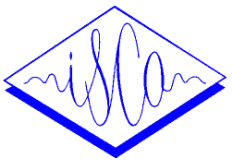

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

STUDY ON DETECTION OF PROSODIC PHRASE BOUNDARIES IN SPONTANEOUS SPEECH

SUN Hui, XU Mingxing, WU Wenhu

Center of Speech Technology,
State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology,
Tsinghua University, Beijing
[sunh, xumx, wuwh]@sp.cs.tsinghua.edu.cn

ABSTRACT

Prosodic information, which has the abilities of disambiguation, improving the parsing of the spoken language and predicting recognition errors, becomes more and more important in speech recognition and understanding, especially in spontaneous speech. In this paper, we investigate the detection of the phrase boundaries by prosodic features in the domain-specified Chinese spontaneous speech. The ultimate goal is to use these detected boundaries in the recognition and understanding component of *EasyFlight*, a Chinese spoken dialogue system for querying and booking flight tickets, to improve the system performance. In the experiment, we use more than 30 prosodic features for detecting four types of the prosodic phrase boundaries and obtain a correct detection rate over 85%.

1. INTRODUCTION

Word, phrase and paragraph are the basic elements for people to express themselves, however, when people explain their opinions they add some other important information, mainly prosody. Prosodic information, which has the abilities of disambiguation, improving the parsing of the spoken language and predicting recognition errors, becomes more and more important in speech recognition and understanding, especially in the spontaneous speech.

In a spoken dialogue system (SDS), users speak to system using spontaneous speech, then the front-end of the SDS gives the recognition results containing only word sequences without syntactic structure information such as the punctuation marks e.g. commas and full stops in written language. Such recognition results make many difficulties in understanding utterances.

Prosodic phrase boundaries can be helpful to solve the above problem because there is a strong correlation between prosodic phrase boundaries and syntactic phrase boundaries [1]. That is to say, prosodic phrase boundaries can play an important role in understanding utterance as punctuation marks do in written language. What's more, prosodic phrase boundaries also have the ability of disambiguation. In fact, there are more ambiguous phenomena in Chinese than in English. See the following sentences:

"**她看见爸爸 | 很高兴。**" (She is so happy to see her father) Vs.
"**她看见 | 爸爸很高兴**" (She sees that her father is so happy.)

These two sentences have the same word sequence, but they are definitely different in the meaning: first sentence means that the girl is happy, while the other one means her father is happy. The prosodic phrase boundaries in different place (marked by symbol '|') can help to distinguish two meanings. So we believe that prosodic phrase boundaries can be very useful in understanding Chinese spontaneous speech.

In this paper, we investigate the detection of the phrase boundaries by prosodic features in the domain-specified Chinese spontaneous speech. The ultimate goal is to use these detected boundaries in the recognition and understanding component of *EasyFlight*, a Chinese spoken dialogue system for querying and booking flight tickets, to improve the system performance.

Although only two types of prosodic phrase boundaries are used in some dialogue systems [2], the characters of our spontaneous dialogue corpus used in the experiment, in which the utterances are relatively short, make us believe that more detailed prosodic phrase boundaries will improve usefulness of the detected boundaries. So four different types of prosodic phrase boundaries were defined and used in our data labeling, i.e. B0 for normal boundary, B1 for prosodic word boundary, B2 for intermediate boundary and B3 for international boundary, and all of them are to be detected.

The prosodic boundaries can be marked by some acoustic cues, such as preboundary lengthening, pauses, boundary tones and change of speaking rate [3]. Since these cues usually can be reflected by different duration, F0 and energy features, we use them for detecting prosodic phrase boundaries. In the experiment, we use more than 30 prosodic features for detecting four types of the prosodic phrase boundaries and obtain a correct detection rate over 85%.

The paper is organized as follows: In Section 2 experiment data and the method of labeling prosodic phrase boundaries are given.

Then the prosodic features used in the detection of phrase boundaries are presented in detail in Section 3. Experiment results and discussion are presented in Section 4. Finally, some conclusions are given in Section 5.

2. DATA LABEL

2.1 Experiment Data

The experiment here uses a spontaneous spoken dialogue corpus, which is collected for evaluating *EsayFlight* system. Some representative and frequently used sentences are selected from the real human-to-human dialogue data, which cover many sentence types in this domain including flight tickets booking and all kinds of querying, e.g. time querying, price querying and so on. According to the text of these selected sentences, twelve native speakers are required to record the sentences in spontaneous style instead of reading them, and each speaker records 100 sentences.

2.2 Data Label

As for labeling prosodic phrase boundaries, there are two methods.

One of them is strictly dependent on the prosodic criteria, which use many prosodic features such as F0 contour, Energy contour etc. This kind of label method can give accurate prosodic phrase boundaries, however, it needs much time. In addition, prosody is variable and can be affected not only by syntactic phrase boundaries, but also by many other factors, e.g. speaking style, speakers' state etc. So accurate labeling of prosodic phrase boundaries is lack of robustness and may be not so good.

Based on above reasons, we choose another method, which labels the boundaries mainly based on the syntactic phrase boundaries because there is strong correlation between prosodic phrase boundaries and syntactic phrase boundaries.

In the experiment, data are labeled mainly according to the syntactic criteria by using the wave files and their transcriptions. At the same time, some spoken phenomena are considered. Four different types of prosodic phrase boundaries were defined and used in our data labeling (Figure 1 is some examples of label result):

- B0 for normal boundary
- B1 for prosodic word boundary
- B2 for intermediate boundary
- B3 for intonational boundary .

The labeling of prosodic phrase boundaries is on the basis of knowing syllable boundaries in the sentence, and here B0 means syllable boundary. In fact B0 is not a real prosodic boundary, so prosodic boundaries mentioned later in this paper mean boundary type B1, B2 and B3. B0 is a default boundary type, which is not marked in Figure 1. B1 represent the prosodic word boundary. Prosodic word is the basic element in the prosodic hierarchical structure, and it often indicate two-syllable words or three–syllable words in Chinese. B2 means intermediate boundary. Intermediate boundary is composed of one or several prosodic words. There is little pause at intermediate boundary. Finally, B3 represents intonational boundary, which contains one or several intermediate boundaries. At intonational boundary there is obvious pause.

1. 八点 $_{B1}$五十五 $_{B1}$那班 $_{B2}$是什么 $_{B1}$机型 $_{B3}$？
2. 八点 $_{B1}$左右 $_{B2}$有哪些 $_{B3}$？
3. 从北京 $_{B1}$到乌鲁木齐的 $_{B1}$票价 $_{B2}$是多少 $_{B3}$？
4. 到北京 $_{B2}$都有 $_{B1}$哪些航班 $_{B3}$？
5. 对 $_{B3}$订 $_{B1}$三张票 $_{B3}$。
6. 今天上午 $_{B2}$到深圳的 $_{B1}$航班 $_{B2}$有哪些 $_{B3}$？
7. 你好 $_{B3}$请问 $_{B2}$到上海的 $_{B1}$航班 $_{B2}$都有 $_{B1}$几点的 $_{B3}$？

Figure 1: Some examples of the label result

3. PROSODIC FEATURES

The prosodic boundaries are often marked by pause, preboundary lengthening and F0 reset. Generally, there are silence periods at a prosodic boundary. The duration of the syllable before the boundary may be longer than this syllable's duration at other place in the sentence. F0 reset is another important cue for prosodic boundaries. F0 of each syllable in the same prosodic hierarchical structure will decline, and then F0 is reset after the prosodic boundary (see figure 2, F0 reset is between "tell me" and "result").

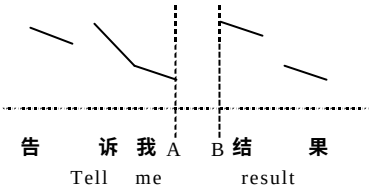

Figure 2: An illumination of some F0 related features

3.1 Duration related features

Duration related features mainly describe the pause character at the prosodic boundaries and the preboundary lengthening phenomena.

In the experiment, the primary duration related features include:

- **Silence duration**: the duration of the silence after a hypothetical prosodic boundary
- **Syllable duration**: the duration of the syllable before or after a hypothetical boundary
- **Syllable duration difference**: the duration of the syllable before a boundary minus the duration of the syllable after a boundary

3.2 F0 related features

Pitch is one of the most primary prosodic information. It is an important feature for detecting prosodic phrase boundaries, e.g. F0 reset is a good indication of prosodic boundary.

In the experiment, we get the original frame-based F0 by "get_f0" function of ESPS/WAVES (with default parameter settings, such as frame length is 0.01s), which is an autocorrelation-based pitch tracker. Besides the original F0 value, linear regression coefficients of frame-based F0 in a syllable are computed for later use. In detail, the main F0 related features are as follows:

- **F0 reset:** F0 reset at the hypothetical prosodic boundary
- **F0 slope:** One of the linear regression coefficients, which reflect the trend of the F0 contour in a syllable
- **F0 slope difference:** The difference between the F0 slopes between two consecutive syllables (the syllables before and after the hypothetical boundary)
- **F0 onset and F0 offset** (see *Figure 2*): the F0 value at point A is the F0 offset of the syllable "**我**"; the F0 value atpoint B is the F0 onset of the syllable "**结**"。
- **F0 mean:** the mean of the F0 value in a syllable.
- **F0 mean difference:** the difference between F0 means of two consecutive syllables.
- **F0 max:** the maximum of the F0 in a syllable
- **F0 min:** the minimum of the F0 in a syllable
- **F0 range:** F0 range in a syllable, which is the F0 max minus F0 min

3.3 Energy related features

Though energy related features are not so important as the duration features and F0 features, they are also useful in the detection of the prosodic boundaries.

On the basis of the frame-based energy value, we mainly use the following features:

- **Energy mean:** the mean value of the energy in a syllable
- **Energy mean difference:** the difference between the energy mean of the two consecutive syllables
- **Energy max:** the maximum of the energy in a syllable
- **Energy mean difference:** the difference between the energy max of the two consecutive syllables

3.4 Other features

In addition, some other features are also useful in this task, e.g. position information of hypothetical prosodic boundaries. A hypothetical boundary closer to the beginning of the sentence is less likely to be a prosodic boundary.

Here, features that reflect the position information are mainly as follows:

- Distance to begin (in second)
- Distance to begin (in word)
- Distance to End (in second)
- Distance to End (in word)

3.5 Feature normalization

Prosodic features such as duration, F0 and energy are less robust mainly because of their variability across speakers. The experiment results are far away from good when using the above features without normalization (totally 25 features)[4]. So we normalize some features according to a certain speaker by the following formula

$$\bar{f}_i = \frac{f_i - m_i}{s_i}$$

where f_i is the features before normalization, m_i is the mean of f_i, and s_i is the variance of f_i.

The original features together with some normalized features, totally 38 features, are used in the experiment.

4. EXPERIMENT RESULTS AND DISCUSSION

4.1 CART decision tree

As mentioned before, the experiment is on the basis of knowing syllable boundaries in the sentences, that is to say the task of detection of the prosodic phrase boundaries change to the task of classifying each known syllable boundaries into one of four boundary types defined in Section 2.

In the experiment, we choose CART [5] (classification and regression trees) as the classifier for several reasons. CART decision trees make no assumptions about the shape of feature distributions, so there is no need to convert features to some standard scale. The main reason is that CART decision trees can offer the importance of the each feature, and then we can know which features are most useful in the classification of the prosodic boundaries and further optimize features.

4.2 Experiment results

In our corpus, the number of each boundary type is not equal. Among them, B0 accounts for 60% of the total boundaries, B1 is 8.4%, B2 is 17% and B3 is 12.6%. Because the ultimate goal is to use prosodic boundaries in the understanding component of SDS, the boundary types of B1, B2, and B3 need to be better

recognized. So we suppose that all types have equal probability in the training procedure of CART.

There are totally 12 speakers' data. The experiment uses 11 speakers' as training set and another one's data as test set. Table 1, and Table 2 give the classification result of training set and test set.

Table 1a. The classification results of the training set (overall correct rate is 99.2%)

Predict / Real	B0	B1	B2	B3	Correct rate(%)
B0	5109	45	33	4	98.420
B1	11	1439	2	0	99.105
B2	0	3	688	1	99.422
B3	2	0	0	1068	99.813

Table 1b. The classification results of the test samples in training set (overall correct rate is 97.4%)

Predict / Real	B0	B1	B2	B3	Correct rate(%)
B0	1295	15	13	3	97.662
B1	8	338	5	0	96.296
B2	3	5	168	1	94.915
B3	2	0	0	269	99.262

(Test samples in Table 1b means the data that CART use to adjust the produced decision trees. These test samples are part of the training set. In the experiment, 20% of the training set is used as test samples.)

Table 2. The classification results of the test set (overall correct rate is 89.4%)

Predict / Real	B0	B1	B2	B3	Correct rate(%)
B0	583	4	6	0	98.314
B1	3	75	86	0	45.732
B2	2	0	77	0	97.468
B3	1	0	0	121	99.180

Table 2 indicates that correct rate of the test set reach 89.4%. The data in test set belongs to the new speaker whose data are not contained in the training set. So this result is considerably good and shows that our features and CART classifier have satisfying classification ability.

4.3 Discussion

Although the overall correct rate is 89.4%, some problems exist in the results of Table 2. We can see that that correct rate of B1 is bad, only about 46%. In the test set, many B1 boundaries are wrongly recognized as B2 mainly for several reasons. Firstly, B1 boundary and B2 boundary may be similar in prosodic features. For example, in sentences "**八点** B1 **左右** B2 **有哪些**": B1 and B2 can be distinguished by syntactic criterion, but the whole sentence is so short that there is no distinction between them in prosodic features. Secondly, we have mentioned before that the four boundary types do not have the amount of the samples. The data set of B1 is the smallest, so not having enough data to train is another reason for above problems.

At the same time of training, CART gives the importance of features used in the experiment. Of all the features, distance to end, silence duration, and F0 reset is the most important, and this result is consistent with the result of Chinese prosody research [6].

5. CONCLUSION

This paper presents a method for detecting the prosodic phrase boundaries in Chinese spontaneous speech, and the result is impressive. From the analysis of the result, we find that the prosodic features used to detect boundaries can be further optimized to achieve higher performance. This will be done in the near future.

6. REFERENCES

[1] Price, P., M. Ostendorf, S. Shattuck-Hufnagel and C. Fong (1991) "The use of prosody in syntactic disambiguation", *Journal of the Acoustic Society of American*, 90: 2956-2970 , 1991.

[2] Kompe R., Kiesling A., Niemann H., Noth E, Batliner A., Schachtl S., Ruland T., Block H. "Improving Parsing of Spontaneous Speech with the Help of Prosodic Boundaries." *In Proc. ICASSP'97*, volume 2, pp. 811, Munich, Germany, 1997.

[3] Colin W. Wightman, Mari Ostendorf, "Automatic Labeling of Prosodic Patterns", *IEEE Transactions on speech and audio processing*, vol. 2, NO. 4. October 1994.

[4] Hun S. "Study on Prosodic Boundary Detection and Dialogue Act Classification with the Help of Prosody in spontaneous speech", *Master thesis*, Tsinghua University, 2002

[5] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., "Classification and Regression Trees." *Wadsworth and Brooks, Pacific Grove, CA*. 1984

[6] Wang P., Yang Y.F., Liu S.N., "Acoustic features of Chinese prosodic hierarchical boundary structure", *The Proceeding of 5th national Conference On Modern Phonetics*, P161-165, Beijing, 2001