The Toshiba Mandarin TTS System for the Blizzard Challenge 2009

Jian Li, *Jian Luan*, *Lifu Yi*, *Xiaoyan Lou*, *Xi Wang*, *Liqiang He, Jie Hao*

Research and Development Center, Toshiba (China) Co., Ltd., Beijing, China

{lijian,luanjian,yilifu,louxiaoyan,wangxi,heliqiang,haojie}@rdc.Toshiba.com.cn

Abstract

This paper introduces the Toshiba Mandarin Text-to-Speech (TTS) system submitted to the Mandarin benchmark of the Blizzard Challenge 2009. The basic framework keeps unchanged with the system in 2008 and we modify the system in several aspects: automatically find bad units in the database when preparing the speech corpus, add a G2P procedure after the text analysis to increase accuracy of the predicted pinyin for heteronyms[1], introduce prosody layer information into the prosody modeling and modify the fusion methods to fuse units in frequency domain. The subjective evaluation results show that these modifications improve the performance.

1. Introduction

Since 2005 Blizzard Challenge has been held every year. The Challenge provides a platform to evaluate different TTS technologies based on common databases.

The Blizzard Challenge 2008 extended the evaluation language to Mandarin. At that time we just finished developing our new Mandarin TTS using Toshiba's own "plural unit selection and fusion" method [1], and we participated in the Challenge to benchmark our Mandarin TTS system. Our method combines the ideas of both conventional unit selection method and Toshiba's closed-loop training method [2] so that the synthetic speech sounds not only clear at the segment level but also smooth and stable throughout the whole sentence. The evaluation result showed that our Mandarin TTS system achieved high performance in blizzard Challenge 2008 [3].

However, the speech quality is not good enough. As pointed in [3], several aspects can be improved for our system. Compared with last year's system, trials in the following four aspects were made:

- In the corpus preparation stage, we use automatically way to find bad units
- After pronunciations (pinyins) are given by text analysis, a G2P (grapheme-to-phoneme) procedure is added to increase the pinyin accuracy for the heteronyms.
- Prosodic layers information, mainly PW (prosodic word) and PP (prosodic phrase), is used to train prosodic models.
- In back-end, the units are fused in frequency domain instead of in time domain

The evaluation result shows that these modifications improve the system performance.

This paper is organized as follows. Section 2 is an overview of the Toshiba's Mandarin TTS system, the system tuning and the speech corpus preparation. In section 3 the improvements made for the Blizzard challenge 2009 and the method to cope with the synthetic speech vial telephone channel for task MS2 are introduced. In section 4 the evaluation results of our system are analyzed and some discussions are presented. Finally the conclusions are addressed in section 5.

[1] Some Chinese characters have more than one pronunciation. These Chinese characters are called as heteronyms in the paper.

2. System Overview

Toshiba's TTS system consists of two major parts: the front-end and the back-end. The front-end predicts the prosodic information. The back-end generates speech waveform and in back-end some parameters need to be optimized according to the language or the speaker.

For our system, the speech corpus preparation mainly consists of labeling the phoneme boundaries, segmenting the training sentence into words, and providing syntactic information about the words.

In this section the front-end, back-end, system parameter tuning and the steps to prepare the database will be briefly introduced.

2.1. Front-end

2.1.1. Text analysis

The text of each prompt sentence has already been normalized. The normalized text is syntactically analyzed to obtain:

- The word segmentation of the sentence.
- The POS (part of speech) of each word in the sentence
- The pronunciation (pinyin with tone) of each Chinese character in the sentence.

From the above basic information, some other linguistic information can be further calculated, such as the position of a Chinese character in a word, the position of a word in the sentence, the length of the word, and so on. The phonetic and linguistic information are the basis for the prosodic prediction. They are referred to as attributes hereinafter.

In this year, a G2P module is added after text analysis to increase the accuracy of predicted pinyin for heteronyms. This procedure will be introduced in section 3.

2.1.2. Pause

Pause model is to predict pauses from a sequence of contextual linguistic attributes for each segmented word. A generalized linear model (GLM) [4] is used to predict pause for our TTS system [5].

The error distribution of pause is assumed to obey a Bernoulli distribution in our system. Accordingly, the Logistic GLM [5] is applied to handle the Bernoulli distribution.

For each segmented word, the attributes of the left three words and the right two words are used as the attribute set, which include POS and word length. The attribute set is automatically selected by stepwise regression, which is a totally data-driven method. Open tests show the proposed method outperforms CART.

After pause prediction, the distance of a Chinese character to the next and previous pauses can be calculated, which are important for pitch prediction and duration prediction.

2.1.3. Duration

Quantification Method Type 1 (QMT1) is used for duration modeling. QMTI is a 'sums-of-product' like method and it handles discrete variables. For duration prediction, the linguistic and phonetic attributes, such as the part of speech, the tone of the phoneme, distance to previous and next pause, and so on, are discrete variables of QMT1.

2.1.4. F0 contour

In our system a codebook-based F0 contour model [6][7] is used. F0 contour is expressed by two parts, one is the shape of the contour, which is called as *F0 pattern* in our system, and the other is the offset level of a F0 contour.

In training phase, for each tone a codebook of representative F0 patterns is firstly obtained from the speech corpus by vector quantization clustering method. Then, for a F0 contour c_i in the speech corpus, the approximation error e_{ij} is calculated if it is generated by the representative F0 pattern P_j in the codebook of the corresponding tone. The approximation errors and corresponding attributes are used to train QMT1 model.

In prediction phase, firstly the attributes for each syllable S_i are calculated from the text analysis results. Then the approximation error $\hat{e}_{ij}$ is calculated using the QMTI model. The representative F0 contour with the minimal error is selected.

Similarly, the offset of a F0 contour can be trained and predicted using QMT1 method.

After the optimal F0 pattern is selected and the F0 offset level for the syllable is also predicted, they are combined to generate the F0 contour for one syllable. Then the F0 contour will be expanded or contracted by the predicted duration. And finally, we concatenate the F0 contours of all syllables and thus generate the pitch contour of the whole sentence.

2.2. Back-end

In the back-end, the "plural unit selection and fusion" method is used. It uses two steps to create a speech unit: unit selection and unit fusion. In the unit selection step, multiple speech units, rather than a single unit as in conventional unit selection method, are selected for each target segment according to their target costs and concatenation costs to the neighboring units. In the unit fusion step, the selected multiple units are averaged to generate a new fused speech unit. Then the fused speech units are modified according to the predicted prosody and concatenated with each other to generate the speech waveforms.

2.2.1. Selection of multiple speech units

Target cost and concatenation cost are used to select the multiple units for a target segment. The target cost includes the cost of phonetic context to the target segment, duration cost, pitch cost at the beginning and ending point and so on. The concatenation cost is mainly calculated from MFCC spectral at the concatenation position.

Firstly a pre-selection step is applied to select the most possible candidates using only target cost. The number of candidate units is well controlled after the pre-selection to reduce the computation burden in the following unit-searching step. Then in the unit-searching step, dynamic programming is used to search for an optimal path of primary speech units through a sentence using both target cost and concatenation cost. After that, secondary speech units are selected based on the target cost and their concatenation cost with the neighboring primary speech units. The number of secondary speech units can vary from speaker to speaker as well as according to the system configuration, such as memory footprint. Basically, more secondary speech units can help to improve the robustness in the synthetic speech.

2.2.2. Unit fusion

The selected multiple units for the same target segment are fused together to generate a new speech unit. The unit fusion is based on pitch-cycle waveforms.

Firstly the selected units are lengthened or shortened by aligning their pitch marks to fit the predicted pitch marks in the target segment. The pitch-cycle speech waveforms of the primary and secondary speech units are then averaged. In 2008 the fusion is done in time domain while in this year the fusion is done in frequency domain. In section 3.3, the new fusion method will be introduced.

After fusion, an LSP formant-emphasis filter is used to make the fused speech unit sound clearer.

2.2.3. Unit concatenation

The pitch-cycle speech waveforms of the fused speech unit are aligned according to the predicted pitch mark and overlapped with each other to form the final speech waveform of the synthetic speech.

2.3. System tuning

Two important parameters in back-end need to be tuned. One is the number of units to be fused, *N*, and the other is the strength of the formant emphasis filter that is applied to the fused pitch-cycle waveforms [8].

When N equals to 1, our method is similar to conventional unit selection method. When N is increased, the stability of synthetic speech is increased while the clearness of the segments may be degraded. The formant emphasis filter is used to improve the clearness of the fused segments. The parameter of the emphasis filter is related to N. So the task of system tuning is to find a good combination of these two parameters, as judged by a subjective evaluation. From an internal subjective evaluation the combination of fusing 6 units and using a medium setting of the formant emphasis filter was selected as the best combination. This combination is different from that used for English in 2007 [8] and Mandarin in 2008 [3].

2.4. Speech corpus preparation

The word segmentation and syntactic information of the training data can be done by the text analysis of front-end, which is introduced in section 2.1, so only the phoneme boundary labeling will be introduced here. The steps are the same with those in last year except one new step is added to automatically find the bad units in the database. The steps that are same with last year's will be simply introduced below and the new step will be introduced in section 3.

2.4.1. *Manual check of text and pinyin*

This step is to correct the text and pinyin that are different from actual pronounced syllables in the recorded speech. These differences are caused by mispronunciation by speaker or some other reasons.

2.4.2. *Automatic phoneme segmentation*

The forced alignment tool of HTK [9] was used to segment the speech automatically.

In our system the boundaries between the closure and the release of plosives or affricates are important. Since it is difficult to model the closures using HMMs and segment closure and plosives by forces alignment, we just set the closure/release (CR) boundary at one third of the way through the plosive/affricate. Of course, these CR boundaries are not good and need to be refined.

2.4.3. *Automatic refining the segmentation*

The performance of forced alignment is not good enough for the purposes of TTS. So an automatic refinement tool has been developed as a post-process to refine the phoneme boundaries. The tool focuses on refining boundaries between voiced and unvoiced phonemes (VU boundaries) and CR boundaries (as discussed in the previous section). The features in time domain, such as energy and zero crossing rate, are used to refine the segmentation.

2.4.4. *Manual check of the segmentation*

In this step the boundaries between voiced initial, /l/, /m/, /n/, /r/ and their following finals, and the bad VU and CR boundaries after automatic refinement, are manually checked. The boundaries between two finals are not refined because these boundaries are not easy to identify, even manually.

After above steps, a tool is used to automatically find the bad units in the database. It will be introduced in section 3.1.

3. Changes in 2009

Our system achieved high performance in 2008. As pointed in [3], the system can be further improved in several aspects. This section introduces the modifications made in speech corpus preparation, in front-end and in back-end respectively.

3.1. Automatically find bad units

This is an additional step in speech corpus preparation stage. For a unit selection-based system, it is important to ensure that the selected units are correct and good. If abnormal units are selected, the quality of the synthetic speech will be degraded. The figure below shows an abnormal initial /h/ which contains impulse noise. If it is selected, the noise will be perceived in the synthetic speech.

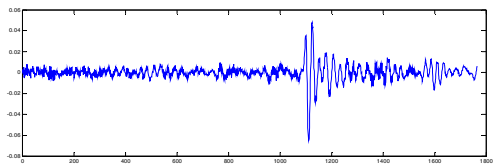

Figure 1. Waveform for segment 'h'

The initials are quite different with finals in both time domain and frequency domain, so different methods are applied to automatically find bad initials and bad finals.

3.1.1. *Find bad finals*

Firstly a mono-phone HMM is trained for each final using isolated training method of HTK with the labeled boundaries. Then for each final instance in the database, its likelihood score is calculated. The units whose scores deviate the average score too much are regarded as 'bad' units and will be removed. The cases of the found 'bad' units include label errors, pronunciation mistakes, retroflexed syllables and so on.

3.1.2. *Find bad initials*

The bad initials containing impulse noise are the focus. Firstly the short-time energy is calculated with frame length of 2ms and frame shift of 1ms. Then the delta-delta coefficient of the frame energy is calculated. The initials with big delta-delta coefficients can be considered that impulse noise is included. This method can also find some badly pronounced initials that have sharp energy changes.

This is a very rough method to find bad initials and finals. Some units found by this method are actually good units and some bad units in the database cannot be found. Better methods should be surveyed for this task in the future.

3.2. G2P procedure

The total number of Chinese characters is fixed so the lexicon can contain all Chinese characters and their pinyins. Some Chinese characters have more than one pinyin and are called as heteronyms. Currently in our system, if a heteronym is included in a segmented word, the correct pinyin can be given using the information of the word entry in the lexicon. But in other cases, no special methods have been applied to help finding the correct pinyin.

The pronunciations for the heteronyms in sentences are determined by their meanings. Language understanding is a hard task in itself so in the new G2P procedure, the context information of heteronyms is used to help selecting the correct pinyin.

Firstly some important heteronyms that are frequently used are selected. And a corpus is prepared covering the different contexts for these heteronyms as possible as we can. Then the commonly used machine learning methods including GLM, TBL, C4.5 and CART are tested and for each important heteronym the best method is selected. The experiment shows that the relative error reduction of the heteronym pinyin prediction is more than 50%.

3.3. Change in prosody modeling

Prosody layers play an important role in the speech communication process. Generally, the bottom-up levels of Mandarin prosody hierarchy consist of syllables, prosodic words (PW) and prosodic phrases (PP) [10]. Some important prosody changes that have great impact on the naturalness of the synthetic speech, such as perceived silence pauses, syllables lengthening and pitch resets are usually associated with PW and PP boundaries. So in this year, the prosody layers are introduced to improve the performance.

3.3.1. *Utilize prosody layers*

Among the Mandarin prosody hierarchy, PW and PP are used in prosody modeling. New attributes are calculated from them, such as whether current syllable is in the boundary of a PW, whether it is in the boundary of a PP, and the distance from the previous and to the next PW and PP. The

performance using prosody layers (NEW) is compared with the performance without prosody layers are listed in Table 1.

Table.1: The open test performance comparison between using and not using prosody layers

	RMSE	Corr
Duration (for finals)	26.86 ms	0.8004
NEW_Duration (for finals)	25.10 ms	0.8875
F0	29.44 Hz	0.8254
NEW_F0	28.06 Hz	0.8422

RMSE in Table 1 means the root mean squared error and Corr represents the correlation. The results show that both the duration model and the F0 model are improved by using the prosody layers.

3.3.2. Prosody layer prediction

In our system, the predicted pauses are used as prosodic phrases. Pause prediction is introduced in 2.1.2 and here the prosodic word prediction is introduced.

Conditional random fields (CRF) method [11] is adopted to label the prosodic words from lexical word sequences given by the text analyzer.

For a lexical word w_i segmented by text analyzer, the task is to determine whether the lexical word boundary is a prosodic word boundary or not. An attribute window covers the previous 3 lexical words and next 3 lexical words is used for attribute extraction. The POS and word length of each word in the attribute window are extracted as the attribute set. The utterances in the Blizzard corpus are split into two parts with 70% for training and 30% for open test. The CRF performance is shown in the Table 2.

Table 2: The CRF performances for prosodic words

	Precision	Recall	F-score
Close test	0.9994	0.9993	0.9993
Open test	0.9332	0.9289	0.9310

3.4. Change in back-end

Unit fusion is a key module in Toshiba TTS system back-end to generating high quality synthetic speech. Instead of sub-bands waveform averaging, a more precise fusion method in frequency domain is proposed in this year. As well known, human hearing is mostly sensitive to the magnitude spectrum of speech signal. For this reason, directly averaging the magnitude frequency can eliminate the influence of phase difference and achieve more effective fusion for the subject perception.

Firstly the magnitude spectrum and phase spectrum of each pitch-cycle waveform are obtained by FFT. Then the magnitude spectrums of selected multiple units are averaged on logarithm scale while the phase spectrum of a primary unit (as described in section 2.2.1) is directly used. Finally, the fused pitch-cycle waveform is reconstructed by IFFT.

At the boundary of units, the fused pitch-cycle waveforms need further smoothing for the consistency. The first pitch-cycle waveform of one unit is used to smooth its previous unit while its last one is used to smooth the next unit. A fade-in/fade-out weight scheme is applied for smoothing. Similar with the fusion, the smoothing process is also conducted in frequency domain. The magnitude spectrum is smoothed through a weighted averaging on logarithm scale. Specially, the phase spectrum may also be weighted averaged for low frequency in smoothing.

A small-scale internal evaluation was carried out to compare the performance of new fusion method with the old one. 40 sentences are randomly selected out of the training data for 7 listeners. The preference result is shown in Fig. 2. It indicates that new fusion method remarkably outperforms the old one.

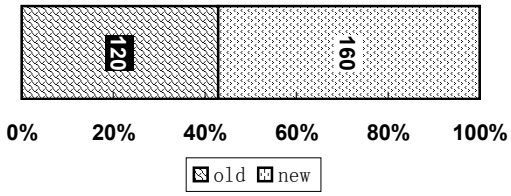

Figure 2: Preference results of subject evaluation on unit fusion

3.5. Method for synthetic speech via telephone channel

From experiments it was observed that the high frequency part of synthesis speech was heavily degraded and sounded like buzzing noise after transmitted vial the telephone channel. Thus, a simple magnitude spectrum modulation was applied to attenuate the high frequency part. The modulation function is defined as the equation 1.

$$w(f) = 1 - \alpha * \sin\left(\frac{f}{F} * \frac{\pi}{2}\right) \quad (1)$$

where f is the frequency index, F is the Nyquist frequency and α is an adjusting factor. With the growth of α, the high frequency can be gradually attenuated which leads to the increase of intelligibility but the decrease of similarity. The tradeoff between the intelligibility and the similarity should be considered.

4. Evaluation result

There are three tasks in this year's Mandarin benchmark. The hub task (MH1) is to synthesize speech to be evaluated in the quiet environment using all data. The other twos are spoken tasks: one is to use the first 10, 50 and 100 sentences of the database to build the voice (MS1) and the other is to synthesize speech to be transmitted via telephone channel (MS2). We participated in the hub task MH1 the 2nd spoken task MS2.

Three aspects of the synthetic speech were evaluated in the challenge: the naturalness of the synthetic speech, the intelligibility of the synthetic speech, and the similarity to the original voice.

4.1. Hub task MH1

Two HMM-based systems participated in Mandarin benchmarks in both 2008 and 2009 and they are used as reference systems. The speaker-adaptive system in this year is same with the entry HTS_2007 that was submitted to Blizzard 2007 while the speaker-adaptive system in 2008 was named as HTS_2008 that improved based on HTS_2007 for English. It is unknown whether HTS_2007 is the same with HTS_2008 for Mandarin and the result shows the performance of the speaker-adaptive systems are not stable in two years. So only HTS_2005, a speaker dependent system is used as reference system.

A 5-point scale mean opinion score (MOS) was applied to rate both the naturalness of the synthetic speech and the similarity to the original speaker. Figure 3 and 4 show the naturalness and similarity for our system and HTS_2005 in 2008 and 2009. We find both systems achieve improvement in 2009 while our system seems to make bigger progress than HTS_2005. This indicates that the modifications mentioned in section 3 are effective.

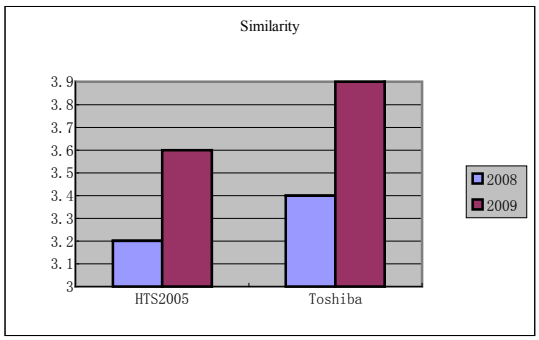

Figure 3. Similarity to the original speaker

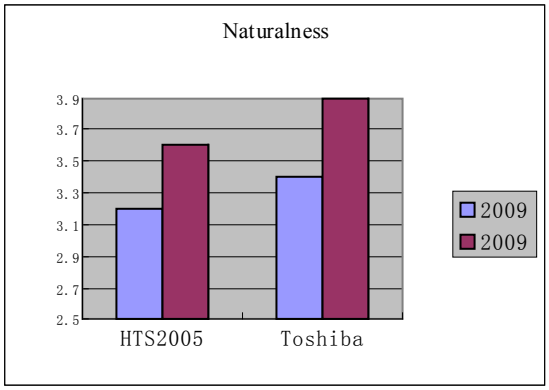

Figure 4. Nauturalness

Semantically unpredictable sentences (SUS) were used to evaluate the intelligibility of TTS systems. Subjects transcribed the sentences in Chinese characters. Since in Mandarin, one Chinese character may have several pronunciations, Pinyin-without-tone Error Rate (PER) and Pinyin-with-Tone Error Rate (PTER) are calculated as well as Character Error Rate (CER).

Since Mandarin is a tonal language, we think PTER is more important to reflect the intelligibility of a system. Figure 5 is the PTER result of two systems in 2008 and 2009. The performance of both systems in 2009 is worse than that in 2008. One possible reason is that the length of sentences in 2009 is longer than that in 2008 and more deletion errors are introduced.

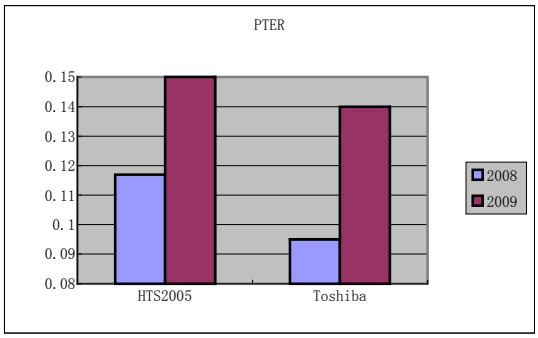

Figure 5. Intelligibility by PTER

4.2. Spoken task MS2

In task MS2, the synthetic speech is transmitted via a simulated telephone channel. A simple method to attenuate the high frequency band as described in section 3.3 is applied. This paper compares the result of MH1 and MS2. In the figures, A is the real speech, C is HTS_2005, D is HTS_2007 and F is Toshiba's system.

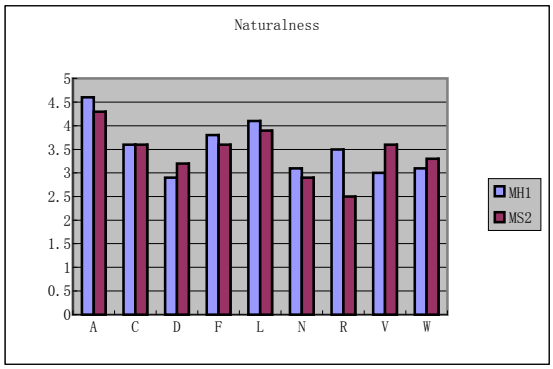

Figure 6. Naturalness for MS2

For naturalness test, some systems including real speech receive lower score in MS2 than in MH1 while some systems such as D, V and W get higher score in MS2. This shows that the effect of telephone channel on synthetic speech is complex and cannot simply to conclude that the telephone channel will degrade the speech quality. More detailed research should be done for telephone channel.

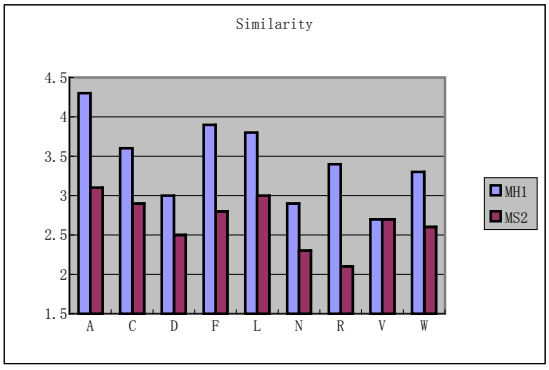

Figure 7. Similarity to the original speaker for MS2

For similarity to the original speaker, all systems receive lower score in MS2 than in MH1 except system V. The degradation of our system F is 1.1 and it is relative large comparing with other good systems such as C and L. The

major reason is that frequency modification changes the characteristic of the speaker.

As for intelligibility evaluation, some systems receive lower or equal PTER in MS2 than in MH1, which is contrary to the common sense that the noise introduced by the telephone channel will degrade the intelligibility of the speech. One of the reasons may be that in noise environment subjects paid more attention on listening.

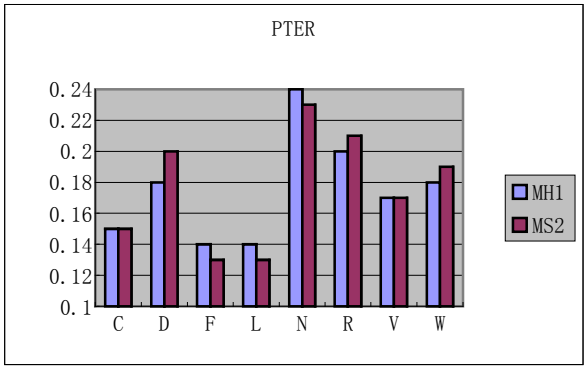

Figure 8. Intelligibility for MS2 by PTER

4.3. Some discussions

In both tasks MH1 and MS1, our system achieves high performance. However there is room for improvement.

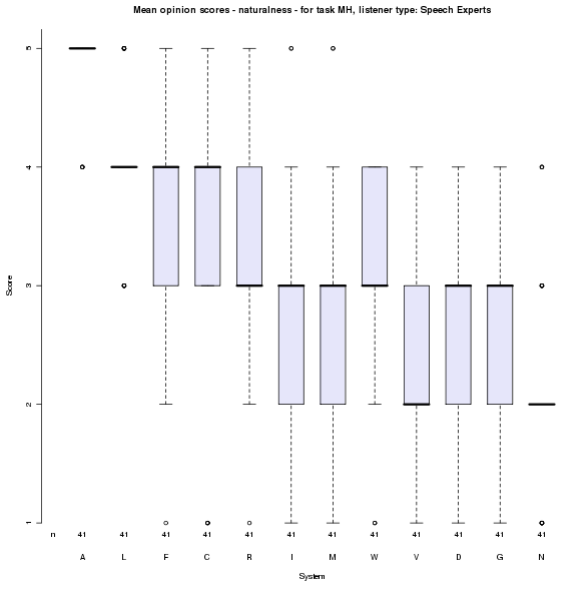

Figure 9. MOS score for MH naturalness by speech experts

Figure 9 is the MOS score for MH naturalness by speech experts. It shows that the scores for system A (real speech) and system L are very concentrated. While for our system (F), the scores are scattered. Although more sentences of our system are scored as 5 than system L, some sentences of our system are scored as 2 and 1 while for system L no sentences are scored as 2 and 1. It shows that our system is not robust and we need analyze the sentences of low scores for further improvement.

The intelligibility result of version 5, which corrected the answer of one sentence, is compared with that of version 4. The CER, PER and PTER of all systems decrease and the relative error reduction is about 10%. For example the CER of system F, G, L decrease from 0.22, 0.27 and 0.21 to 0.20, 0.23 and 0.19 respectively. This shows the intelligibility result is very sensitive even to a big error rate from only one sentence. In the SUS sentences, about 10% sentences whose CER are larger than 50%. The big error rates are mainly caused by the deletion errors while the deletion errors are basically caused by subjects rather than TTS systems. The effect of these sentences with big error rates on word error rate may not be ignored. So the detailed information such as the rate of deletion errors, substitution errors and insertion errors are important to help analyzing the intelligibility of the TTS systems.

5. Conclusion

This paper describes the Toshiba entry for the Mandarin benchmark of the Blizzard Challenge 2009. Keeping the basic framework of the system unchanged, we modify the system in the corpus preparation stage, the front-end as well as the back-end. The results show that our system achieves high performance and these modifications improve the quality of the synthetic speech.

6. References

[1] Mizutani, T. and Kagoshima T., "Concatenative Speech Synthesis Based on the Plural Unit Selection and Fusion Method", *IEICE Transactions on Information and Systems*, Vol. E88-D, No.11, pp.2565 – 2572, 2005.

[2] Akamine, M. and Kagoshima, T., "Analytic Generation of Synthesis Units by Closed Loop Training for Totally Speaker Driven Text to Speech System (TOS Drive TTS)," Proc. ICSLP'98, pp.1927-1930, 1998.

[3] Jian Li, Dawei Xu, Lifu Yi, Xiaoyan Lou, Jian Luan, Xi Wang, Liqiang He, Jie Hao, "The Toshiba Mandarin TTS System for the Blizzard Challenge 2008", Blizzard 2008, Sep. 2008

[4] McCullagh, P., Nelder, J. A., "*Generalized Linear Models*", Chapman & Hall press, 1989.

[5] Yi, L., Li, J. Lou, X. and Hao, J., "Phrase Break Prediction Using Logistic Generalized Linear Model", *Proc. of Interspeech 2006*, pp1308 –1311, 2006.

[6] Takehiko Kagoshima, T., Morita, M., Seto, S., Akamine, M., "An F0 Contour Control Model for Totally Speaker Driven Text to Speech System", *Proc. of ICSLP1998*, pp1975-1978, 1998.

[7] Suh, C. K., Kagoshima, T., Morita, M., Seto, S., Akamine, M., "Toshiba English Text-To-Speech Synthesizer (TESS)", *Proc. of Eurospeech1999*, pp.2111-2114, 1999.

[8] Buchholz, S., Braunschweiler, N., Morita, M., Webster, G., "The Toshiba entry for the 2007 Blizzard Challenge", *Proc. of Blizzard Workshop 2007* (in Proc. SSW6), Bonn, Germany, 2007.

[9] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. "The HTK Book", (for HTK Version 3.4), Cambridge, United Kingdom, 2006.

[10] Li Aijun, "Chinese prosody and prosodic labeling of spontaneous speech", *Proc. of Speech Prosody 2002*, pp. 39-46, 2002.

[11] J.D. Lafferty, A. McCallum, and F.C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", *Proc. of ICML, 2001*, pp.282-289, 2001.