The WISTON Text to Speech System for Blizzard 2008

Jianhua Tao, Jian Yu, Lixing Huang, Fangzhou Liu, Huibin Jia, Meng Zhang

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

{jhtao, jyu, lxhuang, fzliu, hbjia, mzhang}@nlpr.ia.ac.cn

Abstract

The WISTON system is a large corpus based TTS system with the unit selection method. The text analysis part of this system contains text pre-processing, word segmentation, POS tagging, phonetic transcription and prosody structure prediction. The prosody information (duration, F0, energy) is predicted by the CART model with the input context information. In the unit selection model, we use the mutual prosody constraint as the part of concatenation costs for the path searching while the predicted F0s, durations and energies are used to get the target costs. The spectrum smoothing method is also used for the speech generation. The final system was used to attend Blizzard evaluation for both English test and Mandarin test. Good scores were got based on this system.

Keywords: WISTON, speech synthesis, mutual prosody constraint

1. Introduction

The WISTON TTS system has been developed for a long time. The framework of the system is originally designed for the multilingual speech synthesis with the good component structure which includes two main modules: text processing module and unit selection modules. Text processing is designed to do the following things: text pre-processing, word segmentation, POS tagging, phonetic transcription and prosody structure prediction. To reduce the language specificity, the maximum entropy (ME) method is used for most of text analysis work.

The unit selection module contains many criterions which are used for the selection of speech unit. Due to the mutual prosody constraint between adjacent speech units, we introduce a prosody concatenation model which works together with the spectrum concatenation constraints. The prosody concatenation model helps the system to remove most of the unnatural and unstable prosody parts in the synthesized speech. Finally the natural, fluent synthesized speech can be generated in the WISTON system.

The rest of this paper is organized as follows: section 2 introduces the text processing module. Section 3 introduces the unit selection module, especially the prosody model based on prosodic mutual constraint. In section 4, the paper will introduce and analyze the evaluation results of our system in Blizzard Challenge 2008. The section 5 gives the final conclusion of our work.

2. Maximum Entropy based Text Analysis Module

The Figure 1 shows the framework of the text analysis module of the WISTON system. In the module, the pre-processing part performs the tokenization, digital processing, control symbol processing, markup language processing, etc..

The phonetic transcription information is processed for each word. For English speech, several letter to sound rules are used for the phonetic transcription of unknown words. For Mandarin speech, as all Chinese characters have been labeled with pronunciation (Pinyin) in the lexicon, our work is aiming to solve the homograph disambiguation problems. The ME algorithm is used for Mandarin homograph disambiguation.

Unlike English, the word boundaries are not clear in Chinese sentences. For Mandarin speech, we combine the word boundary detection, unknown word prediction, name entity and address prediction together with the POS tagging by a single ME algorithm, and get a good accuracy for them. For English speech, there is no word segmentation part but the POS tagging is also predicted by the ME algorithm.

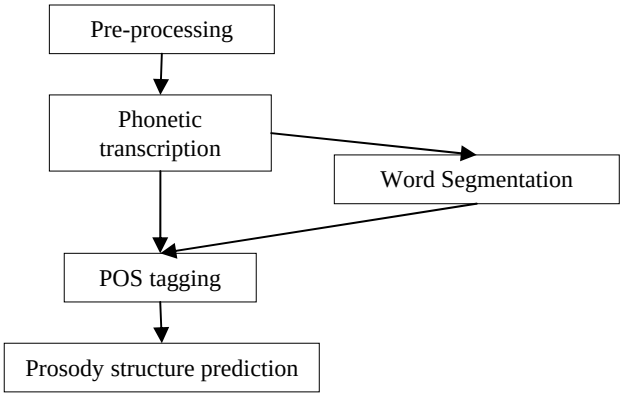

Figure 1: The framework of text analysis of the WISTON system

Breaking sentences into suitable rhythmic units is important to achieve high naturalness in speech synthesis. Various machine learning algorithms have been employed to predict the most likely positions for breaks in a text stream [12-14]. In our work, we separate the prosody structures into four levels: syllable, word (prosody word for Mandarin), minor prosody phrase, major prosody phrase.

The Figure 2 shows a sample of the prosody structure for Mandarin speech. In the figure, the PW means prosody word, NP is the minor prosody phrase and MP denotes the major prosody phrase.

For both English speech and Mandarin speech, the phrase boundaries are predicted by the following steps.

Step 1: we use the maximum entropy (ME) method based minor/major phrasing model to predict phrase boundaries. The ME model is good at classification problem.

Step 2: we try to find the stable blocks in which the minor or major phrase boundaries cannot be inserted. We use these results to reduce the predicted minor/major phrase boundaries.

Step 3: we use the length balance to find the best major phrase breaking points. The prosodic structure is then predicted by

$$P(J_i = status \mid POS_i, POS_{i+1}) \cdot P(J_i = status \mid nLen_i, nLen_{i+1}) \quad (1)$$

Where, POS and nLen denote the part of speech and the length of the major phrase respectively. Subscripts i and i+1 represent both sides of the boundary to be predicted. The major phrase breaks are predicted using decisions based on the POS and a N-gram model of major phrase lengths. Additional length constraints are introduced to enable search within a limited length.

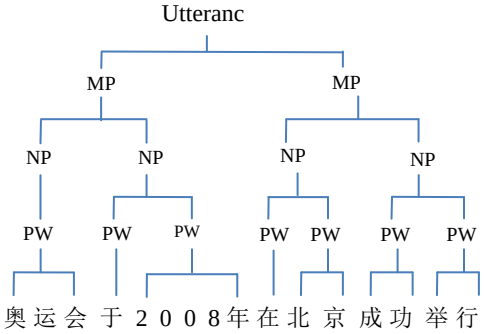

Figure 2: A sample of prosody structure for Mandarin speech.

3. Unit Selection Module

3.1 Unit Pre-Selection

In WISTON system, the basic unit for English part is diphone while the Mandarin part is using syllables.

In the unit selection procedure, there are so many candidate units that selecting the best path from all candidates is a very time-consuming task. In our work, we use a pre-selection procedure. The scale for this pre-selection procedure is contextual information difference (CID), which depicts the difference of contextual information between the candidate unit and the synthesized unit. The contextual information includes the location of the current speech unit in word, phrase and sentence, the phoneme string of the current speech unit, the length of the word, phrase and sentence, the boundary types before and after the current speech unit, etc.

Suppose that the number of contextual information category is n, the formula is as follows:

$$CID = \sum_{i=1}^{n} W_i * D_i \quad (2)$$

Where D_i is the difference of the i th contextual information between the current candidate unit and the predicted target unit, and W_i is the weight of the i th contextual information.

In real application, the candidate unit whose contextual information is most similar with the predicted target unit is not always the most appropriate unit. However, because this procedure is just a pre-selection procedure, we don't expect its results to be very precise. It is satisfying as long as one or more appropriate units could be included in the pre-selection results, which can be achieved by properly setting the number of pre-selection results even with the unreasonable definition of CID.

3.2 Target Cost

The target cost is defined as the difference between the predicted prosody parameters and the real parameters of the candidate units. In our work, we use $F0_M, F0_B$ and $F0_T$, which denote the pitch register and the pitch range. The difference between predicted results and real parameters are $DF0_M, DF0_B$ and $DF0_T$. Then, the target cost is:

$$target\_cost = w_1 * DF0_M + w_2 * DF0_B + w_3 * DF0_T \quad (3)$$

Where w_1~w_3 are the weights.

In our work, a CART model is trained to predict $F0_M, F0_B$ and $F0_T$ by inputting the context information (see definition in 3.1) from the text analysis module. The target cost is presented to depict the overall trend of the pitch contour. For example, pitch declination in naturally read discourses can be realized by all of the three parameters descend as the sentence approaches the end. Minimizing the target cost can make the output pitch contour similar with the natural one on overall trend.

3.3 Concatenation Cost

The concatenation cost which includes spectrum concatenation cost and prosody concatenation cost is trying to make spectrum and prosody smoothing for the synthesized speech. The final concatenation cost will be the sum of the spectrum concatenation cost and the prosody concatenation cost.

For the spectrum concatenation cost, we are simply using spectrum deviation between two speech units.

$$Spectrum\_concatenation\_cost = D(S_n, S_{n+1}) \quad (4)$$

Here, S_n and S_{n+1} denote the spectrum of the both boundaries between speech unit n and speech unit n+1.

To get a good prosody smoothing result, we use seven parameters, shown in Figure 3, to denote the pitch contour of the speech unit. Among these, $F0_M, F0_B$ and $F0_T$ denote the pitch register and the pitch range, which reflect the overall trend of pitch contours. While $F0_S, F0_E, F0_{SD}$ and $F0_{ED}$ can be considered as boundary features, which can be used to measure the naturalness of pitch contours in local area.

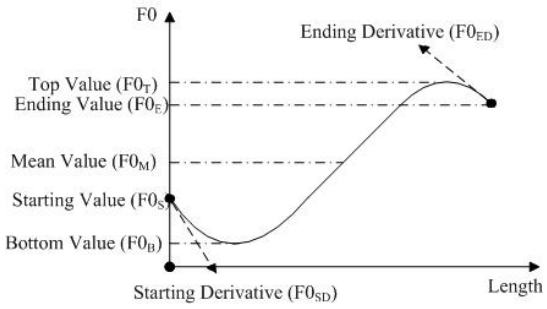

Figure 3: Seven parameters used in the pitch contour parameterization method

Prosodic parameters involved in the concatenation cost include $F0_S, F0_E$, $F0_{SD}$ and $F0_{ED}$, which can be considered as boundary features of a unit's F0 contour. The simplest definition is the difference between $F0_E$ of the previous speech unit and $F0_S$ of the next speech unit, which is based on the assumption that the pitch contour is always continuous on the whole sentence. However, there are always some silences or voiceless phonemes inside the speech, the intonation will be separated into a few isolated parts. But even in this situation, most of the F0 contours are virtually connected across the silences and voiceless phonemes. For example, the pitch end of the previous phoneme tends to stretch out across the span of the silence, reaching the pitch head of the next phoneme. Similarly, the following phoneme's F0 contour also has some impacts on the later portion of the previous unit [8].

In Table 1, we have listed all possible factors which may influence the prosody concatenation between two speech units. Then a Classification and Regression Tree (CART) is used to train a predicting model for the prosody concatenation parameters, $F0_S, F0_E, F0_{SD}$ and $F0_{ED}$.

Table 1: Features used in predicting $F0_S, F0_E, F0_{SD}$ and $F0_{ED}$

Features in predicting $F0_S$ and $F0_{SD}$	Features in predicting $F0_E$ and $F0_{ED}$
Frequently used text information	Frequently used text information
$F0_E$ and $F0_{ED}$ of the previous speech unit	$F0_S$ and $F0_{SD}$ of the following speech unit
Pause duration before current speech unit	Pause duration after the current speech unit
End phoneme of the previous speech unit	Start phoneme of the following speech unit
Start phonme of the current speech unit	End phoneme of the current speech unit

The difference between these predicted values with the CART model and real values can be used to measure the naturalness of pitch contours between two units. The prosody concatenation cost is then defined as (5), in which the cost is the weighted sum of differences between predicted values and real values of four prosodic parameters mentioned above

$$prosody\_concatenation\_cost = w_4 * DF0_S + w_5 * DF0_E + w_6 * DF0_{SD} + w_7 * DF0_{ED} \quad (5)$$

Figure 4 schematically illustrates the procedure to get prosody concatenation cost for Mandarin speech where the basic speech unit is the syllable.

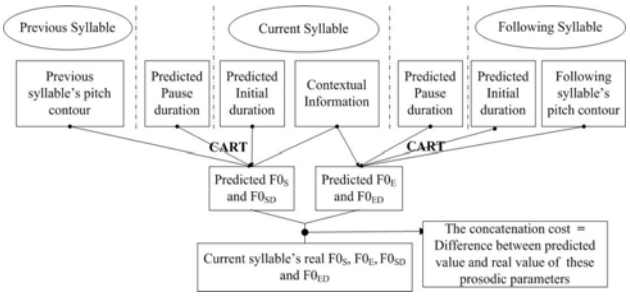

Figure 4: The procedure of the prosody concatenation cost for Mandarin speech

3.4 Best Unit Series Selection

All in all, our cost definition is comprised by two parts: the concatenation cost and the target cost. The formula is as follows:

$$COST = w_8 * target\_cost + w_9 * prosody\_concatenation\_cost + w_{10} * spectrum\_concatenation\_cost \quad (6)$$

The weights are not assigned equally. For instance, the weights related to $F0_S$ and $F0_{SD}$ are normally higher than that for $F0_S$ and $F0_{SD}$. Based on the cost definition in (6), a Viterbi search algorithm will be used to find the best path with the minimum cost. The final unit selection results will be found from this path.

4. System Building for Blizzard 2008

4.1. English and Mandarin Speech database

The English speech database is the Roger UK English Speech Corpus for Blizzard Challenge 2008 [9]. It contains 1132 arctic script utterance, 1390 dialogue-rich utterances, 2880 isolated words, 45 addresses, 1681 sentences with emphasized words, 2449 news items, etc.. In total it has 9509 utterances, around 15 hours. The database includes wave files and scripts with the labeling of phrase boundaries and POS tagging. The English lexicon is also supplied.

The Mandarin Speech database for the Blizzard 2008 is supplied by CASIA [10]. It contains 4500 utterances with the neutral reading style, around 8.5 hours. The database includes wave files, phone list and scripts with word boundaries and POS tagging.

4.2. Building Systems

The English database was separated into two parts, Voice A and Voice B. Voice A is from the full dataset while Voice B is from the arctic subset. The English systems were built on Voice A and Voice B separately.

All English and Mandarin database were labeled with pitch marks at the beginning. Then, the HTK tools are used to segment the speech into phoneme level. The Unilex and some letter to sound rules were used for the text to phoneme conversion in the English parts. As the typical phonetic and tonal structures for Mandarin syllables, we further used pitch contours, energy and zero-across rate contours to improve the segmentation accuracy of the syllable boundaries.

The mandarin database was further labeled with the phrase boundary. The method is described in [11]. Then, the ME method was used for the training of phrasing model for both the English and the Mandarin parts.

4.3. Evaluation results

20 participants attend the evaluation for Voice A, 18 participants for Voice B, and 12 participants for Mandarin voice. The evaluation results from all listeners are given in Table 2, 3 and 4.

Table 2: the results for Voice A from all listeners

	Results
Similarity to original speaker	3.0
Mean of Score	2.8
Word error rates for SUS test	0.40

For the Voice A, the highest similarity to original speaker is 3.3 which was scored by the EUL (paid UK students), the highest MOS is 2.8 scored by the EI (paid Indian students) and the ES (speech experts).

Table 3: the results for Voice B from all listeners

	Results
Similarity to original speaker	3.1
Mean of Score	2.8
Word error rates for SUS test	0.40

For the Voice B, the highest similarity is 3.3 scored by the ES (speech experts) and the highest MOS is 3.0 scored by the EI (paid Indian students) and the EUL (paid UK students)

Table 4: the results for Mandarin Voice from all listeners

	Results
Similarity to original speaker	3.2
Mean of Score	3.6
Word error rates for SUS test	0.20

For the Mandarin speech, the highest similarity is 3.4 scored by both the MR (volunteers) and the MS (speech experts), the highest MOS is 3.6 scored by both the MC (paid participants in China) and the MS (speech experts).

The testing results of the Voice B looks like the same as the results of the Voice A, but actually, the ranks of the Voice A are significantly higher than the ranks of Voice B. They are improved from 9 to 5 for voice similarity, from 14 to 9 for MOS and from 12 to 8 for word error rates, separately. The reason is that the Voice A is more expressive than Voice B and our WISTON system is more suitable for neutral speech processing.

Among all tests, the results of Mandarin Voice are the best scores we've got. They are also better than most of others. The reason is very simple, because our system was mostly tested for Mandarin speech, although it has been designed for multilingual processing. Due to the limitation of English training data, the phrase boundary prediction and spectrum smoothing for English speech have not been fully tested before the Blizzard challenge 2008. We will try to improved them in our future work..

5. Conclusion

This paper introduces the construction of WISTON TTS system, which is designed for the synthesis of both English and Mandarin speech. For Mandarin speech, we have used the mutual constraint of the prosody between speech units for unit selection module. The spectrum smoothing is also used for the concatenation cost for both Mandarin speech and English speech. The ME algorithm is used for the phrasing model for the text analysis part of the system. The evaluation experiments have proved that the Mandarin speech synthesis part of our TTS system has achieved a high score compared with other systems, however the English part still need to be improved in our future work.

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. Reference

[1]. Jian Yu, Wanzhi Zhang, and Jianhua Tao. "A New Pitch Generation Model Based on Internal Dependence of Pitch Contour for Mandarin TTS System", in ICASSP. 2006. Toulouse, France.

[2]. Jian Yu, Jianhua Tao, and Xia Wang. "Pitch Prediction for Mandarin TTS with Mutual Prosodic Constraint", in ISCSLP. 2006. Singapore.

[3]. Black, A.W. and P.A. Taylor. CHATR: a generic speech synthesis system. in COLING. 1994.

[4]. Christophe Blouin, O.R., Paul C. Bagshaw and Christophe. Concatenation cost calculation and optimisation for unit selection in TTS. in Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop. 2002.

[5]. Jithendra Vepa, S.K.a.P.T. New Objective Distance Messures for Spectral Discontinuties in Concatenative Speech Synthesis. in Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop. 2002.

[6]. Tomoki Toda, H.K., Minoru Tsuzaki and Kiyohiro Shikano. Perceptual Evaluation of Cost for Segment Selection in Concatinative Speech Synthesis. in Proc. IEEE 2002 Workshop on Speech Synthesis. 2002.

[7]. Fu-Chiang Chou, C.-Y.T., and Lin-Shan Lee, A Set of Corpus-Based Text-to-Speech Synthesis Technologies for Mandarin Chinese. IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING , VOL.10 , NO.7 , OCTOBER 2002, 2002.

[8]. Greg Kochanski, Chilin Shih, "Prosody Modeling with Soft Templates", Speech Communication, 2003. 39.

[9]. Roger UK English database: http://www.cstr.ed.ac.uk/projects/roger_blizzard2008/

[10]. CASIA Mandarin database: http://www.speakit.cn/corpus/license.html

[11]. Jianhua Tao, "Acoustic and Linguistic information Based Chinese Prosodic Boundary Labelling", TAL2004

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

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

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