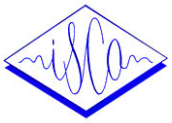
ISCA Archive
http://www.isca-speech.org/archive



International Symposium on Chinese Spoken Language Processing (ISCSLP 2006)
Kent Ridge, Singapore
December 13-16, 2006

Prosodic Word Grouping in Mandarin TTS System

Qing Guo[1], Endong Xun[2], and Nobuyuki Katae[3]

[1] Fujitsu Research and Develop Center China, Beijing
[2] Beijing Language and Culture University, Beijing
[3] Fujitsu Laboratories Ltd., Japan

guoqing@cn.fujitsu.com

Abstract. This paper reports the methodology and results of prosodic word grouping for a Mandarin TTS system developed by the Fujitsu Laboratories. In view of any inner prosodic word break will make speech unintelligible or unnatural, a new prosodic word grouping framework is proposed. The word segmentation result can be regarded as an initial prosodic word sequence with grids inserted into each word boundary. The target of prosodic word grouping is to remove those grids that are assessed needlessly in the prosodic word level. Several prosodic word grouping methods can work together because each makes its own decision. As one of three methods used in our TTS system, a binary prosodic tree method is also proposed in this paper. Lastly, experiment results and discussion are presented.

Keywords: prosodic word, prosodic word grouping, binary prosodic tree.

1 Introduction

Rhythm is an important factor that makes the synthesized speech of a TTS system more natural and understandable. Researchers have found that there is a hierarchical prosodic structure for Chinese prosody, which constitutes the rhythm of Chinese speech[1][2]. The boundaries of prosodic units can be identified by pauses, pitch changes, or duration changes of boundary syllables in the speech. In a TTS system, the prosodic structure provides important information for the prosody generation model to realize all these affects in the synthesized speech.

There are many reports specifying various hierarchical structures for prosodic constituents. Generally, the main prosodic constituents in Chinese speech are prosodic word, prosodic phrase and intonation phrase. Prosodic word is a group of syllables uttered continuously and closely without breaks in the speech. Prosodic word is the lowest constituent in the prosodic hierarchy and should have a perceivable prosodic boundary. In other words, no prosodic boundary can be perceived within a prosodic word and pauses should happen only on the boundaries between prosodic words. Thus, good prosodic word grouping plays an important role in increasing the naturalness of synthesized speech.

Many studies reveal that the prosodic word is quite different from the lexical word. One of the reasons is that the formation of prosodic words is not only based on the

meaning of words but also based on rhythm requirement of speech. A prosodic word can contain more than one lexical word and it can also be a part of a relatively long lexical word. The perceptive experiments in [3] shows that the TTS system using the prosodic word as the basic unit has a much higher intelligibility and naturalness than using the lexical word directly.

In recent years, many methods of prosodic word boundary prediction have been proposed in Chinese, such as the classification and regression tree (CART) method[3][4], rule-driven approach[5], statistical approach[6][7][8][9] and recurrent neural network (RNN) method[10]. In these works, the POS and word length information are mostly used.

In view of any inner prosodic word break will make speech unintelligible or unnatural, a new prosodic word grouping framework is proposed in this paper. The word segmentation result can be regarded as an initial prosodic word sequence with grids inserted into each word boundary. Hence, the target of prosodic word grouping is to remove the grids that are assessed needlessly in the prosodic word level. Several prosodic word grouping methods can work together since each of them generates its own decision. In other words, a grid can be removed by any one of these methods.

Three word grouping methods have been considered in this paper. First, a binary prosodic tree method is proposed. Punctuation marks in the reading text can provide pause guidance in speech. A novel means using punctuation marks to train a language model that possesses speech break information is proposed. This language model is used to generate the binary prosodic tree of an input sentence. Secondly, a statistical probability method proposed in [6] is adopted. Thirdly, a rule based method for frequently used prosodic cohesive words is considered.

This paper is organized as follows. Section 2 describes our prosodic word grouping framework and the three prosodic word grouping methods. Section 3 provides the experiment results and discussion. Finally, we conclude this paper in Section 4.

2 Prosodic Word Grouping

Prosodic word is the lowest constituent in the prosodic hierarchy and should have a perceivable prosodic boundary. In other words, no prosodic boundary can be perceived within a prosodic word and pauses should happen only on the boundaries between prosodic words. Thus, good prosodic word grouping plays an important role in increasing the naturalness of synthesized speech.

In real speech, not all boundaries of prosodic words have breaks. This is tolerable if there is a break at the end of each prosodic word. However, any inner prosodic word break will make speech sound unintelligible or unnatural.

For example, “蹭蹭蹭就爬上了山顶” is separated into “蹭蹭蹭 就 爬 上 了 山顶” by a word segmentation tool in a Mandarin TTS system. Lexical words “就”, “爬”, “上” and “了” are all monosyllables. They should be grouped together into a prosodic word, “就爬上了”. If they are isolated in the prosodic word level, the synthesized speech sounds very unnatural in the part of “就 爬 上 了”. These four words sound like they are uttered one by one and people can perceive uncomfortable breaks in the speech. This is because both pitch prediction and duration prediction in

the prosody generation model are sensitive to a prosodic word boundary or a word boundary. In case that “就爬上了” is a prosodic word, its pitch contour will sound natural since the pitch prediction model can consider more the co-articulation. In addition, the duration prediction model does not lengthen durations of syllables “就”, “爬” and “上”. These 3 syllables are not in the prosodic word boundary when they are within a prosodic word.

In view of any inner prosodic word break will make speech sound unintelligible or unnatural, a new prosodic word grouping framework is proposed in this paper.

Initially, the input text is processed by a word segmentation and pos-tagging module. This word segmentation result can be regarded as the initial prosodic word sequence with grids inserted into each word boundary. The target of prosody word grouping is to remove grids assessed needlessly in the prosodic word level. Three prosodic word grouping methods, namely binary prosodic tree method, statistical probability method and a rule based method can work together because each makes its own decision. A grid might be removed by any one of these three methods.

In this section, each of these three methods will be discussed.

2.1 Binary Prosodic Tree

Punctuation Marks vs. Speech Break. Punctuation marks play an important role in the writing text of a given language. The main function of modern English punctuation, however, is logical: it is used to make clear the grammatical structure of the sentence, linking or separating groups of ideas and distinguishing what is important in the sentence from what is subordinate. However, what has been isolated by punctuation marks might not be an integrated syntactic component in Chinese. One important role of punctuation marks in Chinese is to imply speech break. In a long Chinese sentence without punctuation marks, to decide where a speech break should be inserted, we just need to predict the most likely position where the punctuation mark is inserted. Therefore, a statistical language model is used to estimate the probability of inserting a punctuation mark at a given word boundary in the sentence.

Language Model Training. Generally speaking, a large scale training corpus is needed to train a language model. It is quite difficult to obtain a large scale training corpus with the prosody structure annotated. Here we use a novel method to avoid the lack of training corpus. First, a language model is trained from a raw corpus by using the connotative relationship between punctuation marks and speech breaks. Second, another language model is trained from a corpus with the prosody structure annotated. Third, these two language models are linear combined together.

A sentence with words segmented can be represented by a word sequence as follows:

$$W = w_1 w_2 \cdots w_n$$

The probability of the word sequence can be decomposed as a linear combination of a trigram language model and a bi-gram model:

$$P(W) = \lambda \times P_G(W) + (1-\lambda) \times P_D(W) \tag{1}$$

Here, $P_G(W)$ is a trigram model, which is trained from a raw corpus by replacing each punctuation mark with a symbol Δ that is used to represent speech breaks in the speech.

$$P_G(W) = P(w_1) \times P(w_2 \mid w_1) \times \prod_{i=3}^{n} P(w_i \mid w_{i-2}, w_{i-1}) \tag{2}$$

$P_D(W)$ is a bi-gram model, which is trained from an annotated corpus by inserting symbol Δ in all the prosodic word boundaries and those boundaries beyond prosodic word boundary.

$$P_D(W) = P'(w_1) \times \prod_{i=2}^{n} P'(w_i \mid w_{i-1}) \tag{3}$$

Binary Prosodic Tree Building. Binary prosodic tree is proposed in this paper and is built for an input sentence by the language model discussed above. The binary prosodic tree is used to group the prosodic words.

In a word sequence $W = w_1 w_2 \cdots w_n$ that is derived from the word segmentation module, a potential break exists in the place between any word pairs $w_i w_{i+1}$. In fact, there are $n-1$ potential break candidates in a sentence with n words.

By inserting a break symbol Δ into each potential break candidate, a new sentence $W_i^{'} = w_1 w_2 \cdots w_i \Delta w_{i+1} \cdots w_n$ is obtained. By employing equation (1) to calculate each $P(W_i^{'})$, we can find the most likely break candidate, namely, $\arg\max_i P(W_i^{'})$, $i \in [1, n-1]$.

The growing algorithm of the binary prosodic tree can be described as follows:

Step 1: For each potential break candidate position i, $i \in [1, n-1]$, compute the probability of a new sentence $W_i^{'} = w_1 w_2 \cdots w_i \Delta w_{i+1} \cdots w_n$ with a break symbol Δ is inserted in position i.

Step 2: Find the most likely break inserting position $j = \arg\max_i P(W_i^{'})$, $i \in [1, n-1]$.

Step 3: Split the sentence to 2 sub-sentences by j, i.e.,

$leftsent = w_1 w_2 \cdots w_j$

$rightsent = w_{j+1} w_{j+2} \cdots w_n$

Step 4: Grow the prosodic binary tree for left sub-sentence $w_1 w_2 \cdots w_j$

Step 5: Grow the prosodic binary tree for right sub-sentence $w_{j+1}w_{j+2}\cdots w_n$

Step 6: If the number of words in current sentence is 1, stop.

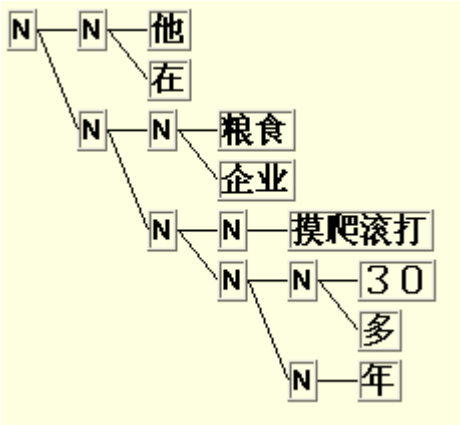


Fig. 1. The binary prosodic tree of Sentence 1.

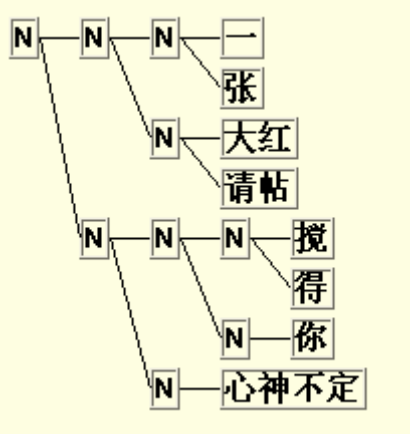


Fig. 2. The binary prosodic tree of Sentence 2.

Figure 1 and Figure 2 present two exemplified binary prosodic trees of the following two segmented sentences (with grids being added into all lexical word boundaries) by our binary prosodic tree building algorithm.

Sentence 1: 他 | 在 | 粮食 | 企业 | 摸爬滚打 | ３０ | 多 | 年

Sentence 2: 一 | 张 | 大红 | 请帖 | 搅 | 得 | 你 | 心神不定

We named the binary tree obtained from the above algorithm as a binary prosodic tree since it is shown that the structure of the binary tree of a sentence has a good consistency with the prosody structure of that sentence. We also applied the binary tree as a pruning strategy in prosodic phrase boundary prediction based on the

assumption that a prosodic phrase of a sentence should be corresponding to a sub-tree of the binary tree of that sentence[11].

According to the binary prosodic tree in Figure 2, those grids that exist between every 2 leaf nodes with the same father node will be removed. Using the binary prosodic tree word grouping method, the prosodic word grouping result of Sentence 2 is as follows,

Sentence 2: 一张 | 大红请帖 | 搅得 | 你 | 心神不定

2.2 Statistic Probability Model

A statistic probability model proposed in [6] is used in our prosodic word grouping framework as well.

Part-of-speech (POS) and word length information are used to predict prosodic word boundary. By assuming that the POS is independent from the word length information when predicting the prosodic words, the probability of two adjacent lexical words being grouped together into a prosodic word can be estimated using the following formula.

$$\begin{aligned} & P(J_0 \mid POS_0, POS_1, wLen_0, wLen_1) \\ &= P(POS_0, POS_1, wLen_0, wLen_1 \mid J_0) \cdot P(J_0) / P(POS_0, POS_1, wLen_0, wLen_1) \\ &\approx P(J_0) \cdot P(POS_0, POS_1 \mid J_0) \cdot P(wLen_0, wLen_1 \mid J_0) / P(POS_0, POS_1, wLen_0, wLen_1) \\ &= (1/P(J_0)) \cdot P(J_0 \mid POS_0, POS_1) \cdot P(J_0 \mid wLen_0, wLen_1) \end{aligned} \quad (4)$$

In equation (4), $P(J_0)$ is a constant value. Thus, the following formula is used to compute the probability that the two adjacent lexical words are bundled into a prosodic word.

$$P(J_0 = 0 \mid POS_0, POS_1) \cdot P(J_0 = 0 \mid wLen_0, wLen_1) > \eta \quad (5)$$

If the probability is larger than η, it is recognized that these two words should be combined into a prosodic word.

The Maximum Likelihood Estimate is used when training the probability $P(J_0 = 0 \mid POS_0, POS_1)$ and $P(J_0 = 0 \mid wLen_0, wLen_1)$.

2.3 Prosodic Word Grouping Rules

Lastly, prosodic word grouping rules are carefully designed for frequently used prosodic cohesive words, such as suffix morpheme, structure auxiliary words, location words and verb tendency phrases. Often these words have a fixed prosodic word grouping manner under some constraints. If these words are not grouped into their appropriate prosodic words, the synthesized speech sounds very unnatural. There are many approaches for a verb to form its reduplicative compound in Chinese. For

example, “V 一 V”, “V 了 V” and “V 了一 V” (“谈一谈”, “想了想”, “读了一读”). After word segmentation, there are grids within these phrases, for example, “谈 | 一 | 谈”. However, in the natural prosodic level, these reduplicative compounds are treated as a whole prosodic individual word. Therefore, rules are designed for these verb reduplicative compounds in order to deal with the grids in the phrases.

3 Experiment Results and Discussion

3.1 Corpus

The training corpus of the trigram language model used for binary prosodic tree building was the Chinese People Daily 1998 Corpus, which is transcribed from a Chinese newspaper, with word segmentation and POS-tag annotated for natural language processing purpose.

CLDC-SPC-2005-011(http://www.chineseldc.org) is a large phonetically and prosodically enriched Mandarin speech corpus, of which the prosody structure is labeled by a well-trained annotator. Usually sentences with 7-25 syllables are selected in this corpus. It contains a total of 6084 utterances. The statistical probability model used in the paper was trained from this corpus.

An independent test corpus was used in this paper’s experiments. There were 400 sentences in the test set with an average number of Chinese characters per sentence at about 37 and the average number of lexical words in a sentence was about 23. These figures are more consistent with the actual cases. The prosody structure was labeled by a well-trained annotator from the text and then modified by listening to the speech corpus recorded by a female graduate student majoring in Chinese literature.

3.2 Results and Discussion

Precision and recall statistics were calculated to evaluate the performance of prosodic word grouping in this paper.

5101 prosodic word boundaries were annotated in the test set. 4713 prosodic word boundaries were predicted by our prosodic word grouping module in this framework. 4049 of them were consistent with the human annotated results. Thus, the overall precision rate was 85.91% and the recall rate was 79.38%.

Table 1 describes the result obtained from those correctly grouped prosodic words, which are composed of only two lexical words. In this table, +, * and – are used to represent the binary prosodic tree method, the statistical probability method and the rule based method respectively. Grouping methods used “+*” means that a prosodic word is grouped by both the binary prosodic tree method and the statistical probability method. Grouping methods used “+” means that a prosodic word is grouped only by the binary prosodic tree method. Other grouping methods used are similarly defined.

Table 1. The result of prosodic words that are composed of two lexical words.

Grouping methods used	Number of correctly grouped prosodic words	Percentage
+	758	46.7%
+*	664	40.9%
+*-	142	8.7%
+-	1	0.06%
*	51	3.1%
*-	8	0.5%
-	None	0%
Total correctly grouping number	1624	100%

From this table we can see that the binary prosodic tree method is most effective. It contributes 96.4% to the grouping. The statistical probability method can be complementary to the binary prosodic tree method. It further contributes 3.6% to the grouping. The rule based method is the weakest one.

There are two remarks for the above analysis. First, the analysis is only conducted for prosodic words consisting of two lexical words. For prosodic words consisting of more than two lexical words, it's difficult to make such analysis. For example, in prosodic word "走(zou3)/v +* 出(chu1)/v *- 了(le5)/u", we cannot count the exact contribution that each method makes. Second, the analysis shows that the statistical probability method demonstrates less contribution in grouping two lexical words. It will be helpful when more lexical words need to be grouped, since the binary prosodic tree method can only combine two consecutive lexical words. The rule based method is also important in that it can assure the grouping of prosodic words with certain fixed pattern.

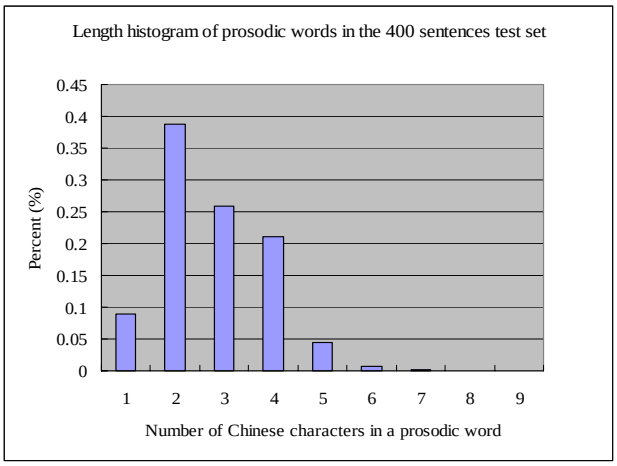


Fig. 3. Length histogram of prosodic words in the test set.

Figure 3 shows the length histogram of prosodic words in the test set of 400 sentences. The average number of Chinese characters in a prosodic word in the test set was about 2.8. Our data are quite different from that in [3][4]. They assumed that a prosodic word is primarily composed of disyllable or tri-syllable. However, many relative long units, in particular, those units with two lexical words such as “宽阔明亮” and “筹措资金”, were annotated as prosodic words by listening perception in our test set. The following is a transcribed sentence in our test set:

400. 他们/r | 着意/d 揣摩/v | 专家/n 意图/n ，/w ||| 反复/d 征求/v | 专家/n 意见/n ，/w ||| 因为/c ‖ 只有/c 专家/n | 满意/v 了/y ，/w ||| 作品/n ‖ 才/c 有/v 希望/n | 获奖/v 。/w@

The prosodic word grouping module in this paper needs to resolve these kinds of prosodic word groupings also. An annotator is asked to check all of the wrong predicted prosodic boundaries. About half of them belong to prosodic word boundary insertion errors that occur within those relative long prosodic words, for example, “宽阔 | 明亮” and “这位 | 导演”. Although perception experiments show that it is better to group these kinds of prosodic words, not grouping is also acceptable in synthesized speech. In fact, sometimes people also regard them as two different prosodic words in the speech for emphasis purposes or for poem style.

On the other hand, our strategy for removing grids may form a long predicted prosodic word that is actually composed of two or more prosodic words. This decreased the recall of the proposed prosodic word framework. However, this approach can significantly reduce the number of insertion errors; these errors make speech sound unintelligible and unnatural.

4 Conclusions

In this paper, a new prosodic word grouping framework with a grid removing strategy by several prosodic word grouping methods working together was proposed. The intention of the grid removing strategy was to decrease the number of prosodic word boundary insertion errors since any inner prosodic word break will make speech sound unintelligible or unnatural. The word segmentation result was regarded as an initial prosodic word sequence with grids inserted into each word boundary. Several prosodic word grouping methods working together to remove grids assessed needlessly in the prosodic word level were used. As one of three methods used in our prosodic word grouping module, a binary prosodic tree method was also proposed in this paper. Lastly, experiment results and discussion were presented.

References

1. Shih, C. L.: The Prosodic Domain of Tone Sandhi in Mandarin Chinese. PhD Dissertation, UC San Diego (1986)
2. Li, A. and Lin, M.: Speech corpus of Chinese discourse and the phonetic research. International Conference on Spoken Language Processing (2000)

3. Qian Y., Chu, M. and Peng, H.: Segmenting unrestricted Chinese text into prosodic words instead of lexicon words. International Conference on Acoustic, Speech and Signal Processing, Salt Lake City (2001)
4. Chu M. and Qian Y.: Locating boundaries for prosodic constituents in unrestricted Mandarin texts. Journal of Computational Linguistics and Chinese Language Processing, 6(1), (2001) 61-82
5. Tao, J., Dong, H. and Zhao, S.: Rule learning based Chinese prosodic phrase prediction. International Conference on Natural Language Processing and Knowledge Engineering, Beijing (2003)
6. Dong H., Tao J. and Xu b.: Prosodic word prediction using the lexical information. International Conference on Natural Language Processing and Knowledge Engineering, Wuhan (2005)
7. Shao Y., Han, J., Liu T. and Zhao Y.: Prosodic word boundaries prediction for Mandarin text-to-speech. International Symposium on Tonal Aspects of Languages with Emphasis on Tone Languages, Beijing (2004) 159-162
8. Dong M., Lua K.T. and Li H.: A probabilistic approach to prosodic word prediction for Mandarin Chinese TTS. 9th European Conference on Speech Communication and Technology, Lisbon, Portugal (2005)
9. Qin Shi and XiJun Ma: Statistic prosody structure prediction. International Conference of the IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Ca. (2002)
10. Ying, Z., and Shi, X.: An RNN-based algorithm to detect prosodic phrase for Chinese TTS. International Conference on Acoustic, Speech and Signal Processing (2001)
11. Xun E., Qian Y., Guo Q., Song R.: Applying binary trees as pruning strategy to identify prosody phrase break. Journal of Chinese Information Processing, Vol. 20, No. 3 (2006)