Acoustic and Linguistic Information Based Chinese Prosodic Boundary Labelling

Jianhua Tao

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
jhtao@nlpr.ia.ac.cn

Abstract

The paper analyzes both acoustic and linguistic features with different Chinese prosodic boundaries. Then a rule-learning approach was used to do the prosodic boundary labelling. In the paper the prosodic boundaries are classified into four levels, full intonational boundary with strong intonational marking with/without lengthening or change in speech tempo, prosodic phrase boundary with rather weak intonational marking, prosodic word boundary and phone foot boundary. Candidate acoustic and linguistic features related to prosodic boundary were extracted from the corpus to establish an example database. Based on this, a series of comparative experiments is conducted to collect the most effective features from the candidates. Results show that the selected candidates characterize the boundary features efficiently. Final experiments show that rule-learning approach introduced in the paper can achieve better prediction accuracy than the non-rule and RNN based methods and yet retain the advantage of the simplicity and understandability

1. Introduction

When people make speech communication, the information they exchanged includes not only the speech wave of the phones but also the structure of how the speech wave is organized. Here, we say that the speech is organized to a certain structure means that, each sentence of the speech was divided into several blocks by breaks, and each block include many phones, which has a certain duration mode. Such structured information was commonly called by prosodic rhythm. It cover the features of duration, intensity and pitch, which reflect speaking rate, accent and tone. Prosodic rhythm is very important both for the naturalness of the utterance but also for understanding the utterance. It makes it possible to divide a long utterance to some short prosodic phrases, which are more suitable for understanding by people or processing by computer. The break mode of the utterance provide important cue for syntactic disambiguation. For these reasons, research on prosodic rhythm is widely noticed in the field of speech synthesis and speech understanding. Prosodic phrase boundary location is a basic problem in the field of prosodic rhythm research.

A lot of methods have been introduced to predict prosodic phrase such as Classification and Regression Tree (Wang and Hirschberg, 1992, Yao and Min, 2001, Shen and Tao, 2003), Hidden Markov Model (Paul and Alan, 1998), Recurrent Neural Network (Ying and Shi, 2001). They pointed out that there is a tight relationship between the syntactic structure and the prosodic structure. In their work, they try to map prosodic boundaries with lots of linguistic information, such as part of speech, word length, sentence length, position, etc. It works efficiently in some TTS systems. But it is well known that the syntactic structure is not the only factor to determine the prosodic structure. Many others also studied the acoustic parameters. Li (2000) presented some statistic result of prosody on dialogue. The syllabic duration, accent and F0 range for stressed and unstressed syllable are statistically analyzed respectively. Lin (2000) showed the relation between breaks and prosodic structure. He pointed out that there are two types break can be apperceived in mandarin speech: break with silent pause and break with filled pause. Pause is always created by major break, the syllable before the break has an elongate duration and the pitch has a transition from the syllables before the break to the syllable followed. And distinguish between minor and major break is the range of F0.

Since both acoustic and linguistic information provide important cue in prosodic phrase boundary detection, the principle idea in our work is to combine them together and build them into the prosodic boundary labelling system. The whole paper was organized as following. Section 2 introduces the corpus used in the paper. Both acoustic and linguistic features related to prosody boundaries are analyzed. Section 3 describes the method for prosody boundaries labelling and evaluation. In section 4, candidate acoustic and linguistic features related to prosodic boundary were extracted from the corpus to establish an example database. Based on this, a series of comparative experiments is conducted to collect the most effective features from the candidates. Section 5 presents some evaluation and discussion of the labelling system.

2. Corpus and features

2.1. Data Corpus

To do the research, a large mandarin speech corpus, designed for synthesis and labelled with prosodic ties, is used in our research. The corpus contains 601 sentences and around 9000 syllables. It was read by 4 speakers, in which two are men and two are women. All are standard mandarin speakers. The speech was labelled with syllable in SAMPLA-C system, and labelled prosody accent, boundary and tone in C-ToBI system. In the corpus, prosodic boundary was labelled by B0, B1, B2, B3. They are,

- B3: full intonational boundary with strong intonational marking with/without lengthening or change in speech tempo.
- B2: prosodic phrase boundary with rather weak intonational marking.
- B1: prosodic word boundary.
- B0: phone foot boundary (default, not marked explicitly).

Normally, B3 related to the sentence mark such as comma, full stop, etc. It could be easily determined by them. Therefore, the following work will only be focused on the labelling of B0, B1 and B2.

2.2. Features

Normally, acoustic features are extracted from the specific speech signal interval that belongs to the prosodic unit, describing its specific prosodic properties, and can be fed directly into a prosodic boundary classifier. Within this group we can further distinguish as follows.

- $F0_{\max,t}$, $F0_{\min,t}$, $F0_{mean,t}$, $F0_{range,t}$: Minimum, maximum, mean and range of fundamental frequency (F0) of the syllable previous to current break.
- $\Delta F0_{\max,t}$, $\Delta F0_{\min,t}$, $\Delta F0_{mean,t}$, $\Delta F0_{range,t}$: Differential of minimum, maximum, mean and range of fundamental frequency (F0) in the specific context on break t. For example, $\Delta F0_{\max,t} = F0_{\max,t} - F0_{\max,t-1}$
- DF_t: deviation of F0 in break t.

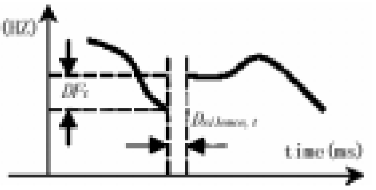

Figure 1 *The F0deviation and silence in break t*

- $D_{silence,t}$: Duration of silence between syllable t and t+1.
- $D_{B0,t}$, $D_{B1,t}$, $D_{B2,t}$, $D_{B3,t}$: Distance from the last break B0, B1, B2 and B3.
- R: Speaking rate.
- $E_{mean,t}$: Mean energy of the syllable previous to current break
- $\Delta E_{mean,t}$: Differential of mean energy of the syllable in the specific context on the time axis. $\Delta E_{mean,t} = E_{mean,t} - E_{mean,t-1}$

Table 1 lists the mean values and standard deviation of the acoustic features from the corpus.

Table 1 *The mean value and standard deviation of acoustic features in one speaker's data*

Feature	**B0**		**B1**		**B2**	
	Mean	**Deviation**	**Mean**	**Deviation**	**Mean**	**Deviation**
$D_{silence,t}$ (ms)	4.1	18	16.2	22	26.1	67
$F0_{\max,t}$ (HZ)	260	79	243	81	182	75
$F0_{\min,t}$ (HZ)	121	42	119	61	118	36
$F0_{mean,t}$ (HZ)	230	84	210	83	180	76
$F0_{range,t}$ (HZ)	130	68	109	64	78	58
$\Delta F0_{\max,t}$ (HZ)	18	31	-2.3	26	-5	24
$\Delta F0_{\min,t}$ (HZ)	6	23	0.2	21	3	20
$\Delta F0_{mean,t}$ (HZ)	2.3	21	-5	14	-18.1	12
$\Delta F0_{range,t}$ (HZ)	-3.2	58	-8	42	-10.3	36
DF_t (HZ)	7.8	23	27.3	32.1	52	42.1
$D_{B0,t}$ (ms)	182	79	191	89	212	91
$D_{B1,t}$ (ms)	276	170	418	280	653	343
$D_{B2,t}$ (ms)	623	426	620	345	1240	411
$E_{mean,t}$	62	12	52	11	51	12
$\Delta E_{mean,t}$	5	3	-1.2	3	-1	4

On the other hand, prosodic information is highly interrelated with 'higher' linguistic information, i.e., the underlying linguistic information strongly influences the actual realization and relevance of the measured acoustic prosodic features. In this sense, we speak of linguistic prosodic features that can be introduced from other knowledge sources, as lexicon, syntax or semantics; usually they have either an intensifying or an inhibitory on the acoustic prosodic features. The linguistic prosodic features can be further divided into two categories, lexical prosodic features and linguistic prosodic features..

Due to the lack of sophisticate method in syntactic and semantic parsing, we do not consider syntactic/ semantic prosodic features in the work, but some basic linguistic information which could be acquired easily and reliably. They are,

- T_t: Syllable tone previous to current break.
- POS_t: Part of speech of the word previous to current break.
- $L_{word,t}$, $L_{sentence}$, $L_{B1,t}$, $L_{B2,t}$: Length of previous lexicon word, length of sentence, length from the previous break B1 and B2. They behave as the limitation features.

With these parameters, it is still an open question, which prosodic features are the most relevant for the classification problems and how the features are interrelated. We therefore try to be as exhaustive as possible, and leave it to the classifier to find out the relevant features and the optimal weighting of them. This part of work will be described in section 4.

3. CART model based prosody boundary labeling

3.1. Method

To generate a system, rules always the easiest method. It has some obviously advantage of simplicity and understandability. But for a large database, the rules might be changed or, at least, have to be adapted to new data set.

The automatic classification and regression tree (CART) is an effective method to solve the problem. Suppose we have a string of acoustic and linguistic features: $f_1, f_2, \cdots f_T$, the boundary between f_t, f_{t+1} is labelled as B_0, B_1, B_2, B_3. Assume the label of a boundary is determined by its contextual linguistic information and neighboured acoustic information represented by a feature vector $\overrightarrow{F_t}$, prosodic boundaries can be viewed as a classification problem that in essence can be handled with any trained classifiers, taking the feature vector $\overrightarrow{F_t}$ as input and giving the most probable boundary label as output.

In order to reduce the computing complexity, all of features are normalize to a certain range 0~1. The stop criterion of CART is,

$$C = \sum_b \left(\frac{n_b}{n_t}\right) \times \left[\sum_c -\left(\frac{n_{bc}}{n_b}\right) \log_2 \left(\frac{n_{bc}}{n_b}\right)\right] \tag{1}$$

Where, b : Number of branches of the node, n_b : The sample instances in branch b, n_t : Total number of instances in all branches, n_{bc} : Number of instance of class c in branch b, c : Number of class. The later part in bracket of (1) is the entropy of the branch. So the stop criterion is when the C less than a certain valve, or the reduction of C less than a valve after a splitting process, otherwise split the node.

3.2. Evaluation Parameters

As a classification task, prosodic boundary prediction should be evaluated with consideration on all the boundary labels. The rules induced from examples are applied on a test corpus to predict the label of each boundary. Compared to manual labelling results, it is easy to generate a confusion matrix shown as follows:

Table 2 *Confusion matrix*

True labels	**Predicted labels**		
	B0	***B1***	***B2***
B0	C_{00}	C_{01}	C_{02}
B1	C_{10}	C_{11}	C_{12}
B2	C_{20}	C_{21}	C_{22}

C_{ij} are the amount of boundaries whose manual labelling is B_i but predicted as B_j. Then, the evaluation parameters for prosodic phrasing can be got as.

$$\mathrm{Re}c_i = C_{ii} / \sum_{j=0}^{2} C_{ij} \qquad \mathrm{Pr}e_i = C_{ii} / \sum_{i=0}^{2} C_{ij} \qquad (i, j = 0,1,2) \tag{2}$$

$\mathrm{Re}c_i$ is the recall rate of boundary label B_i. $\mathrm{Pr}e_i$ defines the precision rate of B_i.

4. Feature Selection

The purpose of the labelling scheme is not only to optimize a stand alone prosodic classification but to optimize its usefulness for acoustic and linguistic analysis in particular. There are many acoustic and linguistic cues for prosodic boundary location. But from discussion and statistic results in section 2, it shows not all of the acoustic parameters do the same effect on the prosody boundaries. Feature selection is crucial to the classification.

4.1. Acoustic Features

From table 2, it is not difficult to find that the most important feature to classify the prosodic boundary is the silence duration $D_{silence,t}$. It shows that the longer silence is usually related to higher breaks. But from the table, we still can find the deviation of silence is not very small. It means not all of the breaks comply with the above rule. To get more knowledge, another group of statistic results was got in the table 3. It shows the silence distribution of the breaks. The table convinces the above discussion.

Table 3 *Statistic results of breaks in different silence duration*

Silence **Boundary**	**<5 ms**	**5~20 ms**	**>20 ms**
B0	93.1%	5.7%	1.2%
B1	36.7%	42.1%	21.2%
B2	12.3%	28.4%	59.3%

From table 2, we still can find $F0_{mean,t}$, $\Delta F0_{mean,t}$, DF_t, $D_{B0,t}$ also do the important roles for the boundaries, they have obvious rules and the deviation of them are limited. $F0_{mean,t}$ trends to be lower and $\Delta F0_{mean,t}$ might be related to some negative values in prosodic word and phrase boundaries. The syllable duration previous to current break is usually lengthened and DF_t is enlarged before the prosodic phrase boundaries. Furthermore, $D_{B1,t}$, $D_{B2,t}$ could also be used for the time limitation, as we know, people can not do the expulsion of breath in speech for a long time without any inspiration.

Therefore, all of the selected acoustic parameters are,

$$\vec{A}_t = (D_{silence,t}, F0_{mean,t}, \Delta F0_{mean,t}, DF_t, D_{B0,t}, D_{B1,t}, D_{B2,t}) \tag{3}$$

4.2. Linguistic Features

To get more efficient linguistic parameters, all of the possible features are extracted from the corpus at each boundary to establish an example database. Based on example database, the experiments in which the parameters are added into the input vectors step by step are conducted to show which parameters are more efficient to final prediction results. The precision rate and recall rate are calculated from the training set of the corpus for each step. The results are listed in table 4.

Then we can got some results,

Tone seems to be less useful than what we thought before.

Part-of-speech and word length are the basic and useful feature.

Neighboured information is much helpful for boundary prediction.

Sentence length seems to be not important

It is very interesting that tone seems to be not as important for prosodic boundary detecting as we though usually, even Chinese is a tonal language. The labeling results of phone foot boundary are observably improved with the help of the lexical words boundaries which was induced from linguistic analysis.

Then, all of the selected parameters are,

$$\vec{S}_t = (POS_t, POS_{t+1}, L_{word,t}, L_{word,t+1}, L_{B1,t}, L_{B2,t}) \tag{4}$$

4.3. Window size selection

The above discussion has mentioned that neighbored linguistic information serves as helpful information for prosodic boundaries prediction. How about acoustic parameters? Do they have the same phenomena? To give better answering, another group of experiments was carried out, which is shown in table 6. Here, “Left one” means the window used for CART model covers one syllable previous to the break, “Left two” means the two syllables previous to the break are included, etc.

It is then obvious that neighbored information of both linguistic and acoustic parameters are really helpful for prosodic boundaries prediction, but the information in right side is less important than that in left side. For most case, window size of 2+1 (left two and right one) is enough for parsing. Larger size seems to be helpful, but it greatly enlarge the time consuming with no significant improvement on the results.

Table 4 *Results of feature selection in linguistic parameters*

Steps	**Parameters selected**	**Pre$_0$**	**Rec$_0$**	**Pre$_1$**	**Rec$_1$**	**Pre$_2$**	**Rec$_2$**
Step 0	$\vec{A}_t$	0.822	0.953	0.781	0.590	0.680	0.591
Step 1	$\vec{A}_t$, T_t	0.831	0.947	0.742	0.591	0.702	0.582
Step 2	$\vec{A}_t$, POS_t	0.922	0.971	0.851	0.763	0.798	0.749
Step 3	$\vec{A}_t$, POS_t, POS_{t+1}	0.924	0.978	0.889	0.778	0.814	0.786
Step 4	$\vec{A}_t$, POS_t, POS_{t+1}, $L_{word,t}$	0.931	0.985	0.923	0.817	0.865	0.833
Step 5	$\vec{A}_t$, POS_t, POS_{t+1}, $L_{word,t}$, $L_{word,t+1}$	0.928	0.985	0.927	0.803	0.853	0.831
Step 6	$\vec{A}_t$, POS_t, POS_{t+1}, $L_{word,t}$, $L_{word,t+1}$, $L_{sentence}$	0.918	0.979	0.893	0.774	0.827	0.787
Step 7	$\vec{A}_t$, POS_t, POS_{t+1}, $L_{word,t}$, $L_{word,t+1}$, $L_{sentence}$, $L_{B1,t}$	0.937	0.984	0.922	0.811	0.834	0.832

Step 8	$\vec{A}_t, POS_t, POS_{t+1}, L_{word,t}, L_{word,t+1}, L_{sentence}, L_{B1,t}, L_{B2,t}$	0.935	0.985	0.923	0.823	0.882	0.839

Table 5 *Result of experiments with different window size*

Steps	**Windows Length**	**Pre$_0$**	**Rec$_0$**	**Pre$_1$**	**Rec$_1$**	**Pre$_2$**	**Rec$_2$**
Step 1	Left one	0.921	0.972	0.853	0.760	0.791	0.748
Step 2	Left one, right one	0.937	0.985	0.921	0.825	0.880	0.839
Step 3	Left two, right one	0.942	0.987	0.931	0.838	0.890	0.854
Step 4	Left three, right one	0.939	0.985	0.921	0.832	0.892	0.843
Step 5	Left one, right two	0.911	0.974	0.865	0.763	0.847	0.762
Step 6	Left two, right two	0.930	0.987	0.937	0.817	0.883	0.847
Step 7	Left three, right two	0.924	0.985	0.921	0.807	0.891	0.828
Step 8	Left one, right three	0.922	0.978	0.896	0.785	0.845	0.797
Step 9	Left two, right three	0.921	0.988	0.938	0.805	0.892	0.840
Step 10	Left three, right three	0.944	0.987	0.934	0.843	0.891	0.859

5. Evaluation and Discussion

To evaluate the classification model, the corpus is divided into two parts, training set (1/4) and testing set (3/4). Training set contains all of the sentences listed in the corpus, 1/4 from each speaker. With selected feature and 2+1 window size, the testing results of four speakers' data are list as following,

Table 6 *Results by different test corpus*

Speaker	**B0**		**B1**		**B2**	
	Pre	**Rec**	**Pre**	**Rec**	**Pre**	**Rec**
M1	0.935	0.990	0.951	0.831	0.897	0.867
M2	0.953	0.980	0.884	0.866	0.933	0.829
F1	0.939	0.989	0.947	0.831	0.876	0.865
F3	0.933	0.987	0.942	0.816	0.862	0.851

The results show the high labelling accuracy. The confusion matrix is also listed here,

Table 7 *The statistic results of confusion matrix among B0, B1 and B2 from four speakers*

True labels	**Predicted labels**		
	B0	***B1***	***B2***
B0	21025	1035	260
B1	252	7259	287
B2	33	365	3212

From table 4, we know B2 was potential classified as B1, and B0 is more likely to be labelled as B1 than B2. It obeys the rules of Gaussian distribution.

Table 8 *Evaluation results from different methods*

Tests	**Pre$_0$**	**Rec$_0$**	**Pre$_1$**	**Rec$_1$**	**Pre$_2$**	**Rec$_2$**
CART + acoustic + linguistic	0.942	0.987	0.931	0.838	0.890	0.854
CART + acoustic	0.832	0.955	0.788	0.603	0.686	0.603
CART + linguistic	0.899	0.963	0.814	0.697	0.712	0.677
RULE + acoustic + linguistic	0.892	0.970	0.845	0.726	0.831	0.727
RNN + acoustic + linguistic	0.901	0.984	0.922	0.769	0.883	0.806

Table 8 presents the comparing results among CART (with/without acoustic or linguistic parameters), rule based and RNN. It is of course possible to adapt the classification to various demands, e.g., in order to get better labelling rates for the boundaries if more false alarms can be tolerated. Similar classification experiments with syntactic –prosodic boundaries are reported in (Wang and Hirschberg, 1992; Ostendorf et al., 1993), where HMMs or classification trees were used. The authors rely on perceptual–prosodic labels created on the basis of the ToBI system (Beckman and Ayers, 1994); for such labels, however, a much smaller amount of data can be obtained than in our case, cf. Section 2. Our recognition rates are high, however, that the studies cannot be compared in a strict sense because they considerably w.r.t. several factors

6. Conclusions

In the paper, we describe an effective method to generate rules for Chinese prosodic boundary labelling. The main idea is to extract appropriate features from acoustic and linguistic information and to apply rule-learning algorithms to automatically induce rules for predicting prosodic boundary labels. In order to find the most effective features, a series of feature selection experiments is conducted. The acquired rules achieve a best accuracy rate above 86% for prosodic phrase boundaries on test data and outperform the RNN or rule based methods, which justifies rule-learning as an effective alternative to prosodic phrase prediction. Features from deep syntactic, semantic or discourse information will be paid more attention to (Julia and Owen, 2001). The speech corpus will be enlarged to cover more types of text and speaking styles and more and more effort will be focused on spontaneous speech.

7. References

[1] Wang, M., Hirschberg, J., 1992. *Automatic classification of intonational phrase boundaries.* Computer Speech & Language 6 (2), 175–196.

[2] Julia Hirschberg, Owen Rambow, 2001. *Learning Prosodic Features using a Tree Representation.* Eruospeech2001.

[3] Shen Zhao, Jianhua Tao, Danling Jiang, 2003. *Chinese prosodic phrasing with extended features,* ICASSP2003.

[4] Li Aijun, Lin Maocan (2000). *Speech corpus of Chinese discourse and the phonetic research.* ICSLP2000.

[5] Li Aijun, 1999. *A national database design for speech synthesis and prosodic labelling of standard Chinese,* Proc. of oriental COCDA'99, TaiPei, TaiWan.

[6] Li Aijun, 1999. *Prosodic analysis on spoken dialogues for standard Chinese,* Proc. of the forth national modern phonetics conference.

[7] Lin Maocan, 2000. *Study on breaks and prosodic phrase in mandarin,* Chinese journal of linguistics, V2.4: PP210-217.

[8] Wang Pei, Yang Yufang, 2001. *Prosodic Structure and Syntactic Structure,* Proc. of the third international Conference on Cognitive Science, PP491-496.

[9] Paul Taylor and Alan W Black, 1998. *Assigning phrase breaks from part-of-speech sequences.* Computer Speech and Language, v12.

[10] Yao Qian, Min Chu, Hu Peng, 2001. *Segmenting unrestricted chinese text into prosodic words instead of lexical words.* ICASSP2001.