Prosodic Word Grouping with Global Probability Estimation Method

Qing Guo [1], *Jie Zhang* [1], *Nobuyuki Katae* [2]

[1] Fujitsu Research and Develop Center China, Beijing, P.R.C
[2] Fujitsu Laboratories Ltd., Japan
guoqing@cn.fujitsu.com

Abstract

In this paper, a global prosodic word grouping probability estimation method is proposed. By using some statistical probabilities of four kinds of lexical word position types, the optimum prosodic word grouping path of a sentence can be obtained with the dynamic programming approach. In addition, some statistical and rule punish or encourage strategies have been used to improve the accuracy of prosodic word grouping. Lastly, experiment results and discussion are presented.

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]. 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 speaking, 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. Thus, good prosodic word grouping plays an important role in increasing the naturalness of synthesized speech.

Studies have revealed that the prosodic word is quite different from the lexical word. One of the reasons is that the formation of prosodic word is not only based on the meaning of words but also based on rhythm requirements 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 [2] show that the TTS system using the prosodic word, rather than the lexical word directly, as the basic unit has a much higher intelligibility and naturalness.

In recent years, many methods of prosodic word boundary prediction have been proposed in Chinese, such as the classification and regression tree (CART) method[2], rule-driven approach[3], statistical approach[4][5][6][7], recurrent neural network (RNN) method[8] and binary prosodic tree method[9]. In these works, the POS (part of speech) and word length information are mostly used.

The maximum entropy approach and conditional random field (CRF) approach have been introduced to many tasks of natural language processing, and have achieved a good performance in solving some problems such as Chinese word segmentation[10][11]. These word segmentation systems using the above-mentioned two methods are proven to be much effective in the second and the third SIGHAN Chinese Language Processing Backoff (Backoff-2005 and Backoff-2006). In these kinds of systems, a Chinese character is labeled with a tag indicating its position in the Chinese word that the character belongs to. Then character based tagging method for Chinese word segmentation, either based on maximum entropy or CRF, views Chinese word segmentation as a label tagging problem. Lately, maximum entropy and CRF methods were also introduced to solve prosodic word grouping in Mandarin TTS systems[12][13].

Roughly speaking, these methods predict the break type (existence or not) of each boundary site with some features (e.g. length and POS of adjacent words). Statistical model can be trained from a large scale annotated training corpus. Then, for each potential boundary site in an input sentence, a probability is estimated for each possible break type. The break type with the largest likelihood is determined as the correct type. However, we are aware that the current break type is somewhat dependent on the previous break types. For example, if there is a prosodic break in the previous position, the chance of the current position being a prosodic word break will be less. In view of this fact, Dong implemented a Markov chain model in their CART approach to achieve better results by considering the dependency constraints between breaks[5]. Shi presented a statistical method by combining dynamic program method with the rules[6], and Shao designed a Markov Model combined with transformation based error driven learning algorithm to capture the inherent correlation between breaks from the overall judgment[4].

In this paper, a global probability estimation method is proposed to predict the prosodic word boundary. Firstly, four kinds of position types are defined for lexical words according to their positions in the prosodic words that they belong to. Then, we provide a sentence level global prosodic word grouping probability estimation method based on some statistical probability of position types obtained from a training corpus. The optimum prosodic word grouping path of a sentence can be obtained with dynamic programming approach. In addition, some statistical and rule punish or encourage strategies such as length model and noun phrase correlation factor can be used to improve the accuracy of prosodic word grouping.

The rest of this paper is organized as follows. Section 2 briefly introduces the speech corpus used in our research. Section 3 describes the global prosodic word grouping probability estimation method and some strategies used to improve the accuracy of prosodic word grouping. Section 4 provides the experiment results and discussion. Finally, we end this paper with section 5.

2. Speech corpus

The text source of our speech database comes from Chinese People Daily 1998 Corpus, which is transcribed from a Chinese newspaper with word segmentation and POS-tagging annotated for natural language processing purpose. 3360 sentences with about 200k Chinese characters were selected from the text corpus using greedy algorithm.

The prosody structure used in this paper is composed of four tiers[1]: prosodic word (PW), minor phrase (MIP), major phrase (MAP) and intonation group (IG). Prosodic word is a tone group bearing one word stress. Minor phrase contains one or more prosodic words, bears one phrasal stress and the perceived break between MIPs is longer than that between PWs. MAP contains one or more PWs, bears one phrasal stress and the perceived break between MAPs is longer than that between MIPs. The criterion for prosody structure labeling is listening perception. Major phrases are often marked by commas with incomplete pitch resetting while intonation groups are marked by periods, quotation marks or semicolons with full pitch resetting.

Additionally, three levels of stress have been defined, namely, the stressed, the normal and the neutralized.

The following is a sample transcription of a certain sentence in the speech corpus. “|”, “||”, “|||” and “@” represent PW, MIP, MAP and IG in the transcription respectively. A syllable marked with “_H” means it is a stressed syllable, and a syllable marked with “_L” means it is a neutralized one.

８月(ba1 yve4_H)/t | ２０日(er4 sh%2_H r%4_H)/t | 清晨(qing1_H chen2)/t ，||| 一(yi1)/m 支(zh%1_H)/q 满载(man3 zai4_H)/v || 锅碗瓢盆(guo1_H wan3 piao2_H pen2)/l 、|| 桌椅(zhuo1_H yi3)/n 、|| 调料(tiao2_H liao4)/n 、|| 发电机(fa1 dian4 ji1_H)/n || 等(deng3)/u | 家当(jia1 dang4_H)/n 的(de5_L)/u || 流动(liu2 dong4_H)/vn | 支前(zh%1_H qian2)/vn 车队(che1_H dui4)/n || 从(cong2_H)/p 郑州(zheng4 zhou1_H)/ns | 出发(chu1 fa1_H)/v 了(le5_L)/y 。@

3. Global prosodic word grouping probability estimation

Firstly, four kinds of position types are defined for lexical words according to their positions in the prosodic words that they belong to. These four position types are denoted by B_1, B_2, M and I, in which B_1 means that a lexicon word is at the beginning of the prosodic word that it belongs to; B_2 means that a lexicon word is the second lexical word of the prosodic word that it belongs to; I means that a lexicon word belongs to a singleton prosodic word that has only one lexical word; M means that a lexicon word is at the third or other rear position of the prosodic word that it belongs to.

The original transcription of the training sentences can be formatted easily as follows:

The sentence “晚饭/n 后/f ||| 我们/r | 决定/v | 先/d 去/v 逛逛/v || 张家港/ns 的/u | 市容/n 。@” is formatted to “晚饭/n/B1 后/f/B2 ||| 我们/r/I | 决定/v/I | 先/d/B1 去/v/B2 逛逛/v/M || 张家港/ns/B1 的/u/B2 | 市容/n/I 。@”

The rest of this section is organized as follows. Section 3.1 describes the global probability estimation method in detail. Section 3.2 presents some statistical and rule punish or encourage strategies that have been used to improve the accuracy of prosodic word grouping.

3.1. Global probability estimation

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

$W = w_1 w_2 \cdots w_{n-1} w_n$. Let $pos_i, i = 1,2,\cdots,n$ denote the part-of-speech of w_i.

One possible prosodic word grouping result PW for the sentence can be written as follows,

$$PW = w_1 s_1 w_2 s_2 \cdots w_{n-1} s_{n-1} w_n,$$

in which $s_i \in \{B_1, B_2, M, I\}$, $i = 1,2,\cdots,n$

The target of prosodic word grouping is to find the optimum prosodic word grouping PW^* from all possible paths.

$$PW^* = \max_{s_1,s_2,\cdots,s_{n-1}} P(w_1 s_1 w_2 s_2 \cdots w_{n-1} s_{n-1} w_n) \tag{1}$$

This can be approximately estimated by the following formula.

$$\begin{aligned} PW^* &= \max_{s_1,s_2,\cdots,s_{n-1}} P(w_1 s_1 w_2 s_2 \cdots w_{n-1} s_{n-1} w_n) \\ &\approx \max_{s_1,s_2,\cdots,s_{n-1}} \{P(pos_1)P(s_1 \mid pos_1)P(s_2, pos_2 \mid s_1, pos_1) \\ &\cdots P(s_{n-1}, pos_{n-1} \mid s_{n-2}, pos_{n-2})P(s_n, pos_n \mid s_{n-1}, pos_{n-1})\} \end{aligned} \tag{2}$$

Since $P(pos_1)$ is a constant value here, the above formula can be simplified as follows:

$$\begin{aligned} PW^* &= \max_{s_1,s_2,\cdots,s_{n-1}} P(w_1 s_1 w_2 s_2 \cdots w_{n-1} s_{n-1} w_n) \\ &\approx \max_{s_1,s_2,\cdots,s_{n-1}} \{P(s_1 \mid pos_1)P(s_2, pos_2 \mid s_1, pos_1) \\ &\cdots P(s_{n-1}, pos_{n-1} \mid s_{n-2}, pos_{n-2})P(s_n, pos_n \mid s_{n-1}, pos_{n-1})\} \end{aligned} \tag{3}$$

In order to calculate the above formula, five kinds of probabilities should be estimated from our training corpus. These five kinds of probabilities are described as follows.

(1) The probability that a POS is the POS of a singleton prosodic word, namely,

$$\begin{aligned} &P(s = I \mid pos = pos\_i) \\ &= \frac{C(s = I, pos = pos\_i)}{C(s = I, pos = pos\_i) + C(s = B_1, pos = pos\_i)} \end{aligned}$$

(2) The probability that a POS is the POS of the first lexical word in a prosodic word, assuming that this prosodic word contains at least 2 lexical words, namely,

$$\begin{aligned} &P(s = B_1 \mid pos = pos\_i) \\ &= \frac{C(s = B_1, pos = pos\_i)}{C(s = I, pos = pos\_i) + C(s = B_1, pos = pos\_i)} \end{aligned}$$

(3) The transition probability from position B_1 to position B_2 within a prosodic word.

$$\begin{aligned} &P(s = B_2, pos = pos\_j \mid s_{prev} = B_1, pos_{prev} = pos\_i) \\ &= \frac{C(s_{prev} = B_1, pos_{prev} = pos\_i, s = B_2, pos = pos\_j)}{C(s_{prev} = B_1, pos_{prev} = pos\_i)} \end{aligned}$$

(4)The transition probability distribution from position B_2 or position M to position M.

$$\begin{aligned} &P(s = M, pos = pos\_j \mid s_{prev} = B_2 orM, pos_{prev} = pos\_i) \\ &= \frac{C(s_{prev} = B_2 orM, pos_{prev} = pos\_i, s = M, pos = pos\_j)}{C(s_{prev} = B_2 orM, pos_{prev} = pos\_i)} \end{aligned}$$

(5) The jump probability in a prosodic word boundary.

$$
\begin{aligned}
&P_{jump}(pos = pos\_j \mid pos_{prev} = pos\_i)\\
&= P(s = B_1 orI, pos = pos\_j \mid s_{prev} = B_2 orMorI, pos_{prev} = pos\_i)\\
&= \frac{C(s_{prev} = B_2 orMorI, pos_{prev} = pos\_i, s = B_1 orI, pos = pos\_j)}{C(pos_{prev} = pos\_i, pos = pos\_j)}
\end{aligned}
$$

Trigram jump probability is also tried in our experiments. It can be represented as follows.

$$
\begin{aligned}
&P_{jump}(pos = pos\_j \mid pos_{prev} = pos\_i, pos_{prev-1} = pos\_k)\\
&= P(s = B_1 orI, pos = pos\_j \mid s_{prev} = MorI, pos_{prev} = pos\_i, pos_{prev-1} = pos\_k)\\
&= \frac{C(s_{prev} = MorI, pos_{prev} = pos\_i, s = B_1 orI, pos = pos\_j, pos_{prev-1} = pos\_k)}{C(pos_{prev} = pos\_i, pos = pos\_j, pos_{prev-1} = pos\_k)}
\end{aligned}
$$

In the previous formula (3), if $s_i = B_1 \vee I, i > 1$ then $P(s_i, pos_i \mid s_{i-1}, pos_{i-1})$ is estimated approximately as follows:

$$
\begin{aligned}
&P(s_i, pos_i \mid s_{i-1}, pos_{i-1})\\
&\approx P_{jump}(pos_i \mid pos_{i-1}) P(s_i \mid pos_i)
\end{aligned} \tag{4}
$$

The global prosodic word grouping probabilities of a sentence with various possible grouping paths will be calculated using the above five probability by the dynamic programming approach. The path with the biggest probability will be treated as the optimum prosodic word grouping result.

Remarks. Unlike usually defined character position types in word segmentation task, position E (the end position of a prosodic word) is not defined in our method for prosodic word grouping. In fact, the lexical word of position type E will be labeled as position M or position B_2 in the paper. We have also experimented five position types including E , however, it will cause prosodic word jump too much. We think that the reason is that the jump probability has already characterized the ending of a previous prosodic word.

3.2. Some statistical-based or rule-based strategies

3.2.1. Length model

Prosodic word length model can be used in the global prosodic word grouping probability estimation method as follows:

$$
\begin{aligned}
&PW* = \max_{s_1, s_2, \cdots, s_{n-1}} P(w_1 s_1 w_2 s_2 \cdots w_{n-1} s_{n-1} w_n)\\
&\approx \max_{s_1, s_2, \cdots, s_{n-1}} \{P(s_1 \mid pos_1) P(s_2, pos_2 \mid s_1, pos_1)\\
&\cdots P(s_{n-1}, pos_{n-1} \mid s_{n-2}, pos_{n-2}) P(pos_n \mid s_{n-1}, pos_{n-1})\\
&P(len(PW_1)) P(len(PW_2)) \cdots P(len(PW_j))\}
\end{aligned}
$$

, where $P(len(PW_j))$ can be estimated from the training corpus.

3.2.2. Noun phrase correlation factor

Since our word segmentation module does not provide noun phrase detecting function, noun phrase correlation factors are designed to characterize the correlation between adjacent noun pairs. These factors are helpful in avoiding incorrectly prosodic word boundary insertion within some noun phrases.

$$
\begin{aligned}
&NounPhraseFactor_{forward}(w = word, pos = noun)\\
&= \frac{C(w = word, pos = noun, pos_{next} = noun)}{C(w = word, pos = noun)}\\
&NounPhraseFactor_{backward}(w = word, pos = noun)\\
&= \frac{C(pos_{prev} = noun, w = word, pos = noun)}{C(w = word, pos = noun)}
\end{aligned}
$$

For example, “这/r 位/q 传奇/n 人物/n” will be processed as “这 位 传奇 | 人物” before. After using the new method to compute the correlation factor between two nouns, “传奇” and “人物”, we will find that the correlate factor is very high; hence the prosodic word boundary is not encouraged to be inserted into the place between these two nouns. At last, the processing result will be “这 位 | 传奇 人物” now.

Verb+noun and verb+verb phrase correlation factors are also used in the dynamic programming approach.

3.2.3. Some special Chinese characters and some fixed syntactic patterns

Some rules are designed from linguist knowledge to punish or encourage the probability of the grouped prosodic word according to the context. For example, some rules are designed for some special Chinese characters such as “之”, “所”, “不” and “一”. In addition, some rules are designed for some fixed syntactic patterns with “是”.

Furthermore, the POS tendency verb (vq) is added into our POS set. It is because that a tendency verb will cohere with its preceding verb in the prosodic word level. This will lead another advantage. It will decrease the bias of probability estimation because some “v+v” POS sequences will be replaced by “v+vq” in the training set.

4. Experiment results and discussion

4.1. Test set

An independent test corpus, which is also selected from Chinese People Daily 1998 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. Finally, 5113 prosodic word boundaries were annotated in the test set.

4.2. Results and discussion

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

Table 1. *The experiments result 1*

Results methods	Precision	Recall rate
Bigram jump probability	86.01%	83.94%
Trigram jump probability	86.23%	82.51%
Bigram jump with length model	87.63%	83.39%
Trigram jump with length model	87.82%	82.19%

Table 1 describes the experiments result of prosodic word grouping with the proposed global probability estimation method. At first, a comparison was made between the performance of using bigram jump probability and that of using trigram jump probability. From the table we can see that trigram jump probability can improve the precision slightly, however it degrades the recall rate to a relative large extent. We think that the reason is less of training data for trigram jump probability estimation. Then, the length model was proven to improve the precision about 1.5% with slightly degradation on the recall rate. At last, we decide to use bigram jump probability in our baseline system.

Table 2 gives the experiment result with all statistical and rule encourage or punishment strategies being used. From the table we can see that these statistical and rule encourage or punishment strategies contribute to the improvement of prosodic word grouping much. Table 2 also gives the result when a module of automatic word segmentation and POS tagging is applied. It is observed that the result is better than that of our previous binary prosodic tree method[9] in terms of both the accuracy and the memory cost.

Table2. *The experiment result 2*

results methods	Precision	Recall rate
Baseline	86.01%	83.94%
Baseline + Length model	87.63%	83.39%
Baseline + All strategies	89.56%	84.37%
Automatic WordSeg+POSTagging	87.52%	82.84%
Binary Prosodic Tree Method	85.91%	79.38%

Figure 1 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 [2]. 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@

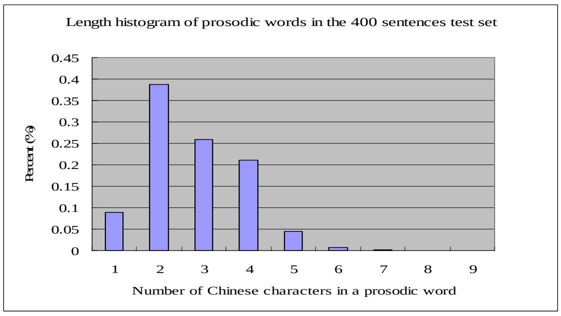

Fig. 1. *Length histogram of prosodic words in the test set.*

The prosodic word grouping method in this paper needs to solve 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.

5. Conclusions

In this paper, we provide a sentence level global prosodic word grouping probability estimation method based on some statistical probability of position types of lexical words. The optimum prosodic word grouping path of a sentence can be obtained with dynamic programming approach. Finally, some statistical and rule punish or encourage strategies have been used to improve the accuracy of prosodic word grouping. The experiment result is quite promising.

6. References

[1] Li, A.; Lin, M., 2000. Speech corpus of Chinese discourse and the phonetic research. *International Conference on Spoken Language Processing*.

[2] Chu M.; Qian Y., 2001. Locating boundaries for prosodic constituents in unrestricted Mandarin texts. *Journal of Computational Linguistics and Chinese Language Processing*, 6(1):61-82.

[3] Tao, J.; Dong, H.; Zhao, S., 2003. Rule learning based Chinese prosodic phrase prediction. *International Conference on Natural Language Processing and Knowledge Engineering*. Beijing.

[4] Shao Y.; Han, J.; Liu T; Zhao Y., 2004. Prosodic word boundaries prediction for Mandarin text-to-speech. *International Symposium on Tonal Aspects of Languages with Emphasis on Tone Languages*. Beijing, 159-162.

[5] Dong M.; Lua K.T.; Li H., 2005. A probabilistic approach to prosodic word prediction for Mandarin Chinese TTS. *9th European Conference on Speech Communication and Technology*. Lisbon, Portugal.

[6] Qin Shi; XiJun Ma, 2002. Statistic prosody structure prediction. *International Conference of the IEEE 2002 Workshop on Speech Synthesis*. Santa Monica, Ca..

[7] Chou F.; Tseng C.; Lee L, 1996. Automatic Generation of Prosodic Structure for High Quality Mandarin Speech Synthesis. *International Conference on Acoustic, Speech and Signal Processing*. 1624-1627.

[8] Ying, Z.; Shi, X., 2001. An RNN-based algorithm to detect prosodic phrase for Chinese TTS. *International Conference on Acoustic, Speech and Signal Processing*.

[9] Guo Q.; Xun E.; Katae N., 2006. Prosody word grouping in Mandarin TTS system. *International Symposium on Chinese Spoken Language Processing*. Sigapore.

[10] Low Jin Kiat; Ng Hwee Tou; Guo Wenyuan, 2005. A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*. Jeju Island, Korea, 161-164.

[11] Zhao Hai; Huang Chang Ning; Li Mu, 2006. An Improved Chinese Word Segmentation System with Conditional Random Field. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Sydney, Australia, 162-165.

[12] Zhang X.; Xu J.; Cai L., 2006. Prosodic Boundary Prediction Based on Maximum Entropy Model with Error-Driven Modification. *International Symposium on Chinese Spoken Language Processing*. Sigapore.

[13] Kang H.; Liu W., 2006. Prosodic Words Prediction from Lexicon Words with CRF and TBL Joint Method. *International Symposium on Chinese Spoken Language Processing*. Sigapore.