The WISTON Text-to-Speech System for Blizzard Challenge 2009

Jianhua Tao, Ya Li, Shifeng Pan, Meng Zhang, Hongjun Sun, Zhengqi Wen

National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{jhtao, yli, sfpan, mzhang, hjsun, zqwen}@nlpr.ia.ac.cn

Abstract

This paper describes the WISTON system, a large corpus based TTS system that was submitted to Blizzard Challenge 2009. The text analysis part of this system contains text pre-processing, word segmentation, POS tagging, phonetic transcription and prosody structure prediction, most of which are based on Maximum Entropy (ME) models. In unit selection part, CART models are used to predict the prosodic parameters (duration, F0, energy), then we use concatenation costs and target costs for path searching to find the most suitable units to concatenate. The acoustic processing part is used for smoothing. The final system was used to attend Blizzard Challenge 2009 for both English test and Mandarin test.

Index Terms: Wiston, speech synthesis, prosodic chunk, unit selection

1. Introduction

The large corpus-based speech synthesis has been very popular for its high quality and naturalness speech output [1], and many TTS systems based on this principle are developed, the Wiston TTS system is one of these systems.

The whole system is composed of three parts, the text analysis module, the unit selection module and the acoustic processing module. The framework of this system is designed for multilingual speech synthesis. The aim of the text analysis module is to find the deep information of the utterance, then to get the prosodic structure, which make the synthesized speech sounds natural. The unit selection module is used to find the best units that were selected from a large corpus with prosody constraints, usually, the F0, duration and energy are considered into the constraints. Finally, the acoustic processing module smooths these units and generates the speech.

The rest of this paper is organized as follows: section 2 introduces the text processing module, especially, the prosodic chunk are put forward. Section 3 introduces the Wistion TTS system framework. In section 4, the overall object evaluation and subject evaluation are presented. The results show that with prosodic chunk prediction part, the system performs much better than that without prosodic chunk. The section 5 gives the final conclusions.

2. Hierarchical Text Analysis Module

2.1. The overall structure

Finding the suitable rhythm structure of a sentence is very important to achieve higher naturalness of Text-To-Speech system. For this, the text pre-processing, word segmentation, POS tagging and polyphone disambiguous in this module are conducted before the prosodic prediction part to get more textual information in hopes of gaining better performance. These parts are basically the same as that in Wistion system submitted to Blizzard Challenge 2008, except adding and deleting several rules in pre-processing which are used to changing some symbols to Chinese characters, especially, the digital processing part has been greatly enhanced with elaborated rules [2].

2.2. The prediction of prosodic structure

Various machine learning algorithms have been employed to predict the most likely positions for prosodic breaks in a text stream [3-5]. In our work, we separate the prosody structures into four levels: syllable, word (prosody word for Mandarin), minor prosody phrase and major prosody phrase.

Three Maximum Entropy (ME) models are separately to make boundary or non-boundary decisions for prosodic word (PW), prosodic phrase (PP) and intonation phrase (IP), respectively. Figure 1shows the flowchart of the hierarchical framework.

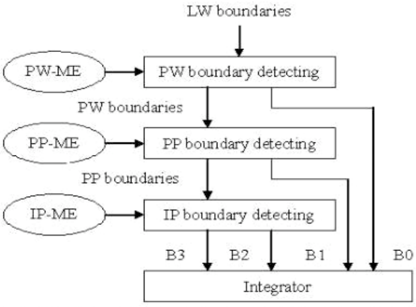

Figure 1 flow chat of the bottom-up hierarchical framework for detecting prosodic boundaries (Mandarin)

In English TTS, we only have two levels, the word and intonation phrase. After these ME models, we can get the probabilities of each syllable boundary than could be a PW, PP or IP boundary, then we use length balance to find the best choice.

The features used to predict the phrase boundary are derived from experience and different in these ME models. On the whole, the features consist of the atomic and the compound features, namely, the word description, the length of the word, the POS, the distance from the current word to the beginning of the utterance and to the end of the utterance and the combination of the forenamed atomic features. Usually, the feature window is 5 or 7.

The feature selection is simple, closely influenced by the developer's experience. However, it is not the best method to get better results. It is reported that efficient feature set can greatly improve the prediction performance on F-score [6]. The attribute sets of each feature are also important to the precision of the ME model. Too many attributes may cause the data sparseness problem, while too few attributes may decrease the discriminative ability of the model. Therefore, two primary studies on these two aspects have been done. The one is automatically select the features for each prediction task.

The other is construct a small but effective attribute sets. The latter work start with POS set. The primary experiment results confirmed with [6] and are very promising. These works are not embodied in the system submitted to Blizzard Challenge 2009, for it needs time to be more perfect and stable.

2.3. Prosodic chunk

2.3.1. The concept and perception of prosodic chunk

In a recent study, Chu et al [7] reported the variability in the rhythmic organization of a sentence. They found that the variability of higher level prosodic units is larger than that of lower level ones. They also reported that the existence of a relatively sable prosodic unit that normally cannot contain an internal break. Such stable unit, referred as prosodic chunk in this work can be grouped into prosodic phrases in various ways.

For example, we have got three records by professional speakers of the sentence “他们不知摔了多少跟头”. The prosodic organizations are as follows,

Hf2000: 他们#1不知#2摔了#1多少#1跟斗。
Hf2002: 他们#1不知#2摔了#2多少#1跟斗。
Zt2002: 他们#2不知#1摔了#3多少#1跟斗。

The #n denotes the prosodic level, the bigger the number n is, the higher level it represents. #2 and above denote the perceptible break in a sentence, while #1 denotes the foot, correspond to the prosodic word. As discussed before, prosodic chunk is the stable structure in which cannot be inserted a break. Then we got the prosodic chunk by compared the three different prosodic organizations,

Prosodic chunk: 他们#不知#摔了#多少跟斗。

With these concepts, the rhythmic organization of a sentence is then decomposed into two steps as shown in Figure 2. First, prosodic words are grouped into prosodic chunks by identifying the no-break positions. Second, prosodic chunks are grouped into prosodic phrase. The first step should be precise, especially on the recall of no-break positions, because any break inserted to a no-break location will significantly decrease the naturalness of synthesized speeches. The second step may be approximate because of the variability in the rhythmic organization.

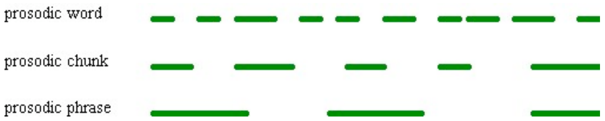

Figure 2. A hierarchical prosodic structure

This work also investigates the variability of the break allocation within Chinese sentences by perceptual experiment. The results confirm the existence of prosodic chunks [8].

Since the prosodic chunk is the stable rhythm block in a sentence, we can get various prosodic organization of a sentence by placing breaks in certain chunk boundaries. They may equally sound “natural”. Therefore, predicting the prosodic chunk boundary is the first step before predicting the prosodic phrase boundary.

2.3.2. Pre-Experiment

Predicting the prosodic chunk boundary is similar to the task that predicting the prosodic phrase boundary. To achieve a better result, we use Conditional Random Field (CRF) model, rather than ME model. The training corpus contains 9323 sentences, while 1702 sentences in the developing corpus and 900 sentences in the open test corpus.
In table 1,

w0, the current word
t0, the POS and length of the current word.
-n, the n-th word before the current word.
n, the n-th word after the current word.

Table 1. The feature template sets used in CRF model

	Word description(w)	POS and length (t)
set 1	w-2,w-1,w0,w1,w2,(w-1,w0), (w0,w1)	t-2,t-1,t0,t1,t2,(t-1,t0),(t0,t1),(t1,t2), (t-2,t-1,t0),(t-1,t0,t1),(t0,t1,t2)
set 2	w-2,w-1,w0,w1,w2,(w-1,w0), (w0,w1)	t-2,t-1,t0,t1,t2,(t-1,t0),(t0,t1),(t1,t2),
set 3	w0,w1, (w-1,w0), (w0,w1)	t-2,t-1,t0,t1,t2,(t-1,t0),(t0,t1),(t1,t2), (t-2,t-1,t0)
set 4	w0,w1, (w0,w1)	t0,t1,(t-1,t0),(t0,t1),(t1,t2)
set 5	w0,w1, (w0,w1)	t0,t1,(t0,t1)
set 6	w0,w1, (w0,w1)	t0,t1
set 7	w0,w1	t0,t1,(t0,t1)
set 8	w0,w1	t0,t1
set 9		t0,t1
set 10	w0,w1	
set 11	(w0,w1)	
set 12		(t0,t1)
set 13		(t-1,t0,t1)

Table 2. Open test results for boundary 0

	Precision%	recall %	F-score %	Feature number
set 1	89.26	83.81	86.45	1407700
set 2	89.20	84.30	86.68	1199665
set 3	89.61	84.13	86.78	1122560
set 4	89.49	84.11	86.72	569585
set 5	89.76	83.29	86.40	547730
set 6	89.13	82.14	85.50	536630
set 7	89.53	83.08	86.19	195005
set 8	88.51	81.82	85.03	183905
set 9	84.91	78.61	81.64	975
set 10	84.87	77.95	81.26	182955
set 11	69.30	68.15	68.72	352750
set 12	87.70	80.62	84.01	11125
set 13	84.15	80.15	82.10	69485

Table 3. Open test results for boundary 1

	Precision%	recall %	F-score %	Feature number
set 1	88.20	92.31	90.21	1407700
set 2	88.51	92.22	90.33	1199665
set 3	88.44	92.56	90.45	1122560
set 4	88.41	92.47	90.40	569585
set 5	87.92	92.76	90.27	547730
set 6	87.15	92.36	89.68	536630
set 7	87.77	92.59	90.12	195005
set 8	86.89	91.90	89.32	183905
set 9	84.56	89.35	86.89	975
set 10	84.17	89.40	86.71	182955
set 11	76.02	76.98	76.50	352750
set 12	86.08	91.38	88.65	11125
set 13	85.40	88.49	86.92	69485

The open test results in table 2 and 3 are very promising, which can provide a confident precipitator in prosodic structure prediction. Therefore, the prosodic chunk boundaries are predict first in the bottom-up hierarchical framework, then detect the prosodic phrase boundaries with the consideration that prosodic chunk cannot contain an internal break.

3. Wiston System Framework

Wiston system is a Text to speech system based on unit selection and waveform concatenation methods [2]. There are three main modules in wiston system, which are text analysis module, unit selection module and acoustic processing module. The text analysis is based on maximum entropy algorithm. In unit selection stage, firstly the prosodic parameters are predicted by CART based prosody trees, secondly the target cost of each candidate unit and the concatenation cost between each adjacent candidate units are computed, and finally the path which has minimum summation of target cost and concatenation cost will be found. The candidate units on this path are chosen for concatenation. In acoustic processing stage, the candidate units are concatenated and the concatenation points are smoothed. The duration, F0 and energy of each unit are then modified to the predicted values by PSOLA algorithm. The synthesized speech waveforms are finally generated.

4. Experiments

To evaluate the performance of synthesized voices with the prosodic chunk integrated, the following objective and subjective evaluation are carried out.

4.1. Experiment Conditions

The database for building Wiston system consists of 10000 phonetically balanced Chinese sentences, which are all manually labeled. The first 8000 sentences are used to train the system and for unit selection. The left 2000 sentences are used for open testing.

4.2. Objective Evaluation

For a unit-selection text to speech system, prosodic features are important factors that can affect the process of unit selection and then the naturalness of synthesized voices. Meanwhile, the prosodic features are obtained according to the results of text analysis. Therefore we will focus on how much the text analysis front with prosodic chunk can improve the accuracy of the predicted prosodic feature.

For this objective evaluation, we randomly select 100 pairs of sentences from the 2000-sentence testing database. The manually labeled syllable boundaries are used to compute the reference values of syllable duration and pause duration. The F0 mean of each syllable which is computed from waveform directly are used as reference of mean F0. Two text analysis fronts are evaluated, one with prosodic chunk and the other without prosodic chunk. The predicted syllable duration, pause duration and mean F0 according to the results of the above two text analysis front-ends will be evaluated. Table 4 shows the mean error and mean relative error between the predicted syllable duration, pause duration, F0 mean and the reference ones.

Table 4. Mean error and mean relative error between the predicted and reference prosody features

		syllable duration	pause duration	F0 mean
system A	mean error	36.65 ms	34.50 ms	20.94 Hz
	mean relative error	0.16	**0.22**	0.08
system B	mean error	36.47 ms	34.50 ms	21.45 Hz
	mean relative error	0.16	**0.20**	0.08

As we can see from table 4, the mean error and mean relative error of syllable duration and F0 mean of system A and system B is quite similar. This means the prosodic features and naturalness of each synthesized syllable of the two systems are quite similar. However, the pause duration of the two systems shows some difference. The mean error of pause duration of two systems seems equal, which means that the statistic mean pause duration of the testing cases are almost the same. However, the mean relative error of pause duration shows 2 percent decrease from system A to system B. This means the difference of pause duration between the two synthesized voices do exist. And with the prosodic chunk in system, the prediction of pause duration can achieve a certain improvement.

4.3. Subjective Evaluation

To get a direct contrast of the synthesized voices with and without prosodic chunk, the ABX test is carried out. 20 pairs of sentences are synthesized by two systems, one with prosodic chunk and the other without. Eight speech experts are asked to evaluate these sentences. That is, make a decision from each pair which one sounds more natural, or just equal. The results of this test are given in table 5.

Table 5. ABX test results (A: system without prosodic chunk in text analysis, B: system with prosodic chunk in text analysis)

Prefer A	Prefer B	Equal
21%	43%	36%

As can be seen from the table 5, the two synthesized voices sound equal in 36% cases, which means the results of text analysis of these two systems are quite equal. However, for the left 64% cases when the results of text analysis are not equal, system with prosodic chunk performs much better than without prosodic chunk. The result is very encouraging.

5. Conclusions

This paper introduces the Wiston TTS system, a large corpus based TTS system, which build for English and Mandarin. Basically, both of the two systems include three modules, the text analysis module, the unit selection module and the acoustic processing module. The prosodic structure and prosodic parameter prediction are of great importance to build a high-natural TTS system. To achieve this, we carried out some experiments on the perception and prediction of prosodic chunk which is relatively sable prosodic unit that normally cannot contain an internal break. The results are promising and we can use prosodic chunk as a constraint to predict the PP and IP phrase boundaries. We also conducted object and subject evaluation. The results show that the accuracy of pause duration prediction achieves a certain improvements, which contribute to the exciting improvements of naturalness of synthesized voices.

In the future, we will go on with the prosodic structure prediction as discussed in subsection 2.2 to achieve a more natural TTS system.

6. Acknowledgements

This work was partially supported by Hi-Tech Research and Development Program of China (Grant No.: 2006AA01Z138) and National Natural Science Foundation of China (Grand No.: 60575032).

7. References

[1] A. Hunt and A. Black, Unit selection in a concatenate speech synthesis system using a large speech database, in proc. ICASSP. 1996, pp. 373-376.

[2] Jianhua Tao, Jian yu, etc, The WISTON Text to Speech System for Blizzard 2008.

[3] Wang. M.Q. and Hirschberg. J., 1991. Predicting intonational phrasing from text. Association for Computational Linguistics 29th Annual Meeting, 285-292.

[4] Ostendorf. M. and Veilleux. N., 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. Computational Linguistics 20(1), pp. 27-54.

[5] Taylor, P. and Black, AW, 1998. Assigning phrase breaks from part-of-speech sequences. Computer Speech and Language 12, 99-117.

[6] Fengjian Li, Guoping Hu and Renhua Wang, 2004, Prosody Phrase Break Prediction Based on Maximum Entropy Model, Journal of Chinese information processing, 18(5), pp. 56-63

[7] Chu, M., Zhao, Y. and Chang, E., 2006. Modeling stylized invariance and local variability of prosody in text-to-speech synthesis. Speech Communication. Volume 48, Issue 6, pp. 716-726.

[8] Min Chu, Honghui Dong, Jianhua Tao, "A Perceptual Study on Variability in Break Allocation within Chinese Sentences", 2006 International Conference on Speech Prosody, May 2006, Germany

[9] Jian Yu, Jianhua Tao, "A Novel Prosody Adaptation Method for Mandarin Concatenation Based Text-to-Speech System", Journal of Acoustical Science and Technology, Vol. 30, No.1, January 2009, pp.33-41

[10] Fangzhou Liu, Jianhua Tao, Qing Shi, "Tree-Guided Transformation-Based Homograph Disambiguation in Mandarin TTS System", 33rd IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP2008, March 2008, Las Vegas, US, pp.2657-4660

[11] Jianhua Tao, Fangzhou Liu, Meng Zhang, Huibin Jia, "Design of Speech Corpus for Mandarin Text to Speech", The Blizzard Challenge 2008 workshop, Oct.2008

[12] Fangzhou Liu, Huibin Jia, Jianhua Tao, "A Maximum Entropy Based Hierarchical Model for Automatic Prosodic Boundary Labeling In Mandarin" , The 6th International Symposium on Chinese Spoken Language Processing, ISCSLP2008, Dec, 2008. Kunming, pp257-260

[13] Donghui Dong, Jianhua Tao, Bo Xu, “Prosodic Word Prediction using a Maximum Entropy Approach”, International Symposium on Chinese Spoken Language Processing, Lecture Notes of Computer Science, ISCSLP2006, Singapore

[14] Jian Yu, Jianhua Tao, "The Pause Duration Prediction for Mandarin Text-to-Speech System" , 2005 IEEE International Conference on Natural Language Processing and Knowledge ngineering (IEEE NLP-KE 2005),Wuhan,China,pp.204-208,2005, LSBN:0-7803-9361-9

[15] Honghui Dong, Jianhua Tao, Bo Xu，"Chinese Prosodic Phrasing with a Constraint-based Approach"，INTERSPEECH 2005-EUROSPEECH，Lisbon, Portugal

[16] Honghui Dong, Jianhua Tao,Bo Xu，"Prosodic Word Prediction Using the Lexical Information"，2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering，Wuhan, China,pp.189-193,2005,LSBN:0-7803-9361-9

[17] Honghui Dong, Jianhua Tao, "Length Optimized Chinese Prosodic Phrasing Model" , Proceedings of the International Conference on Chinese Computing 2005, pp.48-53,2005, Singapore

[18] Jianhua Tao, "Acoustic and Linguistic Information Based Chinese Prosodic Boundary Labelling" , Lecture Notes of Artificial Intelligence, Springer, 2004,9

[19] Honghui Dong, Jianhua Tao , Bo Xu, "Grapheme-to-Phoneme Conversion in Chinese TTS System",ISCSLP2004,pages:165-168

[20] Jian Yu, Wanzhi Zhang and Jianhua Tao, "A new pitch generation model based on internal dependence of pitch contour for mandarin TTS system" , ICASSP 2006, Toulouse, France