The NTUT Blizzard Challenge 2009 Entry

Yuan-Fu Liao and Ming-Long Wu

Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan

yfliao@ntut.edu.tw

Abstract

This paper describes the process of building HMM-based speech synthesis system (HTS) voices for our participation in the Blizzard Challenge 2009. Out of the two languages required (English and Mandarin Chinese) we only built three Mandarin Chinese voices for main hub (MH) and two spoke (MS1 and MS2) tasks. According to the evaluation results, our MH voice got 3 points for both mean opinion scores (MOS) and similarity tests. Beside, 12.2% and 17% pinyin error rates (without (PER) and with tone (PTER), respectively) and 23% character error rate (CER) were achieved for intelligibility test. Moreover, our MS2 voice achieved 4 and 3 points for MOS and similarity test, respectively. In conclusion, we now have reasonable text-to-speech (TTS) baselines (at least for Mandarin Chinese) for developing our own advanced prosody model in the future.

Index Terms: speech synthesis, HMM, HTS

1 Introduction

The Blizzard Challenge [1] is an open evaluation that compares algorithm performance of different TTS systems built with a common speech database. After two months for voice building, participants are asked to synthesize about thousands of test texts in one week that will be evaluated with respect to naturalness, similarity and intelligibility.

NTUT Speech Processing Laboratory [2] has been working in speech signal processing field, especially, for speech recognition, since 2002. But this has been our first attempt to build a TTS system and also our first participation in an international TTS evaluation campaign. Our main goal is to establish reasonable TTS baselines for developing and verifying our own advanced prosody model [3-4] in the future.

For the sake of completion, this paper is in fact more like a technical report. The organization of this paper is as follows. First, we describe our HTS-based Mandarin Chinese TTS system. In Section 3 the voice building process is explained. The evaluation results are presented and discussed in Section 4. Finally some conclusions are drawn.

2 HTS-based Mandarin Chinese TTS

Basically, our system faithfully follows the framework of HTS [5]. Here we briefly review and discuss the core technologies of HTS and some Mandarin Chinese issues.

2.1 HTS Framework Overview

Fig. 1 shows an overview of the basic HMM-based speech synthesis system. The main issues of this framework are:

(1) How to precisely extract excitation and spectral parameters from speech signal
(2) How to generate label sequences using text-analysis and prosody prediction
(3) How to reliably build as much as possible context-dependent HMMs from a speech database according to the extracted labels
(4) How to optimally generate parameters from those context-dependent HMMs according to the extracted label sequence
(5) How to synthesize speech from the generated parameters

For issue (3) decision tree-based model clustering using minimum description length (MDL) criterion is often used. For issues (1) and (5), the state-of-the-art approach is STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [6-7].

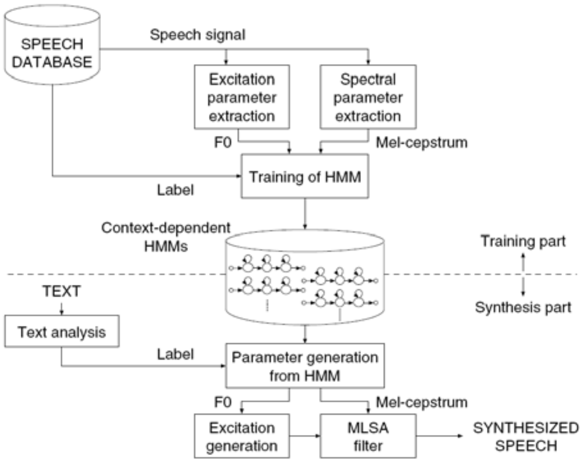


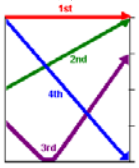
Figure 1: An overview of the basic HMM-based speech synthesis system (adopted from [8]).

2.2 Mandarin Chinese Issues

Mandarin Chinese is a monosyllable and tonal language, i.e., each Chinese character is pronounced as a syllable. Fig. 2(a) shows the hierarchical phonetic structure of syllables. There are five tones in Mandarin Chinese (including a lexicon one). Fig. 2(b) shows the relative pitch contours of the four Chinese tones. There are about 1300 syllables in Mandarin Chinese. If we discard tone difference between syllables, there are only about 412 toneless syllables. These toneless syllables could be further decomposed into 21 initials plus 38 finals.

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

(a)



(b)

Figure 2: The phonetic structure of Mandarin Chinese syllable: (a) hierarchical structure of Mandarin Chinese syllable (consonant, media and nasal are optional), (b) relative pitch changes of the four tones.

2.2.1 Synthesis Units

Since the number of toneless syllables is small (compared to Western language), Mandarin Chinese TTS systems usually use toneless syllables as the basic synthesis units.

However, in our HTS-based system, sub-syllable (initials and finals) are chosen for two reasons. (1) The number of monophones is reduced to only 59 (21 initials plus 38 finals). (2) Initial/final units are very popular recognition units in conventional Mandarin automatic speech recognition (ASR) [9] systems. Beside, the recognition performances of those initial/final-based Mandarin ASRs are usually comparable with the triphone-based ones.

2.2.2 Decision Tree and Question Set

Considering the characteristics of Mandarin Chinese the question set used in the decision tree-based model clustering is composed of 6 layers and listed in Table 1.

Table 1: Hierarchical structure of question set for decision tree-based model clustering.

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 position of syllables in a word
Word	the part-of-speech (POS) of current and surrounding words; the number and position of words in a phrase
Phrase	the number and position of phrases in an clause
Clause	the number and position of clauses in an utterance
Utterance	the number of syllables, words, phrases and clauses in an utterance

3 Voice Building

The whole setting and process explained in the following subsections is fully automatic and applied through all our submissions.

3.1 Speech Parameters

First, excitation (pitch) parameters are extracted with (RAPT) algorithm [10] which is based on normalized cross correlation function (NCCF) and dynamic programming. Then, 24-order mel-generalized cepstrum (MGC) [11] is extracted as the spectral parameters. Beside, their first and second order derivative features are also generated.

3.2 Training Procedures

The steps for voice building are showed in Fig. 3. First, 61 monophone (including silence and short pause) HMMs with 5 states, left-to-right transition and diagonal covariance matrix are trained according to the segmentations given by force-alignment.

Then all possible context-dependent HMMs (CD-HMMs) are generated by duplicating those monophone HMMs, i.e., model expansion. After well retraining those CD-HMMs, decision tree-based model clustering using MDL criterion is applied to shrink the number of models, i.e., model clustering. This expansion-then-clustering step is executed twice. After that, the number of mixtures in each model is increased.

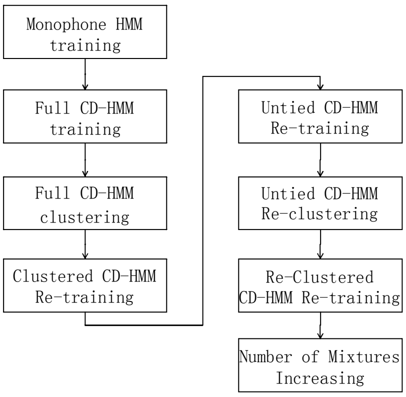

Monophone HMM training
Full CD-HMM training
Full CD-HMM clustering
Clustered CD-HMM Re-training
Untied CD-HMM Re-training
Untied CD-HMM Re-clustering
Re-Clustered CD-HMM Re-training
Number of Mixtures Increasing

Figure 3: The block diagram of the voice building procedure.

3.3 Speech Synthesis

Mel Log Spectrum Approximation (MLSA) [12] filter is applied for speech synthesis.

4 Evaluation

4.1 Mandarin Chinese database

A female Mandarin Chinese database known as 'WJ' was released by iFLYTEK [13]. There are 6000 utterances (about 130,000 Chinese characters) in this database. Only 1000 utterances of this database are labelled manually (including PinYin sequence checking, segmentation and prosodic boundary labelling). The remaining sentences are labelled automatically.

4.2 Tasks

There are three tasks including the main hub (MH) and two spoke (MS1 and MS2) voices:

- MH: build a voice from the full Mandarin database (about 6000 utterances / 130000 Chinese characters)
- MS1: build voices from the specified 'M_SMALL10', M_'SMALL50' and 'M_SMALL100' datasets, which consist of the first 10, 50 and 100 sentences respectively of the full Mandarin database.
- MS2: build a voice from the full Mandarin database suitable for synthesizing speech to be transmitted via a telephone channel. A telephone channel simulation tool [14] is available to assist in system development.

4.3 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 - volunteers
- MS - speech experts

4.4 Tests

All systems were evaluated with respect to naturalness, similarity and intelligibility:

- 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:
  (1) convert any traditional Chinese characters to simplified Chinese characters and calculate character error rate (CER) using a similar procedure to word error rate (WER), treating each character as a word. No spelling correction was used.
  (2) convert each character to pinyin+tone (a one-to-many mapping); the result is a lattice of possible pinyin+tone sequences
  (3) calculate pinyin+tone error rate (PTER), choosing the pinyin+tone path through the lattice that gives the lowest PTER
  (4) strip the tones leaving only pinyin, and calculate pinyin error rate (PER), choosing the pinyin path through the lattice that gives the lowest PER

4.5 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 11 systems for MH, 6 systems for MS1 and 8 systems for MS2 tasks. It must be stressed that System A is natural speech, System C is a standard speaker-dependent HMM-based voice built using a similar method to the HTS entry to Blizzard 2005 [15], System D is a speaker-adaptive HMM-based voice, built using a similar method to the HTS entry to Blizzard 2007 [16]. Finally, Systems E to W are the participants.

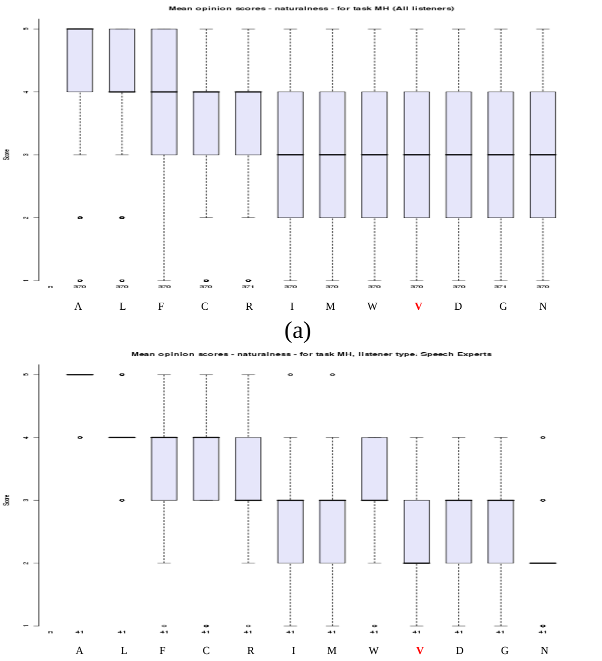
The final results are commented in the following lines comparing our performance (System V) with that of the other participants.

4.5.1 MH Task

4.5.1.1 MH Mean Opinion Score Test

MOS comparative between our (system V) and all other systems is shown in Fig. 4 for (a) all the listeners and (b) speech experts.

Our system got 3 points for all listeners but only 2 for speech experts (this is the worst case for our system). Our guess is that there are indeed some pop noises (overflow or underflow, although not very serious) in our synthetic samples. Therefore, some speech experts may think there are some faults in our system. After further investigation, we found that those pop noises may come from the global variance (GV) algorithm [17]. Because, those pop noises could be removed by either (1) disable the GV option in HTS or (2) reduce the amplitude of the original training waveforms.



(a)

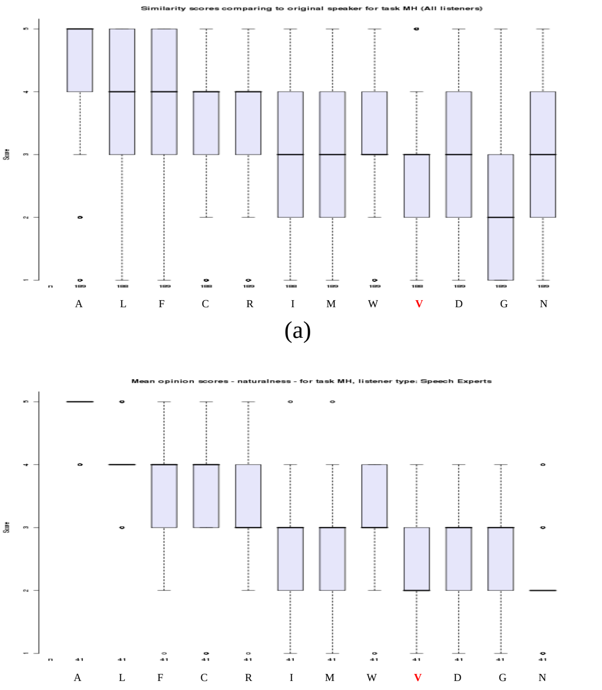
(b)

Figure 4: MOS comparative between our (V) and all other MH systems for (a) all the listeners and (b) speech experts.

4.5.1.2 MH Similarity Test

The boxplots of similarity scores of all systems are shown in Fig. 5 for (a) all listeners and (b) speech experts.

From the figure, we can conclude that our system performs much worse than the average of the rest of the systems in the similarity test. This is in fact the major weakness of our system. Moreover, speech experts gave only 2 points (the worst case) again.



(a)

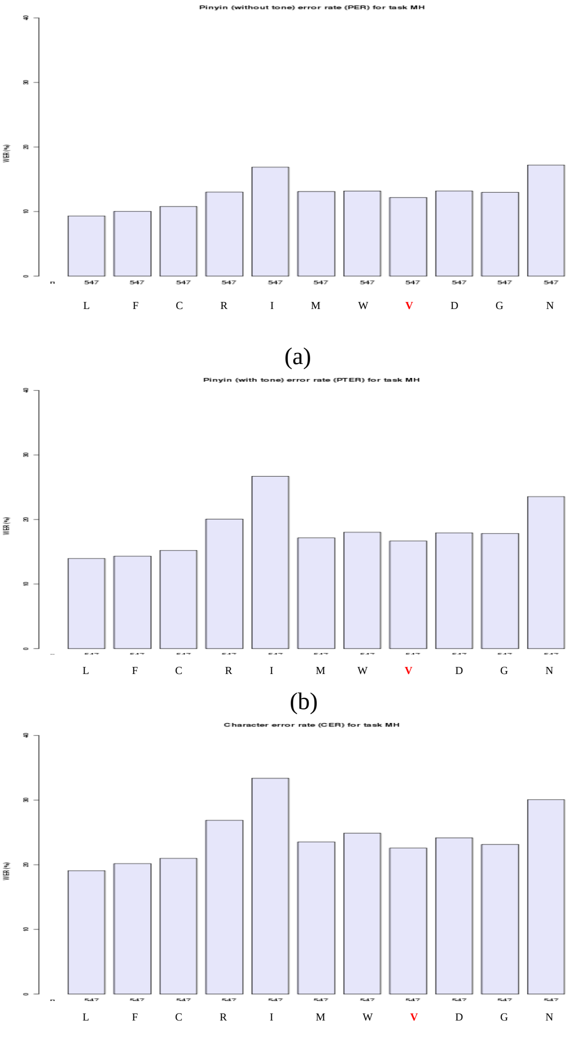
(b)

Figure 5: Similarity comparative between our (V) and all other systems for (a) all the listeners and (b) speech experts.

4.5.1.3 MH Word Error Rate Test

Fig. 6 shows the (a) PER, (b) PTER and (c) CER achieved by all the MH participants for intelligibility test.

According to the test results, our MH voice achieved 12.2% PER, 17.0% PTER and 23.0% CER. The performance of our system is number 4 among all 11 systems. This is the strongest point of our system.



(a)

(b)

(c)

Figure 6: Intelligibility comparative between our (V) and all other MH systems for (a) PER, (b) PTER and (c) CER.

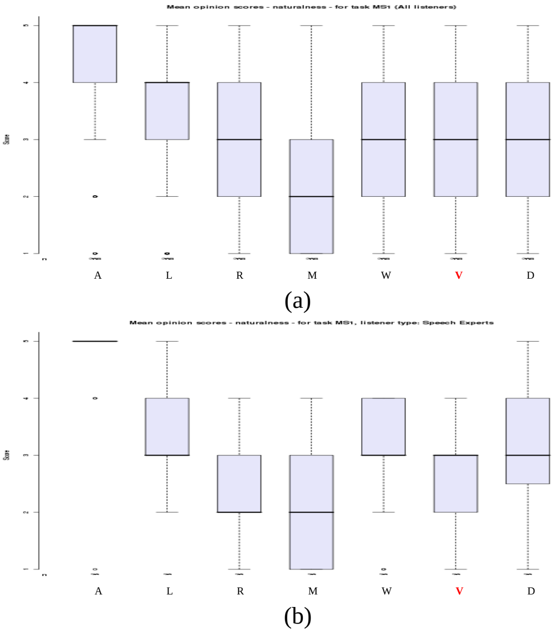
4.5.2 MS1 Task

Although three MS1 voices were built from the specified 'M_SMALL10', M_'SMALL50' and 'M_SMALL100' datasets, respectively, only the M_SMALL100 voice was evaluated. There are only 6 systems for MS1 task.

4.5.2.1 MS1 Mean Opinion Score Test

MOS comparative between our MS1 (system V) and all other systems is shown in Fig. 7 for (a) all the listeners and (b) speech experts. Our system got 3 points from both all listeners and speech experts. But again, speech experts gave more negative opinions (the worst case for our system).

We are very surprising to find that the performance of our MS1 voice built from 'M_SMALL100' dataset is almost comparable with our MH voice. On the other hand, although our voices built from 'M_SMALL50' and 'M_SMALL10' datasets degraded a lot but still acceptable (at least for 'M_SMALL50' case). This may confirm the generalization power of decision tree-based model clustering method.



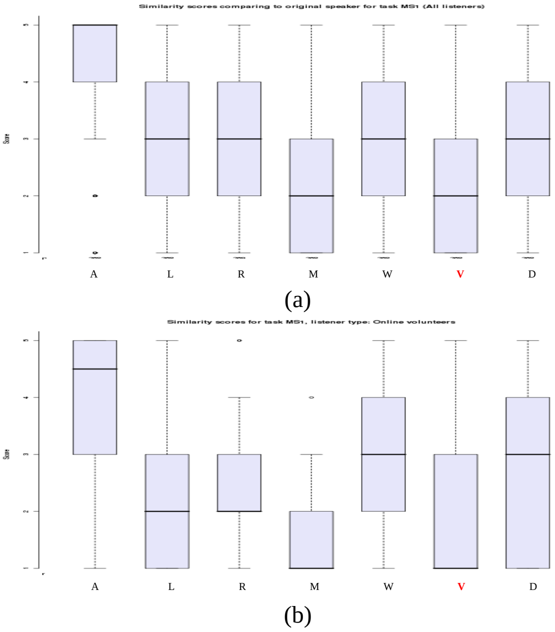
(a)

(b)

Figure 7: MOS comparative between our (V) and all other MS1 systems for (a) all the listeners and (b) speech experts.

4.5.2.2 MS1 Similarity Test

The boxplots of similarity scores of all MS1 systems are shown in Fig. 8 for (a) all listeners and (b) online volunteers. From the figure, we can again conclude that our system performs much worse than the average of the rest of the systems in the similarity test. Moreover, online volunteers gave only 1 point (the worst case).



(a)

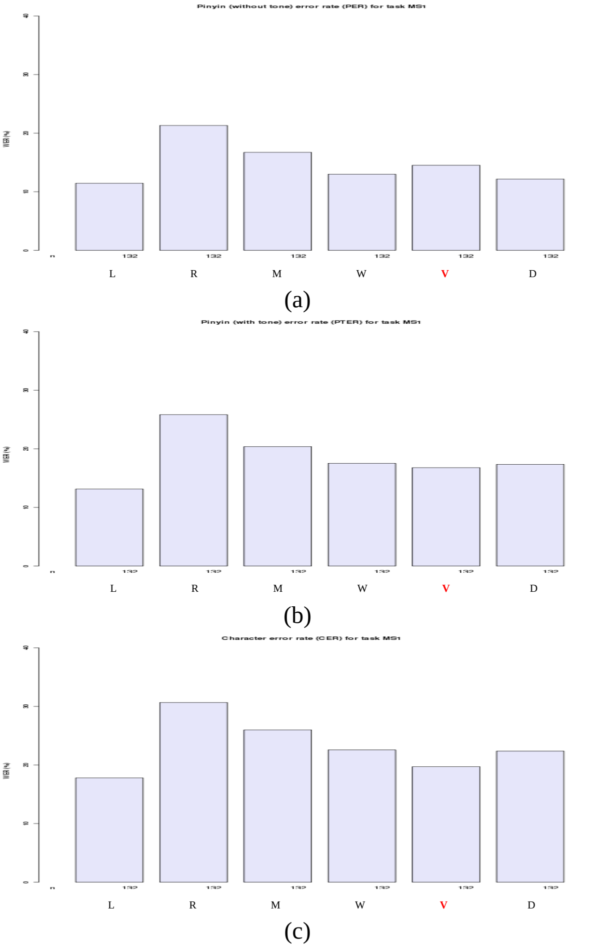
(b)

Figure 8: Similarity comparative between our (V) and all other MS1 systems for (a) all the listeners and (b) speech experts.

4.5.2.3 MS1 Word Error Rate Test

Fig. 9 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 15% PER, 17% PTER and 20% CER. Comparing with our MH results in Fig. 6, there is almost no difference between the intelligibilities of our MS1 and MH systems. This may again confirm the generalization power of decision tree-based model clustering approach.



(a)

(b)

(c)

Figure 9: Intelligibility comparative between our (V) and all other MS1 systems for (a) PER, (b) PTER and (c) CER.

4.5.3 MS2 Task

According the evaluation rules, MS2 voice could be either same as MH one, or built from the full dataset with the purpose of transmitting via a telephone channel.

There are 8 systems for MS2 task. Our MS2 voice is in fact the same as our MH one. The only difference is that it was post-processed using a telephone channel simulation tool supplied by International Telecommunication Union (ITU) [14].

4.5.3.1 MS2 Mean Opinion Score Test

MOS comparative between our (system V) and all other systems for all the listeners is shown in Fig. 10. It is quite surprising that our system got 4 points for all listeners since our MH voice achieved only score 3. Beside our system became number 2 (same score with system C and F) among 8 systems.

A good explanation for that is the pop noises in our synthetic samples is somehow alleviated by the low-pass characteristic of telephone channel. However, our system is now comparable to system C which was built using advanced STRAIGHT vocoder. This may suggest that the major improvement of STRAIGHT come from the preservation of speech information in higher frequency bands. In other words, for the case of telephone channel transmission, it may be fine to use only traditional speech parameter extraction approaches.

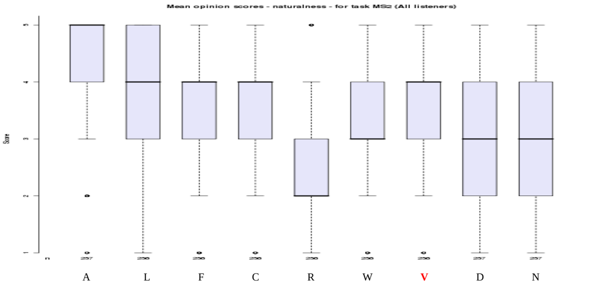
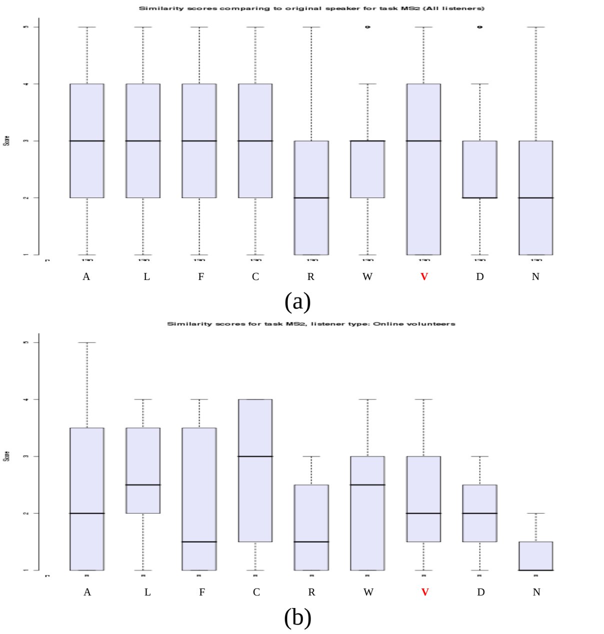


Figure 10: MOS comparative between our (V) and all other MS2 systems for all the listeners.

4.5.3.2 MS2 Similarity Test

The boxplots of similarity scores of all systems are shown in Fig. 11 for (a) all listeners and (b) online volunteers. From the figures, we can conclude that our system in similarity test may be more acceptable in the case of telephone channel transmission.



(a)

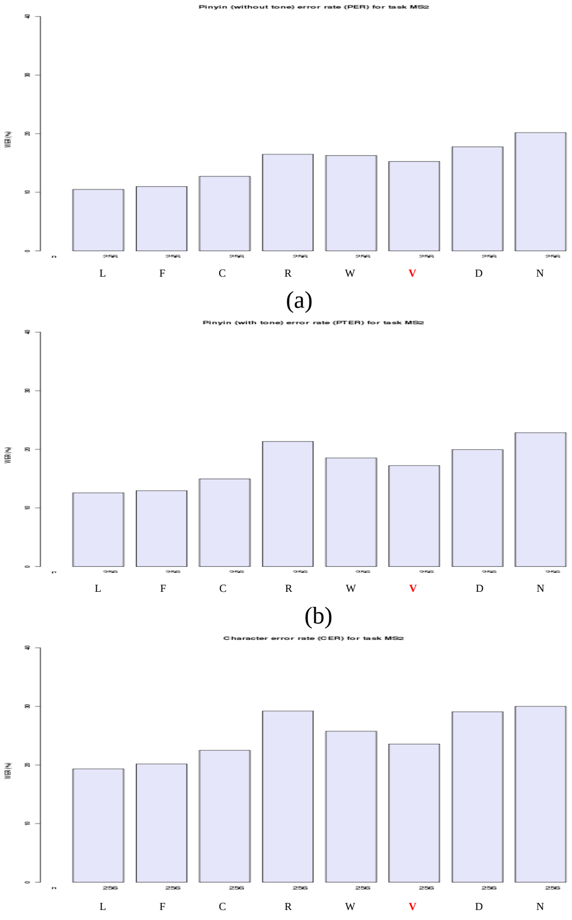
(b)

Figure 11: Similarity comparative between our (V) and all other MS2 systems for (a) all the listeners and (b) online volunteers.

4.5.3.3 MS2 Word Error Rate Test

Figure 12 shows the (a) PER, (b) PTER and (c) CER achieved by all the MS2 participants for intelligibility test.

According to the results, our MS2 voice achieved 15% PER, 17% PTER and 24% CER. This is similar to the results of our MH and MS1 voices and our system is number 4 among all 8 systems.



(a)

(b)

(c)

Figure 12: Intelligibility comparative between our (V) and all other MS2 systems for (a) PER, (b) PTER and (c) CER.

5 Conclusions and Future Works

Blizzard challenge 2009 has been our first attempt to build a TTS system and also our first participation in an international TTS evaluation campaign. The listening tests carried out as part of the Blizzard challenge 2009 have shown that our MH voice got 3 points for both MOS and similarity tests. Beside, 12.2% PER, 17% PTER and 23% CER were achieved for intelligibility test. Moreover, our MS2 voice achieved 4 and 3 points for MOS and similarity test, respectively.

Since, our main goal to participate this challenge is to establish reasonable TTS baselines in order to develop and verify our own advanced prosody model in the future; we consider these results as promising. We will continue to improve our systems and hope to have some contributions in the area of prosody model and prosody prediction.

Acknowledgements

This work was partially supported by the National Science Council, Taiwan, under the projects with contract NSC 96-2221-E-027-100-MY2 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]. Yuan-Fu Liao, Zi-He Chen and Yau-Tarng Juang, “Latent Prosody Analysis for Robust Speaker Identification”, *IEEE Trans. on Speech, Audio and Language Proc.* Aug. 2007.

[4]. Yuan-Fu Liao, Wen-Chieh Chang, Zong-You Xie, Ding-Yun Zeng and Yau-Tarng Juang, “Joint Prosodic and Spectral Modeling for Robust Speaker Verification”, Speech Prosody’2008.

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

[6]. Hideki Kawahara, Ikuyo Masuda-Katsuse and Alain de Cheveigne: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, Speech Communication, 27, pp.187-207, 1999.

[7]. Hideki Kawahara: STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds, Acoustic Science and Technology, Vol.27, No.6, 2006.

[8]. 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.

[9]. Y.-F. Liao, N. Wang, M. Huang, H. Huang and F. Seide, “Improvements of the Philips 2000 Taiwan Mandarin Benchmark System”, ICSLP, pp.298-301, Beijing, Oct., 2000.

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

[11]. 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.

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

[13]. Anhui USTC iFLYTEK, CO., LTD, http://www.iflytek.com/english/index.htm

[14]. ITU-T Recommendation G.191, Software tools for speech and audio coding standardization, http://www.itu.int/rec/T-REC-G.191-200509-I/en

[15]. 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

[16]. 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

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