Computational Linguistics and Chinese Language Processing
Vol. 6, No. 1, February 2001, pp. 61-82

Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts

Min Chu*, Yao Qian+

Abstract

This paper proposes a three-tier prosodic hierarchy, including prosodic word, intermediate phrase and intonational phrase tiers, for Mandarin that emphasizes the use of the prosodic word instead of the lexical word as the basic prosodic unit. Both the surface difference and perceptual difference show that this is helpful for achieving high naturalness in text-to-speech conversion. Three approaches, the basic CART approach, the bottom-up hierarchical approach and the modified hierarchical approach, are presented for locating the boundaries of three prosodic constituents in unrestricted Mandarin texts. Two sets of features are used in the basic CART method: one contains syntactic phrasal information and the other does not. The one with syntactic phrasal information results in about a 1% increase in accuracy and an 11% decrease in error-cost. The performance of the modified hierarchical method produces the highest accuracy, 83%, and lowest error cost when no syntactic phrasal information is provided. It shows advantages in detecting the boundaries of intonational phrases at locations without breaking punctuation. 71.1% precision and 52.4% recall are achieved. Experiments on acceptability reveal that only 26% of the mis-assigned break indices are real infelicitous errors, and that the perceptual difference between the automatically assigned break indices and the manually annotated break indices are small.

1. Introduction

The state-of-the-art text-to-speech (TTS) systems are able to produce very natural synthesized

* Microsoft Research Asia, 5F, Beijing Sigma Center No. 49, Zhichun Road, Haidian District, Beijing 100080, P.R.C., E-mail: minchu@microsoft.com
+ Shanghai Normal University, 100 Guilin Road,Shanghai 200234, P.R.C., E-mail: yqian@shtu.edu.cn

speech if they are provided with a correct phonetic string and with prosodic features extracted from human pronunciation of the string [Chu and Lu, 1996; Dutoit *et al*., 1996]. Automatic prosody generators, however, cannot yet deliver high quality prosody. One of the main obstacles to automatic generation of prosody is the difficulty of identifying the hierarchical prosodic constituents from texts automatically. It has been proven through many experiments [Liberman and Prince, 1977; Gee and Grosjean, 1983; Selkirk, 1984; Ladd and Campbell, 1991] that prosody constituents are not always identical to those of the surface syntax. The relationship between prosody and syntax is not well understood. While, representing prosodic constituents by means of syntactic constituents directly cannot produce very natural prosody, the boundaries of prosodic constituents can be derived from syntactic information. Some early studies used rules to parse prosodic structures. Stochastic models have been used more frequently in recent studies. In some works [Wang and Hirschberg, 1991; Hirschberg and Prieto, 1996; Lee and Oh, 1999], break indices have been predicted using the automatic *classification and regression tree* (CART) from information such as four-word part-of-speech (POS) windows, pitch accent types, the sentence length, the distance from the beginning of the sentence and the end of the sentence, etc. A Markov model is used in works done by Veilleux *et al*. [Veilleux *et al*., 1990], which predicts the most likely sequence of break indices from the input POS sequence based on the assumption that the current index is only related to the previous index. Ostendorf and Veilleux [Ostendorf and Veilleux, 1994] proposed a hierarchical stochastic model for locating prosodic boundaries. Most of the publications on locating prosodic boundaries have focused on alphabet-based languages, such as English, which are very different from Mandarin in nature. Chou *et al*. [Chou *et al*., 1996; Chou *et al*., 1998] presented a top-down procedure for labeling break indices in Mandarin from both acoustic features, such as f_0, duration and energy, and features derived from text transcriptions. They reported that the acoustic features are helpful for predicting prosodic phrases. Since the prosodic boundary detecting approach presented in this paper is meant to be used in the Mandarin TTS system, where no acoustic features are available, only features that can be derived from text transcriptions will be used.

There are many reports specifying various hierarchical structures for prosodic constituents. The intonational phrase (INP) and the intermediate phrase (IMP) are the most commonly accepted levels in English. An English sentence consists of a sequence of INPs and each INP, in turn, is composed of a sequence of IMPs. INPs should have boundary tones at their ends, and IMPs are theoretically marked with phrase accents. Both types of phrases are cued by lengthening of the final syllables. With the above definition of prosodic hierarchy in English, studies have been done on predicting either one of the two prosodic phrases or both. The two prosodic constituents have been referred to as the major phrase and minor phrase in some papers. The word is used as the basic unit in all prosodic-phrase detecting algorithms in English.

Though Ostendorf and Veilleux [Ostendorf and Veilleux, 1994] mentioned the usefulness of considering the prosodic word (LW) rather than the lexical word (LW) as possible sites for break indices, they did not use them in their prosody model since it was difficult to define PWs relative to LWs. Most prosody related studies on Mandarin [Chou *et al*., 1996; Shen and Xu, 2000] have borrowed the two levels of prosodic phrases from English. In addition to the IMP and the INP, Chou *et al*. [Chou *et al*., 1998] defined a breath group boundary and a prosodic group boundary for short paragraphs. The two groups often contain more than one simple sentence. In this paper, only prosodic constituents smaller than sentences will be studied. Only the INP and the IMP are kept. However, our study shows that the PW word is a very important prosodic unit for Mandarin. The surface difference and perceptual difference between the PW and the LW will be introduced in Section 2. These differences show that using PWs instead of LWs as the basic unit of prosody will lead to improved naturalness of the synthesized speech. Thus, in our approach, a three-tier hierarchy is defined for prosody below the sentence level in Mandarin. The PW is the lowest constituent in the prosodic hierarchy. The middle tier is the IMP, which has a perceptive minor break at the end. The INP is the top tier with a major break at the end. The concepts of phrase accent and boundary tone in English are not easy to use in the definition of the IMP and the INP in Mandarin since Mandarin is a tonal language. The degree of break becomes the main cue for identifying them in real speech. The aim of this study was to locate the boundaries for the three-tier prosodic constituents automatically in unrestricted Chinese texts, using only information that can be derived from the texts.

The remainder of this paper is organized as follows. Section 2 discusses the surface difference and perceptive difference between the PW and the LW. Section 3 defines the three-tier prosodic constituents in Mandarin. Section 4 presents the three approaches to locating prosodic boundaries. Experiments and results are given in Section 5. Section 6 gives conclusions.

2. Prosodic Word vs. Lexicon Word

Since in many Asian languages, such as Chinese, Japanese or Korean, texts do not contain any visual cues for word boundaries, word segmentation becomes a basic requirement for almost all text analyses in these languages. Many studies had been done on word segmentation. Chinese has a very flexible list of words. The size of the lexicon used for word segmentation changes from 40,000 items to several hundreds of thousands of items. Most Chinese characters are words by themselves and also parts of longer words. The length of a word in characters ranges from 1 to 10 or more. However, in spoken Chinese, there exists a disyllabic rhythm. Succeeding mono-character words are often uttered as one disyllabic unit of rhythm, and long words are often uttered as several units. The unit of rhythm in Mandarin is referred as the prosodic word, which is defined as a group of syllables that should be uttered closely and continuously.

Although, in real speech, not all boundaries of PWs have breaks, it is tolerable if there is a break at the end of each PW. However, any inner PW break will make the speech unintelligible or unnatural. To distinguish then from PW, words listed in a lexicon used in word segmentation are referred to as lexical words. A PW may contain one or more LWs and it may also be only part of an LW. For example, in the Chinese sentence, "**我买了一本好书** (I brought a good book)," each character itself is an LW. Yet, in natural speech, the sentence is grouped as "**我\买了\一本\好书**." There are four PWs. Since a PW is formed dynamically according to the context, many possible combinations of characters exist in real texts. It is impossible to list all the PWs in a lexicon as has been done for LWs. However, PW strings can be predicted from LW strings [Qian *et al*., 2001].

In an exploratory experiment, three annotators were asked to label the PW boundaries in 1348 utterances, with text transcriptions provided for these utterances. Table 1 lists the main guidelines for labeling PWs in speech. PW boundaries were labeled by both listening to the utterances and reading the text transcriptions. A 96.9% agreement ratio was achieved across three of them. The agreement ratio among at least two of them reached 99.9%. The high agreement ratio shows consistency in PW labeling across different people.

Table 1. *The main guidelines for labeling PWs by listening to the utterances and reading the text transcriptions.*

1.	A disyllabic or tri-syllabic LW is a PW if it has no proclitic or enclitic. Otherwise, it forms a PW with its clitic. Examples of enclitics are "**的、了、着、（楼）上、（地）下、（物理）学、（革命）性**"; examples of proclitics are "**副(所长)、半(正式)**."
2.	A mono-syllabic LW often forms a PW with the LW coming before or after it. Only when a mono-syllabic LW is lengthened enough to balance the disyllabic rhythm does it become a mono-syllabic PW.
3.	All LWs containing more than 3 syllables should be segmented into several disyllabic or tri-syllabic PWs according to their structures. When there are proclitics or enclitics, the clitics merge into the first or last PW in the long LW.

Comparing LW boundaries obtained by a well developed word segmentation tool with the PW boundaries labeled manually, we found that only 70.7% of the LW boundaries coincided with the PW boundaries, and that 6.4% of the PW boundaries are not LW boundaries. Figure 1 shows the histogram of the lengths of PWs and LWs counted in a large corpus. It can be seen that there are less mono-syllabic PWs than mono-syllabic LWs because most of the

mono-syllabic LWs form disyllabic or tri-syllabic PWs with their neighbors dynamically. Only 1.3% of the PWs contain more than three characters, and the longest PW found in the corpus contains 5 characters. They are often disyllabic or tri-syllabic LWs followed by several clitics, such as "**煮熟的了吗？**" The higher ratio of disyllabic PWs shows that the PWs reflects the disyllabic rhythm in Mandarin better than the LWs. If speech is synthesized from LWs, the high ratio of mono-syllabic words will decrease the level of naturalness achieved.

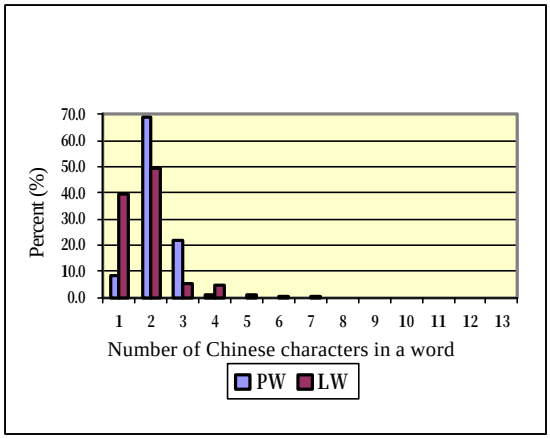


Figure 1 *Histogram of lengths of PWs and LWs in number of characters.*

To investigate the differences between PWs and LWs from the perceptual point of view, a preference experiment was conducted. Speech waveforms were synthesized from two types of input by the MSRCN Mandarin TTS engine [Chu *et al*., 2001]:

A. Sentences were segmented into LW strings, and the LW was used as the basic unit for prosody.
B. Sentences were segmented into PW strings, and the PW was used as the basic unit for prosody.

108 pairs of synthesized speech were played to 15 subjects, who were asked to choose a more natural utterance from each pair. The preference percentages for type A and type B utterances were 21% and 79%, respectively. Speech synthesized from PW strings sounds significantly better than that synthesized from LW strings.

Both the surface difference and perceptual difference between LWs and PWs show that segmenting a sentence into a string of LWs precisely is far from sufficient to generate natural and beautiful prosody in Mandarin TTS systems; it is necessary to re-segment LW strings into PW strings.

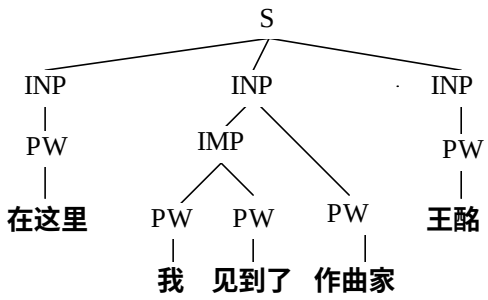
3. Prosodic Constituents in Mandarin

As noted in Section 2, it is very helpful to use the PW instead of the LW as the basic prosody unit. A three-tier instead of the conventional two-tier prosodic hierarchy is defined for a sentence in Mandarin. A sentence consists of one or more INPs. An INP is decomposed into several IMPs and the building blocks for an IMP are PWs. The PW is the lowest constituent in the hierarchy. An INP boundary necessarily coincides with an IMP boundary, and an IMP boundary is inevitably a PW boundary, but the opposite is not true.

Though prosodic constituents should have some relationships with syntactic constituents, the relationships between them are unclear. Figure 2 shows an example sentence "**在这里我见到了作曲家王酪** (We saw Wangming, a composer, here)," which is decomposed into a syntactic hierarchy and a prosodic hierarchy. The differences between them are obvious.

A corpus with both prosodic and syntactic labeled structures was prepared. Three-level prosody boundaries were labeled manually after listening to the speech and reading the text transcriptions. Details about the labeling process will be given in Section 5.1. A block-based robust dependency parser [Zhou, 2000] was used to parse all these sentences into syntactic trees. On one hand, only 56.9% of the INP boundaries and 56.4% of the IMP boundaries coincided with the boundaries of top-level syntactic phrases. On the other hand, less than half of the top-level syntactic phrase boundaries were INP boundaries. Figure 3 shows the percentage of syntactic phrase boundaries that coincided with INP boundaries for 7 major syntactic phrase tags. Since great mismatching exists between prosodic phrases and syntactic phrases, directly mapping syntactic phrases to prosodic phrases will cause many unsuitable breaks in synthesized speech. Section 4 will present three approaches to locating prosodic boundaries.

(a)



(b)

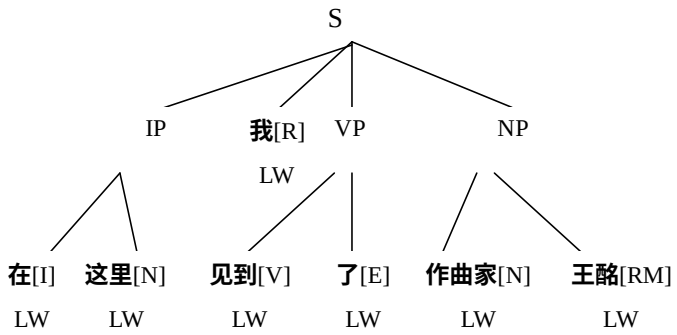


Figure 2 *(a) The prosodic hierarchy and (b) the syntactic hierarchy for the sentence, "**在这里我见到了作曲家王酩** (We saw Wangming, a composer, here)."*

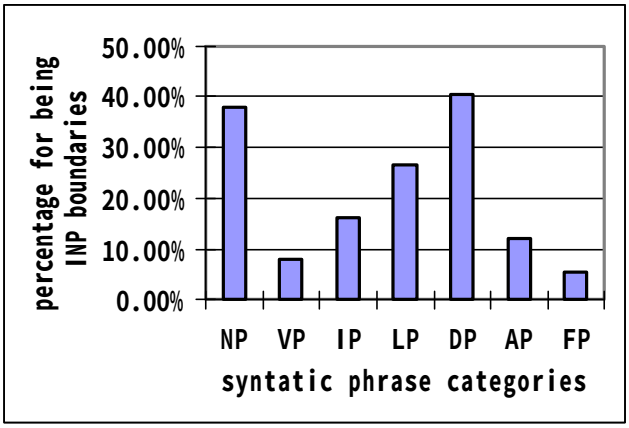


Figure 3 *The percentage of syntactic phrase boundaries that coincided with INP boundaries for 7 major syntactic phrase types. NP - Noun phrase; VP - Verb phrase; IP - Preposition phrase; LP - Post-position phrase; DP - Frame structure; AP - Adjective phrase; FP - Adverb phrase.*

4. Approaches to Locating Prosodic Boundaries

Though, representing prosodic structures by means of syntactic structures directly cannot produce very natural prosody, syntactic information is still helpful for detecting prosodic boundaries. POS has been used in many studies on prosodic phrase prediction. Veilleux *et al*. [Veilleux *et al*., 1990] modeled prosodic group labels and phrase breaks as a six-state Markov chain. Both first- and second- order Markov models were investigated. They reported that using the second-order model did not improve the results. Taylor and Black [Taylor and Black, 1998] used the Markov model in a different way. In their model, state observation probabilities were estimated using a POS sequence model, and the state transition probabilities were estimated using a phrase break model. Wang and Hirschberg [Wang and Hirschberg, 1991] used CART to predict INP boundaries. In Ostendorf and Veilleux' s study [Ostendorf and Veilleux, 1994], CART was used to determine the probability of the occurrence of a minor break at some locations, and a hierarchical stochastic model was used to find the prosodic parse with the highest probability. The Markov model based approaches are based on the assumptions that the current break index is only related to previous indices, and that the state probability and transition probability can be estimated from POS tags of the word sequences. It is difficult to use other syntactic information and length information of phrases and sentences in them. CART based approaches were used in our studies because they can handle data samples with high dimensions, mixed data types and nonstandard data structures. CART based methods also have

the advantage of being comprehensible in the prediction phase. Three predicting models will be presented in this section, and two sets of features will be applied in the training of CART.

Since many Chinese sentences do not have exclusive solutions for LW segmentation and it is possible to have breaks inside some long LWs, each character in a text is assumed to be followed a *potential boundary site* (PBS). Four break indices (BI) are used to label the types of PBS. BI0 represents a non-boundary site. If a PBS is only a PW boundary, it is labeled BI1. BI2 represents an IMP boundary, and BI3 represents an INP boundary. The problem of locating boundaries of prosodic constituents is then changed to the problem of predicting BI for each PBS.

4.1 The basic CART method

CART is used to predict BI for each PBS first. In early CART based approaches [Wang and Hirschberg, 1991; Ostendorf and Veilleux, 1994], features that took continuous values or many discrete values were classified into a limited number of categories first to prevent the excessively dense trees. In many cases, this was done by experts according to their experiences. The number of categories and the way of doing classification would affect the final results. In our approach, the composite-question construction technique [Huang *et al*., 2001] is used to generate complex questions for the tree. The construction of composite questions not only enables flexible clustering of discrete variables, but also produces complex rectangular partitions for continuous variables. Thus, only simple questions about the details of all the features are presented for growing the tree in the training phase.

4.2 The bottom-up hierarchical approach

In the basic CART method, the four BI are treated as being the same, although they have hierarchical relationships. Error analyses show that, sometimes, a BI3 or BI2 is assigned to a non-boundary PBS. This kind of error will decrease not only the naturalness, but also the intelligibility of the synthesized speech. Since PW boundaries can be predicted from LW boundaries with pretty high accuracy [Qian *et al*., 2001], a bottom-up hierarchical approach was proposed. In the new approach, PW boundaries are first detected from all PBS. Then, IMP boundaries are detected only from PBS that are judged to be PW boundaries. Finally, INP boundaries are picked up only from the predicted IMP boundaries. Figure 4 shows the flowchart of the hierarchical approach. Three CARTs were trained separately to make boundary or non-boundary decisions for PWs, IMPs and INPs, respectively. The training procedures for the three CARTs were the same as that described in Section 4.1. However, the data used for training were different. To train the PW-CART, all the PBS with BI0 were treated as non-boundary

samples and all the others were boundary samples. To train the IMP-CART, only PBS with BI1 were used as non-boundary samples, and those with BI2 and BI3 were boundary samples. To train INP-CART, only PBS with BI2 were used as non-boundary samples, and PBS with BI3 were boundary samples.

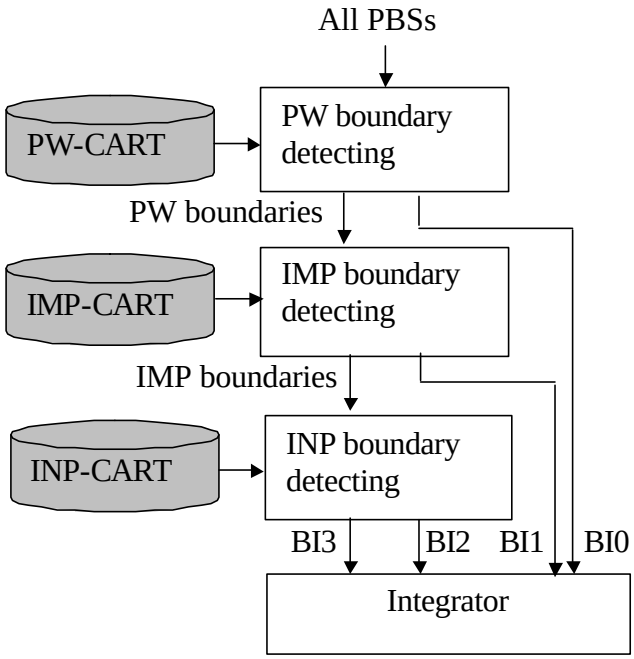


***Figure** 4 Flowchart of the bottom-up hierarchical approach for detecting boundaries of prosody constituents.*

***Table 2.** The average length (ALC) of PWs IMPs and INPs, and their correlation coefficients (CCO) with the lengths of their carrying sentences.*

	PW	IMP	INP
ALC	2.2	3.3	6.7
CCO	0.059	0.155	0.488

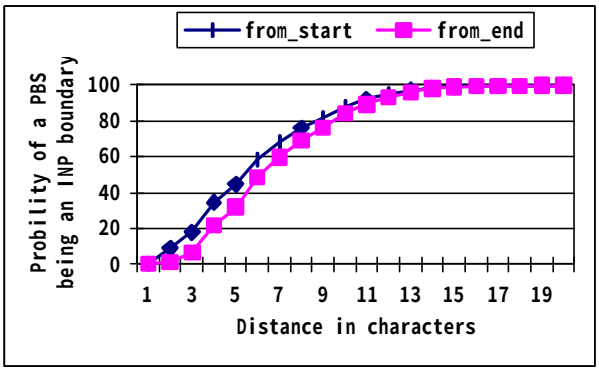


Figure 5 *The probability of a PBS being an INP boundary in terms of its distance to the beginning and the end of a sentence.*

4.3 The modified hierachical approach

In the above two approaches, INP boundaries and IMP boundaries are often confused. We have found that the lengths of sentences in characters, a very important factor that affects the positions of INP boundaries, has not been used sufficiently. The correlation coefficients between the lengths of PWs, IMPs and INPs and the lengths of their carrying sentences, and the average lengths of the three prosodic constituents are listed in Table 2. It can be seen that the length of a PW is uncorrelated with the length of its carrying sentence. The length of an IMP is weakly correlated with the length of its carrying s entence, and the length of an INP is positively correlated with the length of its carrying sentence and tends to increase with it. Statistical results show that the location of an INP boundary is not only related to the length of its carrying sentence, but is related to its distance to the beginning and the end of the sentence. Figure 5 shows two curves revealing the relationship between the probability of a PBS being an INP boundary and its distance to the beginning and the end of its carrying sentence. It is obvious that the PBS at the middle part of a sentence has a higher probability of being an INP boundary than those at the beginning or the end of the sentence. A modified hierarchical approach is proposed based on this observation and another assumption that finding the most likely location for INP boundaries in a sentence one by one is more accurate than finding all INP the boundaries in one loop. In the modified approach, the PW and IMP detecting procedures are the same as those described in Section 4.2. However, the INP detecting procedure is modified to be a recursive detecting method. The output of INP-CART is no longer a boundary or non-boundary decision. Instead, a probability of a PBS being an INP boundary (denoted as P_B) is generated. The use of

INP-CART is similar to that in Ostendorf and Veilleux' s works. P_B for each leaf of the CART is calculated during the training phase. It is defined as the number o f boundary samples over the number of total samples in the leaf. In the prediction phase, when a leaf is selected for an input PBS, its P_B is output as the probability of the PBS being an INP boundary. A confidence measure (denoted as ConfM) for a PBS being an INP boundary is defined by equation (1):

$$ConfM = P_B * P_{start} * P_{end}, \quad (1)$$

where P_{start} and P_{end} are the probabilities of the PBS being an INP boundary in terms of its relative distance to the beginning and the end of the sentence. Their values are defined by the two curves shown in Figure 5, when the distance is smaller than 20. Otherwise they are equal to 1.

The recursive INP boundary detecting algorithm is decomposed into four steps.

Step1: ConfM values are calculated for all PBS that have been detected as IMP boundaries by the IMP-CART. BI3 is assigned to the one with the highest ConfM value, if its ConfM value is larger than the pre-set threshold q_{ConfM} . If no PBS with a ConfM value larger than q_{ConfM} is found, go to Step 4.

Step2: Split the sentence into two parts at the found INP boundary.

Step3: Repeat Step1 and Step 2 for the two new sub-sentences recursively until all paths reach Step 4.

Step4: Stop.

The performance changes with the value of q_{ConfM} , and it is set according to previous experience or experiments. In our case, the best result was achieved when $q_{ConfM} = 0.105$.

5. Experiments

Experiments using the three methods described in Section 4 were carried out on a large speech corpus. The speech corpus, features used, and results from the experiments will be discussed in this section.

5.1 Speech corpus

Since there was no public Mandarin speech database available for this study, we designed and collected a large phonetically and prosodically enriched Mandarin speech corpus. The corpus contains about 12,000 utterances (sentences), which were uttered by a professional female speaker. Prosodic indices BI1 to BI3 were annotated manually by listening to these utterances and reading the text transcriptions. BI1 was annotated according to the guidelines listed in Table 1. BI2 and BI3 were labeled according to the breaks heard. When a minor break was perceived,

a BI2 was assigned. When a major break was heard, a BI3 was assigned. BI2 and BI3 were assumed to correspond to IMP and INP boundaries, respectively. The end of each utterance was labeled with BI3, and each non-boundary PBS was labeled with BI0 automatically.

To check consistency of annotation across different people, an exploratory experiment was carried out. Three annotators were first trained on the same 100 sentences. At this stage, they were required to discuss criteria for annotation so that they could achieve agreement on most of the annotations in the 100 sentences. Then, they were asked to annotate a small subset of the corpus, which included 1,348 sentences and 1,8983 PBS. All three annotators achieved agreement on 82.9% of BIs, and 99.1% of BIs were agreed to by at least two of them. That is to say pretty good consistency existed among the three annotators. To reduce costs, the whole speech corpus was only annotated by one of them.

Investigating the relationship between BI types and punctuation, such as commas, colons and semicolons, we found that there were altogether 5,718 items of punctuation in the corpus (full stops at the ends of sentences were excluded), 5,693 (99.6%) of which were related to BI3 and the rest related to BI2. These kinds of punctuation are referred to as breaking punctuation (BP). Since BPs almost always imply INP boundaries, no learning process is needed for them. All PBS with BPs were assigned to BI3. This has been done in many Mandarin TTS systems. However, placing major breaks only at PBS with BPs is not adequate for synthesizing high quality speech. The ability to predict INP boundaries at PBS without BPs is more important. Thus, accuracy for INP boundaries is calculated using two constraints in this paper. In one constraint, all predicted INP boundaries, including INP boundaries at PBS with BPs, are considered. In the other constraint, only INP boundaries at PBS without BPs are taken into account.

Most of the early studies on detecting prosodic phrases experimented on small databases. Wang and Hirschberg used a 298-utterance corpus, and Ostendorf and Veilleux used 312 sentences in their experiments. Only a limited number of INP boundaries can be found in such small corpora. Thus, only a few features can be used in the training and testing phase to avoid sparsity of training data. A larger training and testing data set was used in this study. 2,583 sentences with 38,499 characters from the corpus we collected were used for training, and another 1,000 sentences with 15,618 characters were used for testing.

5.2 Feature set

Although both acoustic features [Wightman and Ostendorf, 1994; Chou *et al.*, 1998], such as f_0, duration and energy, and syntactic features [Hirschberg and Prieto, 1990; Wang and Hirschberg, 1991; Ostendorf and Veilleux, 1994; Lee and Oh, 1999], such as POS tags and syntactic phrasal

information, have been used to label break indices, only features that can be derived from texts were used in this study. The reason is that no acoustic feature is available when we predict prosodic boundaries in TTS systems. Two feature sets with or without syntactic phrasal information were used. Set 1 is the one without syntactic phrase information. Features used in Set 1 are listed as follows:

1) POS for LWs around each PBS are the most commonly used features in prosodic phrase prediction. A window of three words is used in our approach: two words before and one after the PBS. 26 POS tags are used. Among them, 9 are the normal POS tags used by Zhou's parser [Zhou, 2000]. The others are characters or words that often have special effects on prosodic boundaries. These characters and words are obtained through data analyses and should be considered individually. All 26 tags are listed in Table 3.

Table 3. *The 26 POS tags used in our experiments*

Tags	Explanation	Tags	Explanation	Tags	Explanation
N	Noun	Char1	Mono-syllabic LW "**电**"	Char 10	Mono-syllabic LW "**从**"
V	Verb	Char2	Mono-syllabic LW "**中**"	Word1	Disyllabic LW "**但是**"
A	Adjective	Char3	Mono-syllabic LW "**后**"	Word2	Disyllabic LW "**目前**"
F	Adverb	Char4	Mono-syllabic LW "**的**"	Word3	Disyllabic LW "**今天**"
DM	Place name	Char5	Mono-syllabic LW "**在**"	Word4	Disyllabic LW "**短波**"
RM	Person name	Char6	Mono-syllabic LW "**于**"	Word5	Disyllabic LW "**简讯**"
QM	Organization name	Char7	Mono-syllabic LW "**了**"	Word6	Disyllabic LW "**接着**"
E-I-L-J	Auxiliary, preposition, post-preposition and junction	Char8	Mono-syllabic LW "**等**"	Word7	Disyllabic LW "**就是**"
Other	All other POS	Char 9	Mono-syllabic LW "**着**"		

2) The length in characters of the LW in the window is very important for predicting PW boundaries. It takes 5 discrete values: 1- 4 represent LWs containing 1-4 characters,

respectively. 5 represents all LWs containing more than 4 characters.

3) The distance in characters from the current PBS to the beginning or the end of a sentence. The shorter one among the two is used. As shown in Figure 5, the lengths are divided into four groups, which are <=2, 3-6, 7-10 and >10, respectively.
4) The lengths in characters of the carrying sentences are divided into three groups, which are <=10, 11-20 and >20, respectively.

Set 2 contains all the features in set 1 and the phrasal features listed below:

1) Whether the current PBS is a top-level major syntactic phrase boundary or not.
2) The phrase category for the carrying phrase of the current PBS. The 7 categories used by Zhou [Zhou, 2000] are used. The seven phrase categories are NP - Noun phrase; VP - Verb phrase; IP - Prepositional phrase; LP - Post-position phrase; DP - Frame structure; AP - Adjective phrase; FP - Adverb phrase.
3) The length of the carrying phrase of the current PBS. The lengths are divided into five groups, which are <=5, 6-10, 11-15, 16-20 and >20, respectively.

5.3 Evaluation criteria

There is no commonly accepted measure for evaluating the performance of prosodic parsers. Wang and Hirschberg used accuracy. Accuracy reflects the average performance in both breaking and non-breaking cases. However, what we really care about is the performance in breaking cases. Furthermore, the ratio of the number of breaking samples to that of non-breaking samples greatly affects the overall accuracy. For example, 95% and 94% accuracy for English and Spanish were reported by Hirschberg and Prieto. However, from the CART prediction tree for Spanish given in their paper, we find that only about 16.4% of the total samples had breaks. That is to say if all the samples are predicted to be non-breaking, then 83.6% accuracy is still obtained. The same measure was used by Lee and Oh in their experiments on Korean. Only 85% accuracy was reported. We find the reason for the drop in accuracy is that their testing set contained many more breaking samples (37%). Several measures were used together for evaluation in Taylor and Black's study. They were breaks-correct, the ratio of correctly predicted breaks to all real breaks, junctures-correct, which is the same as the accuracy measure used by Wang and Hirschberg, and juncture-insertion, the total number of insertion errors over the number of data. Juncture-insertion is not an efficient measure. In this study, four measures were used together to evaluate performance. Precision and recall were calculated for each BI type separately, and they are defined by equation (2) and (3), respectively:

$$\Pr e_j = Count(B_{cpj}) / Count(B_{pj}) \ , \tag{2}$$

$$\mathrm{Re}\, c_j = Count(B_{cpj}) / Count(B_{rj}) \ , \tag{3}$$

where j = 0, 1, 2 or 3 denotes the type of BI, $Count(B_{pj})$ is the total number of predicted boundaries for BIj, $Count(B_{cpj})$ is the number of BIj that are predicted correctly and $Count(B_r)$ is the number of real BIj.

Overall accuracy for all BI is calculated using equation (4):

$$Accu = \sum_{j=0}^{3} Count(B_{cpj}) / \sum_{j=0}^{3} Count(B_{rj}) \ . \tag{4}$$

In our study, we found that different types of errors would reduce the naturalness of the synthesized speech to different extents. The larger the BI error, which is defined as the difference between the assigned index and the real one, the larger the decrease in quality. Therefore, an overall error cost is defined by equation (5):

$$ErrCost = \sum W_i Count(E_i) \ , \tag{5}$$

where E_i represents the case where the number of BI errors equals *i*. In our case, only three types of errors, E_1, E_2 and E_3, exist. $Count(E_i)$ is the total number of E_i errors, and W_i represents the weight for E_i. In this study, $W_1 = 0.5$, $W_2 = 1$ and $W_3 = 2$.

5.4 Results

5.4.1 Basic CART method

CART was trained with both feature sets over the same training set. Only simple questions about each individual category of each feature in the feature set were provided manually. Composite questions were constructed automatically. A composite question was formed by first growing the tree with several simple questions and then clustering the leaves into two sets [Huang *et al.*, 2001]. Multiple OR and AND were used to form a composite question for each set. In our case, the depth for search a composite question was five split. The growing of the tree stopped when 40 composite questions had been formed. We have compared the results from 20, 40 and 60 composite questions. 40 was better than 20 in most cases. However, 60 was not better than 40. Thus, 40 composite questions were used in all the training phases for CART in this study.

The four measures obtained by testing the CARTs growing from the two feature sets are listed in Table 4. Column BI3NP shows the precision and recall for BI3 at PBS without BPs. The precision and recall for BI3 in this column is more meaningful than that in column BI3. According to Table 4, feature set 2 produced 1% increase in overall accuracy and 11% decrease in the error cost, compared to set 1. Table 4 also shows that syntactic phrasal information

benefited the precision and recall results for BI2 and BI3 more. However, this improvement was achieved at the cost of using a syntactic parser in on-line systems. Furthermore, the online syntactic parse cannot always provide reliable phrasal information. The phrasal information used in this study was checked manually. If the tags generated from the syntactic parser had been used directly, much worse results would have been obtained. Thus, only feature set 1 was used in the experiments with the other two approaches.

Table 4. *The performance of prosodic boundary prediction with the basic CART method for the two feature sets.*

Feature set	Evaluation Criteria	BI0	BI1	BI2	BI3	BI3NP
Set 1	precision(%)	93.19	63.95	57.41	81.12	65.66
	recall(%)	95.92	69.1	55.68	59.43	39.47
	overall accuracy (%)	82.48				
	overall error-cost	1694.5				
Set 2	precision(%)	95.01	65.06	57.77	83.67	69.85
	recall(%)	95.98	66.13	64.18	60.22	40.64
	overall accuracy (%)	83.41				
	overall error-cost	1508				

5.4.2 Bottom-up hierarchical method

The three CARTs shown in Figure 4 were trained separately from the same training set. Only feature set 1 was used. The precision and recall results for each individual CART are listed in Table 5. The integrated results are listed in Table 6. A significant decrease in the precision and recall performance for BI1, BI2 and BI3 were observed when the outputs from the three CARTs were integrated. The reason may be that errors from PW-CART were promulgated into IMP- and INP-CART, and errors from IMP-CART were promulgated into INP-CART. Comparing Table 6 and Table 4, the same overall accuracy was obtained on feature set 1. However, a 7.2% reduction in the error-cost was achieved, which means that errors with larger BI differences were reduced.

Table 5. *The performance of each individual CART.*

	PW-CART	IMP-CART	INP-CART
Precision (%)	95.74	80.96	84.77
Recall (%)	96.15	87.68	64.90

Table 6. *The integrated results for the bottom-up hierarchical method.*

Feature set	Evaluation Criteria	BI0	BI1	BI2	BI3	BI3NP
Set 1	precision(%)	95.30	65.61	53.27	81.44	67.48
	recall(%)	95.73	58.57	65.61	62.58	44.17
	overall accuracy(%)	82.49				
	overall error-cost	1590				

5.4.3 Modified hierarchical method

The three CARTs trained as described in Section 5.4.2 were also used in this modified version. BI1 and BI2 were predicted step by step as described in the previous section. However, INP boundaries were predicted using the recursive method described in Section 4.3. The final results are listed in Table 7. Comparing Table 7 with Table 6, the precision and recall performance for BI0 and BI1 are unchanged and that for BI2 and BI3 are improved. A 0.6% increase in overall accuracy and a 5.6% reduction in the error-cost are observed. The best precision and recall performance was obtained for BI3 at PBS without BP. All these improvements show that the recursive prediction method benefits the prediction of BI3.

Table 7. *The performance of BI assignment at PBS using the modified hierarchical approach.*

Feature set	Evaluation Criteria	BI0	BI1	BI2	BI3	BI3NP
Set 1	precision(%)	95.30	65.61	54.70	82.68	71.12
	recall(%)	95.73	58.57	65.61	68.10	52.41
	overall accuracy (%)	82.99				
	Overall error-cost	1550.5				

5.5 Experiment on acceptability

While manually annotated break indices are used as a reference for evaluating the results obtained using automatic methods, they are not the only correct indices since the same sentence can be spoken in different ways by human. Two experiments were conducted to evaluate the acceptability of the mis-assigned BI.

5.5.1 Experiment 1

All the errors generated by the modified hierarchical method were presented to three subjects. If at least two of them thought that the mis-assigned break index was acceptable, then, it was considered as a felicitous error. Otherwise, it was considered as an infelicitous error. Among the 2,657 errors, only 698 (26.3%) were infelicitous.

5.5.2 Experiment 2

100 sentences in the testing set were used in this experiment. Two sets of waveforms were synthesized using a data-driven TTS system [Chu *et al*., 2001]. Set A was synthesized from the scripts with manually annotated break indices, and Set B was generated from the scripts with the automatically labeled break indices. The two versions of synthetic waveforms of one sentence formed two pairs of stimuli in the sequence AB, BA. The 200 stimuli were played to 12 subjects, who had to select one from each pair that sounded more natural. The preference rate was calculated as $P_j = count(T_j) / \sum count(T_j)$,
(6)

where $count(T_j)$ is the total number of times type T_j is preferred; j=A or B.

The preference rates for the two sets of synthetic sounds are shown in Figure 6. It can be seen that P_A was higher than P_B, but that the difference between them was not very large. This result shows that our automatic method generated rather natural break indices, which were acceptable in most cases.

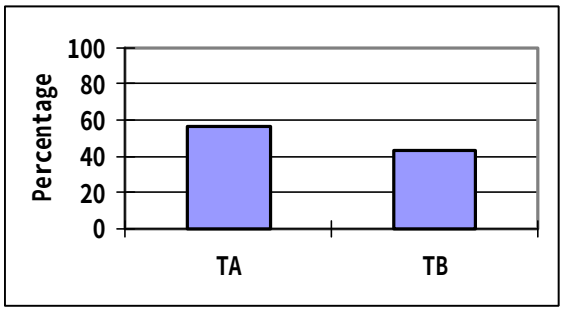


Figure 6 *Preference rates for the two types of synthesized speech. TA, speech synthesized from scripts with manually annotated BI; TB, speech synthesized from scripts with automatically generated BI.*

5.6 Discussion

Three approaches have been proposed in this section for locating the three-tier prosodic boundaries in unrestricted Mandarin texts. Because of differences in language, training and testing corpora, and the definition of prosodic constituent to be predicted, comparing results obtained in different experiments is not easy. The overall accuracy (83%) achieved in our study is not as high as that reported by Hirschberg and Prieto (95%), Lee and Oh (85%). However,

their experiments only involved making decisions between breaks and non-breaks. However, three levels of prosodic boundaries were detected in this study. Another reason for the drop in overall accuracy is the difference in the ratio of the number of break samples to that of non-break samples. In Hirschberg and Prieto' s experiment, about 16.4% of the total samples were breaks. That is to say, if all the testing data are assigned a non-break index, then 83.6% accuracy can still be obtained. In Lee and Oh' s experiment in Korean, the 37% break samples caused a significant drop in accuracy. In our testing data, only 54% were non-boundary samples. Thus, the 83% overall accuracy for the four BI is not poor performance. In all the previous studies, punctuation was used as a very important feature. However, we found that a piece of breaking punctuation almost always implied an INP boundary. Thus, predicting boundaries from non-punctuation PBS should be the focus of studies on locating boundaries. The precision and recall results obtained in several studies on BI3 at PBS with or without BP are listed in Table 8. We derived these results from the tables listed in their papers. From Table 8, the advantages of our method for PBS without BP are obvious.

Table 8. *A comparison of the performance achieved in predicting major breaks with previous results. A "+" means that the corresponding number can not be derived from the original paper.*

Comparing condition	Evaluation Criteria	Hirschberg and Prieto	Lee and Oh	Taylor and Black	Ours
BI3	Precision	92.3%	77.1%	72.3%	82.68%
	Recall	72.4%	85.4%	79.3%	68.10%
BI3NP	Precision	72.1%	+	49.3%	71.12%
	Recall	31.5%	+	54.7%	52.41%

6. Conclusion

This paper has proposed a three-tier prosodic hierarchy, which emphasizes the use of the PW instead of the LW as the basic prosodic unit. Both the surface difference and perceptual difference show the advantages of this prosodic hierarchy. Three approaches to locate the boundaries of prosody constituents in unrestricted Mandarin texts have been presented. The syntactic phrasal information produced a 1% increase in accuracy and an 11% decrease in the error cost for the basic CART method. The improved hierarchical method achieved the best performance on feature set 1. It also produced the best performance in finding INP boundaries. The two acceptability experiments revealed that only 26.3% of the mis-assigned break indices were actually infelicitous errors, and that the perceptual difference between the automatically assigned break indices and the manually annotated break indices was not large.

In this study, modified hierarchical approach, INP-CART was used to generate the probability of each PBS being a boundary. It may not be the best algorithm for generating this probability. A better algorithm may be found in our future work.

Acknowledgements

The authors thank Dr. Ming Zhou for providing the block-based robust dependency parser as a toolkit for use in this study. Thanks go to everybody who took part in the perceptual test. The authors are especially grateful to all the reviewers for their valuable remarks and suggestions.

References

Chou F. C., Tseng, C.Y. and Lee, L.S., "Automatic generation of prosodic structure for high quality Mandarin speech synthesis", *Proceeding of the Fourth International Conference on Spoken Language Processing*, 1996, Philadelphia.

Chou F. C., Tseng, C.Y. and Lee, L.S., "Automatic segmental and prosodic labeling of Mandarin speech database", *Proceeding of the Fifth International Conference on Spoken Language Processing*, 1998, Sydney.

Chu, M, Peng, H., Yang, H. and Chang, E. "Selection non-uniform units from a very large corpus for concatenative speech synthesizer", *Proceeding of the 2001 International Conference on Acoustics, Speech and Signal Processing*, 2001, Salt Lake City.

Chu, M. and Lu, S. N., "A Text-to-speech System with High Intelligibility and High Naturalness for Chinese", *Chinese Journal of Acoustics*, Vol.15, No.1, 1996, pp. 81-90.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. and Verchen, O., "The MBTOLA Project: Towards a set of high quality speech synthesizes free of use for no commercial purpose", *Proceeding of the Fourth International Conference on Spoken Language Processing*, 1996, Philadelphia.

Gee, J. P. and Grosjean, F., "Performance Structure: A Psycholinguistic and Linguistic Appraisal", *Cognitive Psychology*, Vol. 15, 1983, pp. 411-458.

Hirschberg, J. and Prieto, P., "Training intonatinal phrasing rules automatically for English and Spanish text-to-speech", *Speech Communication*, Vol. 18, 1996, pp. 281-290.

Huang, X. D., Acero, A. and Hon, H. W., "Chapter 4: Pattern Recognition", *Spoken Language Processing – A Guide to Theory, Algorithm and System Development*, 2001, Prentice Hall PTR.

Ladd, D. R. and Campbell, N., "Theories of Prosodic Structure: Evidence from Syllable Duration", *Proceeding of the 12nd International Congress of Phonetic Sciences*, 1991.

Lee, S. and Oh, Y. H., "Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems", *Speech Communication*, Vol. 28, 1999, pp. 283-300.

Liberman, M. Y. and Prince, A. S., "On Stress and Linguistic Rhythm", *Linguistic Inquiry*, Vol. 8, 1977, pp. 249-336.

Ostendorf, M. and Veilleux, N., "A hierarchical stochastic model for automatic prediction of prosodic boundary location", *Computational Linguistics*, Vol.20, No.1, 1994, pp. 27-54.

Qian, Y., Chu, M. and Peng, H., 2001. "Segmenting unrestricted Chinese text into prosodic words instead of lexical words", *Proceeding of the 2001 International Conference on Acoustics, Speech and Signal Processing*, 2001, Salt Lake City.

Selkirk, E., *Phonology and syntax: The relationship between sound and structure*, MIT press, 1984.

Shen, X. and Xu, B., "A CART based hierarchical stochastic model for prosodic phrasing in Chinese", *Proc. of the 2nd International Symposium on Chinese Language Processing*, 2000, Beijing.

Taylor, P. and Black, A.W., "Assigning phrase breaks from part-of-speech sequences", *Computer speech and language*, Vol. 12, 1998, pp. 99-117.

Veilleux, N.M., Ostendorf, M., Price, P.J. and Shattuck-Hufnagel, S., "Markov Modeling of Prosodic Phrase Structure", *Proceeding of the 1990 International Conference on Acoustics, Speech and Signal Processing*, Vol.2, 1990, pp. 777-780.

Wang, M.Q. and Hirschberg, J., "Predicting intonational phrasing from text", *Proceeding of Association for Computational Linguistics 29th annual meeting*, 1991, pp. 285-292.

Wightman, C.W. and Ostendorf, M., "Automatic labeling of prosodic patterns", *IEEE Trans. on Speech and Audio Processing*, Vol.2, No.4, 1994, pp. 469-481.

Zhou, M., "A block-based robust dependency parser for unrestricted Chinese text", *Proceeding of the second Chinese Language Processing Workshop Attached to ACL2000*, 2000, Hong Kong.