of the generated voice was normalized before be converted into 16-bit PCM format.

3 Evaluation Settings

3.1 Mandarin Chinese database

A female Mandarin Chinese database was released by Chinese Academy of Sciences. There are 6000 utterances (about 9 hours) in this database.

For prosodic boundary information, only the first 1000 utterances of this database are manually labelled. The remaining sentences are all automatically labelled. On the other hand, the initial/final segmentation and POS of all utterances are all automatically labelled

3.2 Tasks

There are four tasks for Mandarin including two hub (MH1, MH2) and two spoke (MS1 and MS2) ones. Among them, MS2 is a new and interesting task:

- MH1: build a voice from the full Mandarin database (5884 utterances, about 9 hours)
- MH2: build a voice from a subset of the full Mandarin database (utterance 5085~5884, about one hour).
- MS1: build a voice from only 100 utterances of the full Mandarin database (utterance 5085~5184).
- MS2: build a voice from the full Mandarin database which is suitable for synthesizing speech in noisy environment.

3.3 Subjects

The evaluation was conducted online. Hundreds of subjects took the evaluation test. The types of listeners could be divided into four groups including:

- MC - paid participants in China (native speakers of Mandarin)
- ME - paid participants in Edinburgh (native speakers of Mandarin)
- MR - online volunteers
- MS - speech experts

3.4 Tests

All systems were evaluated with respect to naturalness, similarity and intelligibility. Among them, naturalness and similarity were measured using the news subsets of test sentence, but the intelligibility tests were calculate on the SUS subset.

- Naturalness: in each session listeners listened to one sample and chose a score which represented how natural or unnatural the sentence sounded on a scale of 1 (completely unnatural) to 5 (completely natural).
- Similarity: in each session listeners could play 4 reference samples of the original speaker and one synthetic sample. They chose a response that represented how similar the synthetic voice sounded to the voice in the reference samples on a scale from 1 (sounds like a totally different person) to 5 (sounds like exactly the same person).
- Intelligibility: listeners heard synthetic sample utterance by utterance and typed in what they heard. Listeners were allowed to listen to each sentence only once. The procedure for calculation of error rates is as follows:
  (1) calculate character error rate (CER) using a similar procedure as the conventional word error rate (WER), treating each character as a word.
  (2) convert each character to pinyin+tone and calculate pinyin+tone error rate (PTER), choosing the pinyin+tone path through the lattice that gives the lowest PTER
  (3) strip the tones leaving only pinyin, and calculate pinyin error rate (PER), choosing the pinyin path through the lattice that gives the lowest PER

4 Evaluation Results

The evaluation results are reported with boxplots of MOS and similarity scores and barplots of CER, PTER and PER of all systems. In all boxplot figures, the central solid bar represents the median, the shaded box the quartiles, extended lines the 1.5 times quartile range, and the outliers are displayed as circles.

There are in total (excluding the samples of the original voice (natural speech), "A") 11 systems for MH1, 10 systems for MH2, 5 systems for MS1 and 8 systems for MS2 tasks. It must be stressed that System C is a standard speaker-dependent HMM-based voice built using a similar method to the HTS entry to Blizzard 2005 [10-11]. Systems F to V are the participants. The final results are commented in the following lines comparing our performance (System "I") with that of the other participants.

4.1 MH1 Task

4.1.1 MH1 Mean Opinion Score Test

MOS comparison between our (system I) and all other systems is shown in Fig. 2 Our system got in average only 2 points for all listeners which is worse than our 2009 system [12].

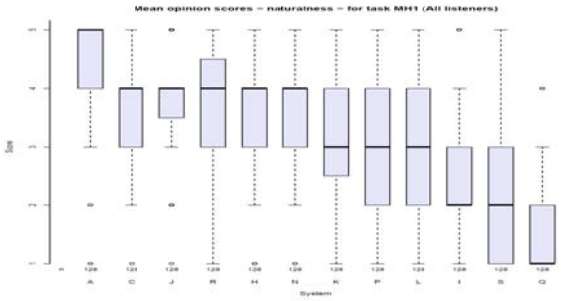


Figure 2: MOS comparison between our (I) and all other MH1 systems.

4.1.2 MH1 Similarity Test

The boxplots of similarity scores of all systems are shown in Fig. 3. From the figure, we can conclude that our system performs worse than the average of all systems in the similarity test. This is still the major weakness of our system (comparing with our last year's results [12]).

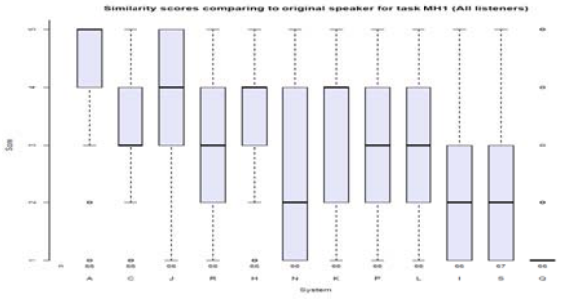
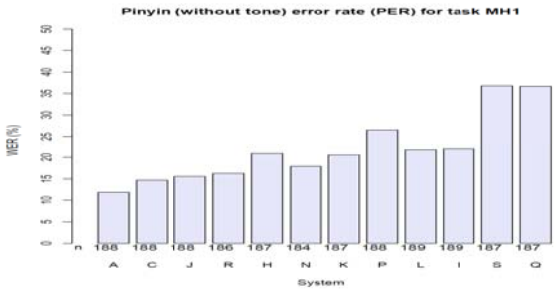


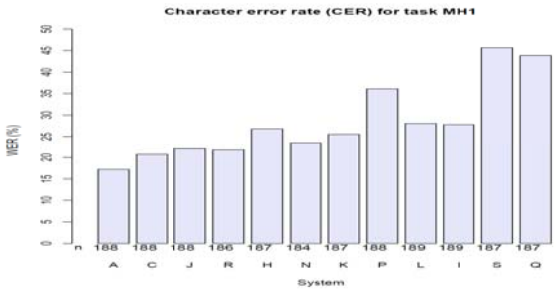
Figure 3: Similarity comparison between our (I) and all other MH1 systems.

4.1.3 MH1 Word Error Rate Test

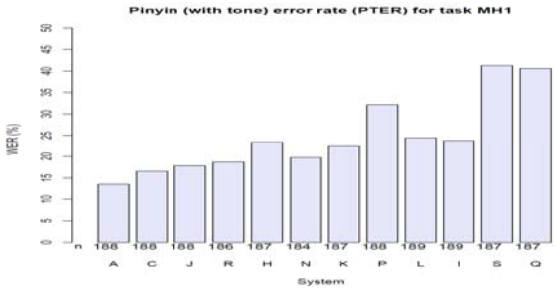
Fig. 4 shows the (a) PER, (b) PTER and (c) CER achieved by all the MH1 participants for intelligibility test. According to the test results, our MH1 voice achieved in average 22% PER, 24% PTER and 28% CER. These results are also worse than the performance of our 2009 system [12].



(a)



(b)



(c)

Figure 4: Intelligibility comparison between our (I) and all other MH1 systems for (a) PER, (b) PTER and (c) CER.

4.2 MH2 Task

4.2.1 MH2 Mean Opinion Score Test

MOS comparison between our (system I) and all other systems is shown in Fig. 5. Our system got in average 2 points for all listeners which is slightly worse than in the MH1 task.

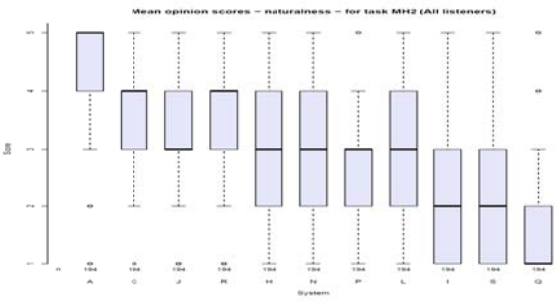


Figure 5: MOS comparison between our (I) and all other MH2 systems.

4.2.2 MH2 Similarity Test

The boxplots of similarity scores of all systems are shown in Fig. 6. From the Fig. 3 and 6, it can conclude that our MH1 and MH2 voices all have the same problems.

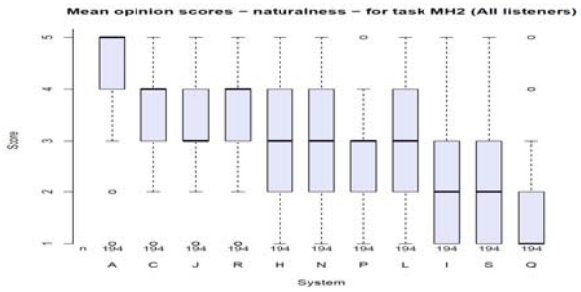
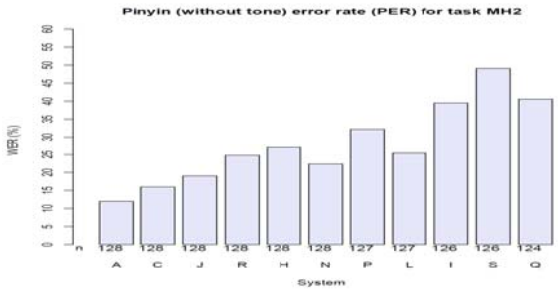


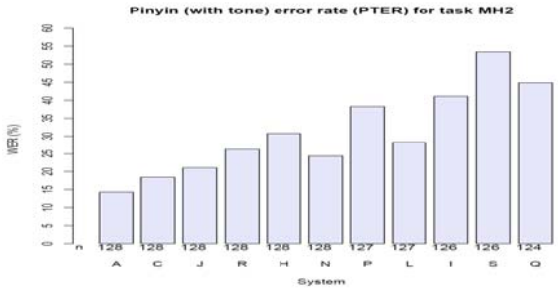
Figure 6: Similarity comparison between our (I) and all other MH2 systems.

4.2.3 MH2 Word Error Rate Test

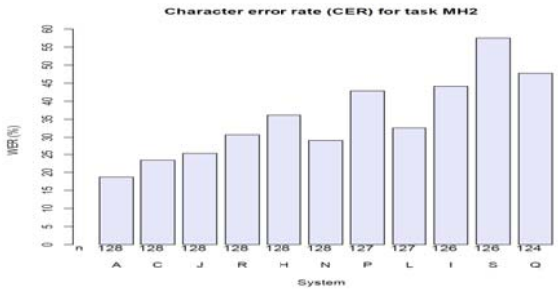
Fig. 7 shows the (a) PER, (b) PTER and (c) CER achieved by all the MH2 participants for intelligibility test. According to the test results, our MH2 voice is worse than our MH1's one. This is mainly the consequence of less (only 1/9) training material.



(a)



(b)



(c)

Figure 7: Intelligibility comparison between our (I) and all other MH2 systems for (a) PER, (b) PTER and (c) CER.

4.3 MS1 Task

4.3.1 MS1 Mean Opinion Score Test

MOS comparison between our MS1 (system I) and all other systems is shown in Fig. 8. Our system got in average 2 points from all listeners.

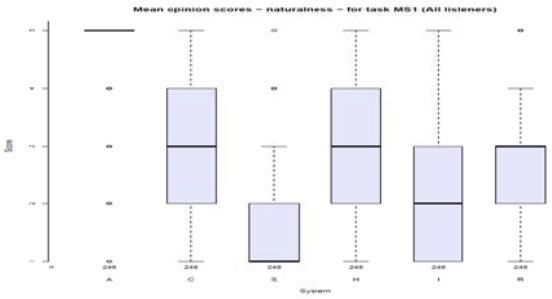


Figure 8: MOS comparison between our (I) and all other MS1 systems.

4.3.2 MS1 Similarity Test

The boxplots of similarity scores of all MS1 systems are shown in Fig. 9. Our MS1 voice still has the same problem with MH1 and MH2's ones. This is again the weakest point of our systems.

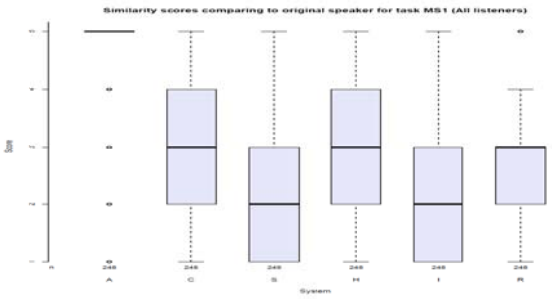
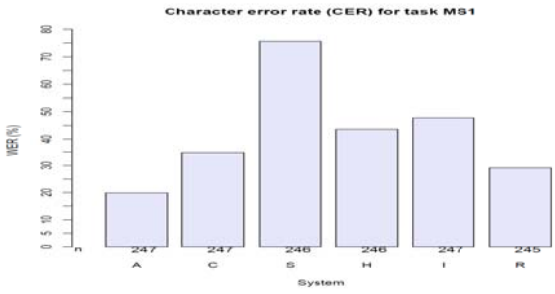


Figure 9: Similarity comparison between our (I) and all other MS1 systems.

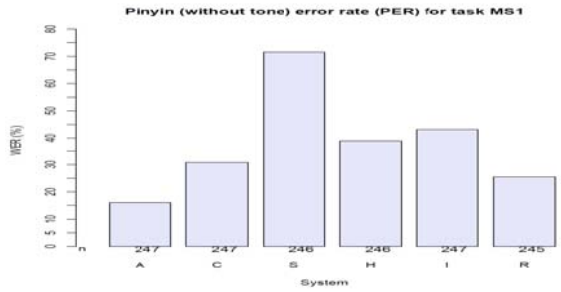
4.3.3 MS1 Word Error Rate Test

Fig. 10 shows the (a) PER, (b) PTER and (c) CER achieved by all the MS1 participants for intelligibility test.

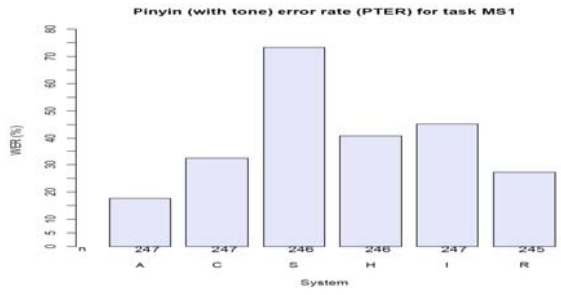
According to the test results, our MS1 voice achieved 43% PER, 45% PTER and 48% CER. Comparing with the results of MH1 and MH2, the error rates were increased dramatically.



(a)



(b)



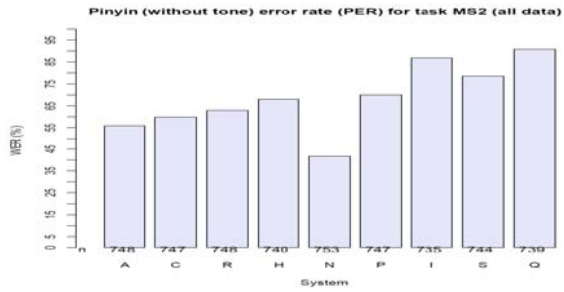
(c)

Figure 10: Intelligibility comparison between our (I) and all other MS1 systems for (a) PER, (b) PTER and (c) CER.

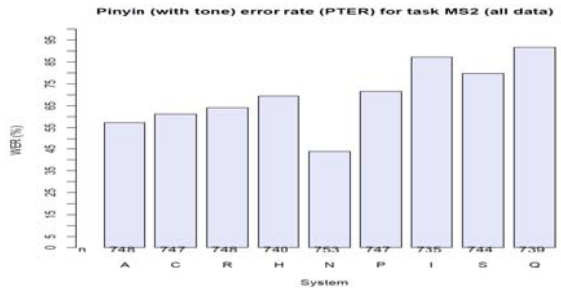
4.4 MS2 Task

Figure 11 shows the (a) PER, (b) PTER and (c) CER achieved by all the MS2 participants for intelligibility test.

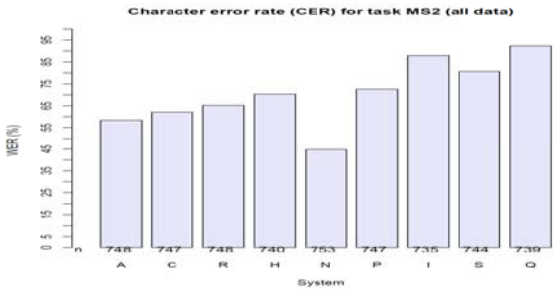
MS2 is a new task, i.e., speech in noise, and was introduced this year. Since, our voice for MS2 task is the same as our MH1's one, the evaluation results in fact shows that our MH1 voice was easily destroyed by environment noise.



(a)



(b)



(c)

Figure 11: Intelligibility comparison between our (I) and all other MS2 systems for (a) PER, (b) PTER and (c) CER.

5 Conclusions and Future Works

From the results of the listening tests, it shows that our MH1, MH2 and MS1 voices all only got in average 2 points for both MOS and similarity tests. And 22% PER, 24% PTER and 28% CER were achieved by MH1 voice for intelligibility test.

Comparing with the results of our 2009 systems [12], it shows that our new system is in fact worse than last year's one. Since we almost didn't change our approach, the difference may come from the different characteristics between 2009's and 2010's corpora. Especially, this year's training data is more dynamic and not always table.

Moreover, a new task, i.e., MS2, speech in noise, was introduced this year. Although, our voice for MS2 task is the same as our MH1's one, the evaluation results shows that our MH1 voice was easily destroyed by environment noise.

Therefore, there is still a lot of room for improvement. In the future, we will try to enhance our system and develop a discriminative training-based algorithm (on going) to increase the robustness of our system to different noises and speaking styles.

Acknowledgements

This work was partially supported by the National Science Council, Taiwan, under the projects with contract 97-2628-E-027-003-MY3 and 98-2221-E-027-081-MY3.

References

[1]. Blizzard Challenge, http://www.synsig.org/index.php/Blizzard_Challenge

[2]. Speech Signal Processing Lab., National Taipei University of Technology, http://www.cc.ntut.edu.tw/~enlab07/

[3]. HMM-based Speech Synthesis System, http://hts.sp.nitech.ac.jp/

[4]. Heiga Zen, Keiichi Tokuda, Tadashi Kitamura, An introduction of trajectory model into HMM-based speech synthesis, Proc. of 5th ISCA Speech Synthesis Workshop, Pittsburgh, June 2004.

[5]. Tomoki Toda and Keiichi Tokuda, Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis, InterSpeech'2005.

[6]. Recent development of the HMM-based speech synthesis system (HTS), APSIPA2009

[7]. K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. Mel-generalized cepstral analysis, a unified approach to speech spectral estimation. In Proc. of ICASSP, pages 1043.1046, 1994.

[8]. Satoshi IMAI, Cepstral analysis synthesis on the mel frequency scale, Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83, 1983.

[9]. D Talkin, A Robust Algorithm for Pitch Tracking (RAPT), Chapter 15, Speech Coding and Synthesis, Elsevier, 1995.

[10]. Heiga Zen, Tomoki Toda, An Overview of Nitech HMM-Based Speech Synthesis System for Blizzard Challenge 2005, Blizzard Challenge 2005, http://www.festvox.org/blizzard/bc2005/IS052192.PDF

[11]. Yamagishi, Junichi / Zen, Heiga / Toda, Tomoki / Tokuda, Keiichi: "Speaker-independent HMM-based speech synthesis system - HTS-2007 system for the Blizzard Challenge 2007", Blizzard Challenge 2007, http://www.festvox.org/blizzard/bc2007/blizzard_2007/full_papers/blz3_008.pdf

[12]. Yuan-Fu Liao and Ming-Long Wu, “The NTUT Blizzard Challenge 2009 Entry”, Blizzard Challenge 2009 workshop