Ni CJ, Zhang AY, Liu WJ *et al.* Automatic prosodic break detection and feature analysis. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 27(6): 1184–1196 Nov. 2012. DOI 10.1007/s11390-012-1295-z

Automatic Prosodic Break Detection and Feature Analysis

Chong-Jia Ni[1] (倪崇嘉), Ai-Ying Zhang[1] (张爱英), Wen-Ju Liu[2] (刘文举), and Bo Xu[2] (徐　波)

[1] *School of Mathematic and Quantitative Economics, Shandong University of Finance and Economics, Jinan 250014, China*

[2] *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

E-mail: nichongjia@gmail.com; Ying_Z1217@163.com; {LWJ, xubo}@nlpr.ia.ac.cn

Received December 27, 2010; revised June 29, 2012.

Abstract Automatic prosodic break detection and annotation are important for both speech understanding and natural speech synthesis. In this paper, we discuss automatic prosodic break detection and feature analysis. The contributions of the paper are two aspects. One is that we use classifier combination method to detect Mandarin and English prosodic break using acoustic, lexical and syntactic evidence. Our proposed method achieves better performance on both the Mandarin prosodic annotation corpus — Annotated Speech Corpus of Chinese Discourse and the English prosodic annotation corpus — Boston University Radio News Corpus when compared with the baseline system and other researches' experimental results. The other is the feature analysis for prosodic break detection. The functions of different features, such as duration, pitch, energy, and intensity, are analyzed and compared in Mandarin and English prosodic break detection. Based on the feature analysis, we also verify some linguistic conclusions.

Keywords prosodic break, intonational phrase boundary, classifier combination, boosting classification and regression tree, conditional random field

1 Introduction

Features of spoken language which cannot be easily identified as discrete segments are variously referred as prosodic features or supra-segmental[1]. Term "supra-segmental" implies a difference between sound units (phones) and features such as pitch and tempo which are likely to be perceived as features extending over longer stretches of speech. The functions of prosody are many and fascinating. While speech-sounds, such as vowels and consonants, function mainly to provide an indication of the identity of words and the regional variety of the speaker, prosody can indicate syntax, turn-taking in conversational interactions, types of utterance, such as questions and statements, and people's attitudes and feelings[1]. They can also indicate word-identity (although only occasionally in English). The prosody of a word sequence can be described by a set of prosodic variables such as prosodic phrase boundary, pitch accent (stress), and lexical stress. Among these prosodic variables, pitch accent and intonational phrase boundary (IPB) have the most salient acoustic correlates, and may be most perceptually robust[2].

Many speech applications can benefit from corpora annotated with prosodic information, but it is very expensive and time-consuming to annotate prosody manually. Therefore, an automatic prosodic annotation algorithm will be very useful for building spoken language understanding systems.

In this paper, we propose a classifier combination method to detect prosodic event. The prosodic event that we consider is prosodic phrase boundary or break. The classifier combination model is obtained by combining different models which are modeled by using all features coming from acoustic, lexical and syntactic evidence. We use boosting classification and regression tree (CART) to model acoustic, lexical and syntactic features. The model can encode well the distributions of acoustic, lexical and syntactic features of syllable. The conditional random fields (CRFs) are more effective for acoustic, lexical and syntactic features, and commendably model the context property of syllable. We verify our proposed method through three different ways. First, we verify the method on Annotated Speech Corpus of Chinese Discourse (ASCCD)[3-4], where 90.34% prosodic break detection precision rate can be achieved and 6.09% is improved when compared with the baseline. The baseline system is the combination of two different systems, where the acoustic features of one system are modeled by neural network,

Regular Paper

Supported by the National Natural Science Foundation of China under Grant Nos. 90820303, 90820011, and the Natural Science Foundation of Shandong Province of China under Grant No. ZR2011FQ024.

and the lexical and syntactic features of the other system are modeled by decision tree. Second, we verify our proposed method on Boston University Radio News Corpus (BURNC)[5]. There are 2.95% precision rate improvement on break detection and 2.33% precision rate improvement on intonational phrase boundary detection respectively when compared with the baseline system. When compared with the previous work on BURNC, our proposed method also has different degrees of improvement. Finally, we use our proposed automatic prosodic break annotation method to label other continuous speeches. When compared with manual annotation, the concordance rate is 92.21%.

In this paper, we also analyze the effects of the duration, pitch, energy and intensity features in prosodic break detection, compare the effect of these features in ASCCD prosodic break detection and BURNC break or intonational phrase boundary detection, and get some significative conclusions.

The paper is organized as follows. Next section will describe the related work. In Section 3, we provide details on the corpora. In Section 4, the features used in Mandarin and English prosodic break detection are introduced, which include acoustic features, lexical and syntactic textual features. In Section 5, the prosodic break detection algorithm is presented. Our experiments and results are introduced in Section 6. In Section 7, we make the feature analysis and compare the differences and similarities between Mandarin and English prosodic break detection. In Section 8, based on the results of feature analysis, we discuss the differences between Mandarin and English prosodic break detection further. The final section gives a brief summary along with future research directions.

2 Related Work

Many approaches have explored the prosodic break detection at the word, syllable and vowel levels based on acoustic, lexical and syntactic information.

Initial attempts at automatic detection of prosodic events were presented in the work by Wightman *et al.*[6] and Ross and Ostendorf[7]. Wightman utilized decision tree to model acoustic evidence (such as pitch, energy and duration evidence), and combined it with a probabilistic model (bi-gram) to detect binary prosodic boundary. Their method achieved 71% accuracy for boundary detection at syllable level on BURNC. The performance of boundary detection is not better than human annotators (95%∼98% for intonational phase boundaries)[6]. Ostendorf *et al.* used a multi-level hierarchical model based on decision tree framework to predict boundary tone types. The three-way boundary tone classifier at intonation phrase level, which is identified as those segments marked with a break index value of 4 or above on Tones and Break Indices (ToBI) break index tier, could achieve 66.9% accuracy rate[7]. More recent related researches were reported by Chen *et al.*[8], Ananthakrishnan and Narayanan[9], Jeon and Liu[10], Sridhar *et al.*[11], and Chou *et al*[12]. Chen *et al.* built Gaussian mixture model (GMM) based on acoustic evidence and artificial neural network (ANN) model based on syntactic evidence at maximum likelihood framework for binary intonational phrase boundary detection, and achieved 93.07% accuracy rate at word level[8]. Ananthakrishnan and Narayanan used a maximum a posteriori (MAP) framework for prosodic event detection. They used an n-gram structure for prosodic language model, and utilized neural network (NN) to model acoustic evidence. When combining acoustic-prosodic model based on NN with lexical and syntactic prosodic model based on n-gram, they could achieve 91.61% binary prosodic phrase accuracy rate at syllable level[9]. Jeon and Liu showed that the neural network classifier achieved the best performance for modeling acoustic evidence, and support vector machines (SVMs) were more effective for lexical and syntactic evidence. The combination of the acoustic and syntactic models yielded 93.3% intonational phrase boundary detection accuracy and 91.1% break index detection accuracy[10]. Sridhar *et al.* described a maximum entropy based automatic prosody labeling framework, and applied the proposed framework to both prominence and phrase structure detection within ToBI annotation scheme. On BURNC, their proposed model achieved pitch accent and boundary tone detection accuracies of 86.0% and 93.1% respectively. The phrase structure detection through prosodic break index labeling provided accuracy of 84% on BURNC[11]. Chou *et al.* proposed an unsupervised joint prosody labeling and modeling (PLM) method for exploiting the prosody of spontaneous Mandarin speech. Many meaningful characteristics of spontaneous-speech prosody were investigated from the parameters of the well-trained prosodic models[12].

Great progresses have been made about the prosodic break detection based on acoustic, lexical and syntactic information in recent years. But there is one shortcoming in most prosodic break detection methods. The shortcoming is the independent assumption between acoustic features and lexical and syntactic features. In fact, the acoustic features and lexical and syntactic features are not independent. In this paper, our proposed method can avoid the independent assumption and achieves better experimental results when compared with other methods.

3 Corpora

Two corpora — ASCCD and BURNC, annotated with prosody are used in our experiments. ASCCD is designed for TTS and labeled with prosodic ties. The text of ASCCD contains 18 pieces of narration or argumentum. Each piece contains 2∼5 sections and 500∼600 syllables. The text was read by 10 speakers, who are M001, M002, M003, M004, M005, F001, F002, F003, F004 and F005 separately (five males and five females). The speech was annotated based on SAMPA-C system to describe sound variation phenomena, such as centralization, reduction, insertion[3]. The break indices and stress are annotated based on C-ToBI system[4]. In the corpus, prosodic boundary is labeled by 0, 1, 2, 3, 4, which stand for syllable boundary in prosodic word, prosodic word break, minor prosody phrase break, major prosody phrase break, and intonation group break respectively. Stress is labeled by 0, 1, 2 and 3, which stand for unstressed, prosodic word (PW) stress, minor prosodic phrase (MIP) stress and major prosodic phrase (MAP) stress respectively. In this paper, we only concern whether the syllable is followed by a prosodic break or not. We do not make distinction between different types of prosodic breaks further. This means that the prosodic word break, minor prosodic break, major prosodic phrase break and intonation group break are regarded as the same type of prosodic break. Table 1 lists the distribution of prosodic break in ASCCD corpus. In [13], Hu described the consistency about ASCCD prosodic break annotation. According to Hu's statistics, there are about 4 282 common words as the non-break annotation according to the annotation files coming from all 10 speakers, which means that there are about 78.41% (4 282/5 461) same non-break annotation pattern for each speaker data. For break annotation, there are more freedoms for each speaker data, which may be lead by speakers or annotators.

Table 1. Prosodic Break Distribution in ASCCD Corpus

Total	Non-Break	Break
87 586	54 614	32 972
100%	62.35%	37.65%

BURNC is used to verify our proposed method for English prosodic break detection[5]. It is a database of broadcast news style read speech that contains the ToBI-style prosodic annotations for part of the data. Data annotated with ToBI-style labels are available for six speakers (f1a, f2b, f3a, m1b, m2b, and m3b), which amounts to 3 hours of speeches. The corpus is annotated with orthographic transcription, automatically generated and hand-corrected part-of-speech (POS) tags, and automatic phone alignments. In break index tiers, the break indices range in value from 0 to 4, where 4 means intonational phrase boundary, 3 means intermediate phrase boundary, and a value less than 3 means phrase-medial word boundary. In BURNC, we take binary break and IPB detection. The values 3 and 4 are grouped to represent there is a break. The phrase boundary tones are annotated at every intermediate phrase boundary or intonational phrase boundary. All of the IPB tones are grouped into one category. Table 2 lists the statistics of BURNC. In [5], Ostendorf *et al.* listed their studies of labeler consistency on a set of three stories containing 1 002 words. They found that boundary tone agreement was 93% for 207 words marked by both labelers with an intonational phrase boundary, and agreement for the five ToBI break index levels was within the uncertainty level for 95% of 989 words. Therefore, there is high consistency about the BURNC annotation.

Table 2. Statistics of Boston University Radio News Corpus

	Female			Male		
	f1a	f2b	f3a	m1b	m2b	m3b
No. Utterances	74	164	33	72	51	24
No. Words	3 993	12 607	2 733	5 059	3 608	2 093
No. Syllables	6 562	20 700	4 422	8 144	5 904	3 354
No. IPBs	748	2 801	437	784	657	292
No. Breaks	1 116	3 914	744	1 247	986	459

In our experiment, we use another Mandarin continuous speech corpus which is provided by a project supported by the High Technology Research and Development 863 Program of China for Mandarin large vocabulary continuous speech recognition (LVCSR) system development, to implement automatic prosodic break annotations. 83 male speakers' data are employed for training (48 373 sentences, 55.6 hours) and 6 male speakers' for testing (240 sentences, 17.1 minutes).

4 Features

In the following subsection, the acoustic, lexical and syntactic features used in Mandarin and English prosodic break detection are introduced. In order to eliminate the natural variations among different speakers, some features must be normalized.

4.1 Features Used in Mandarin Prosodic Break Detection

4.1.1 Duration

The linguistic theories of prosodic break tend to consider syllable duration as one of the fundamental acoustic parameters for detecting syllable prosodic break. Our previous work also indicates that the duration of

the syllable before a prosodic break is lengthened[14]. For every syllable, we extract the following duration related features listed in Table 3.

Table 3. Duration Related Features

Feature Name	Feature Description
SilD	The silence duration after the syllable
SylDur	The duration of the syllable
SylDurRatio	The ratio between the following syllable duration and the current syllable duration
PDur	The duration of pitch discontinuing between the syllable and the following syllable

4.1.2 Pitch

At first, we use the command "To Pitch...", and set time step to 0.01 second, pitch floor to 50 Hz, pitch ceiling to 500 Hz to extract pitch contour with the help of Praa①. And then in order to reduce the effect by both inter-speaker and intra-speaker variation, we use z-score② method to normalize pitch. For each syllable, we compute the statistical features, such as minimum, maximum, range (maximum minus minimum), and mean. We also get an approximation of the pitch contour by using 5-order Legendre polynomial expansion.

Let us suppose $f(t)$ to be a pitch or energy contour (where t represents time), then the Legendre polynomial expansion of $f(t)$ can be approximated as

$$f(t) \approx \sum_{n=0}^{M} a_n P_n(t), \tag{1}$$

where

$$P_n(t) = \begin{cases} 1, & \text{if } n = 0, \\ t, & \text{if } n = 1, \\ \dfrac{2n-1}{n} t P_{n-1}(t) - \dfrac{n-1}{n} P_{n-2}(t), & \text{if } n \geqslant 2, \end{cases}$$

is the i-th Legendre polynomial, a_n is the coefficient of the expansion equation. Each coefficient in (1) represents a certain meaning and models a particular aspect of the contour, such as a_0 stands for the mean of the segment and a_1 is interpreted as the slope.

The previous work indicates that the comparison of adjacent syllable pitch is helpful for the prosodic break detection[8-12]. Therefore, we also compute these features, such as the difference between the pitch mean of the syllable and the pitch mean of the following syllable. Table 4 lists all these pitch-related features used in the prosodic break detection.

Table 4. Pitch-Related Features

Type	Feature Name	Feature Description
Pitch Statistical	Pth_Max	The maximum of the syllable pitch
	Pth_Min	The minimum of the syllable pitch
	Pth_Range	The difference between Pth_Max and Pth_Min
	Pth_Mean	The mean of the syllable pitch
Pitch Contour	Con_Pth_a0, Con_Pth_a1, Con_Pth_a2, Con_Pth_a3, Con_Pth_a4, Con_Pth_a5	The coefficient of 5-order Legendre polynomial expansion
Pitch Comparison	PDlt	The difference between the last non-zero pitch value of the syllable and the first non-zero pitch value of the following syllable
	BPDlt	The difference between the minimum pitch of the syllable and the minimum pitch of the following syllable
	TPDlt	The difference between the maximum pitch of the syllable and the maximum pitch of the following syllable
	PMDlt	The difference between the mean pitch of the syllable and the mean pitch of the following syllable
	PRatio	The ratio between the last non-zero pitch value of the syllable and the first non-zero pitch value of the following syllable

4.1.3 Energy

The methods of computing the energy-related features and pitch-related features are similar. In order to get energy values, we use the command "To Intensity ..." and set minimum pitch to 50 Hz, time step to 0.01 second to extract intensity of speech with the help of Praat. Table 5 lists the energy-related features.

4.1.4 Lexical and Syntactic Features

Predicting prosodic break from text has been studied extensively in the past due to its critical role in text-to-speech system. It has been shown that many factors can affect prosodic break placement. For Chinese words, Packard used the linguistic and cognitive approach to describe in detail[15]. For example, Chinese character "中国 (China)" is a word. It contains two syllables "中 (zhong)" and "国 (guo)". For Chinese "中国人民解放

①Boersma P, Weenink D. Praat: Doing phonetics by computer. http://www.praat.org/, May 2009.

②z-score normalization: $x_{\text{norm}} = \frac{x-\mu}{\sigma}$, where x is a value to normalize, μ and σ are mean and standard deviations which are estimated from all syllable duration, or pitch, energy and intensity for a speaker.

Table 5. Energy-Related Features

Type	Feature Name	Feature Description
Energy	Eng_Max	The maximum of the syllable energy
Statistical	Eng_Min	The minimum of the syllable energy
	Eng_Range	The difference between Eng_Max and Eng_Min
	Eng_Mean	The mean of the syllable energy
	EngRatio	The ratio between the mean of the syllable energy and the mean of the following syllable
Energy Contour	Con_Eng_a0, Con_Eng_a1, Con_Eng_a2, Con_Eng_a3, Con_Eng_a4, Con_Eng_a5	The coefficient of 5-order Legendre polynomial expansion

军 (the Chinese People's Liberation Army)", different segment word systems may generate different segments. It may be believed as a Chinese word "中国人民解放军 (the Chinese People's Liberation Army)", or three Chinese words "中国 (China)" , "人民 (People)" and "解放军 (Liberation Army)", or others. For us, we only use Chinese word segmenter to segment Chinese words. In this work, we first use Stanford Chinese word segmenter to segment Chinese word, then use Stanford postagger to get part-of-speech tags[16-18]. Table 6 lists the lexical and syntactic related features. In Table 6, the different syllables in the same Chinese word have same POS tags.

Table 6. Lexical and Syntactic Related Features

Feature Name	Feature Description
BSeg	Whether the syllable is the boundary of Chinese character or not
Tone	The tone of the syllable
ID	The identification of the syllable
PosTag	The Pos tag of the syllable
Wen	The number of syllables that the Chinese character contains
Hdis	The number of syllables between the syllable and the beginning of Chinese character
Tdis	The number of syllables between the syllable and the ending of a Chinese character

We also compute lexical and syntactic related features in the contextual window. We choose the two syllables before and after the current syllable as the contextual window. So we add "P_" before the feature name in order to represent the lexical and syntactic of the previous syllable, and add "F_" before the feature name in order to represent the lexical and syntactic of the following syllable.

So far, we have listed all these features used in Mandarin prosodic break detection. There are 51 features (4 duration-related, 15 pitch-related, 11 energy-related and 21 lexical and syntactic related features) in total.

4.2 Features Used in English Prosodic Break and IPB Detection

The features that we use in English prosodic break and IPB detection on BURNC almost are the same as the features listed in [10]. Table 7 lists part features used in English prosodic break and IPB detection. In order to reduce the effect by both inter-speaker and intra-speaker variation, both values of pitch and energy are normalized (z-value) with utterance specific means and variances.

We also compute these lexical and syntactic features at the three previous and two next contextual windows, and also add "P_" before the feature name in order to represent the lexical and syntactic of the previous syllable, add "F_" before the feature name in order to represent the lexical and syntactic of the following syllable.

So far, we have listed all these features used in English prosodic break detection. There are 42 features (10 pitch-related, 10 energy-related, 4 duration-related and 18 lexical and syntactic related) in total.

5 Classifiers

The combination of different classifiers is often utilized for the prosodic events detection, which can combine different information sources and different modeling methods, and compound the advantage of different models.

In [10], Jeon and Liu listed (2)∼(5) that are often used for prosodic break detection. We cite directly and list these equations below.

The most likely sequence of prosodic break $P^* = \{p_1^*, p_2^*, \ldots, p_n^*\}$ is

$$P^* = \arg\max p(P|A,S) \tag{2}$$

$$\approx \arg\max p(P|A)p(P|S) \tag{3}$$

$$\approx \arg\max \prod_{i=1}^{n} p(p_i|\boldsymbol{a}_i)^{\lambda} p(p_i|\phi(\boldsymbol{s}_i)) \tag{4}$$

$$\approx \arg\max \lambda \sum_{i=1}^{n} \log(p(p_i|\boldsymbol{a}_i)) + \sum_{i=1}^{n} \log\big(p(p_i|\phi(\boldsymbol{s}_i))\big), \tag{5}$$

where $A = \{a_1, a_2, \ldots, a_n\}$ is the sequence of acoustic feature, $\boldsymbol{a}_i = (a_i^1, a_i^2, \ldots, a_i^t)$ is the acoustic feature vector corresponding to the syllable, $S = \{s_1, s_2, \ldots, s_n\}$ is the sequence of syntactic evidence, $\boldsymbol{s}_i = (s_i^1, s_i^2, \ldots, s_i^l)$ is the lexical and syntactic feature vector corresponding

Table 7. Part of Features Used in English Prosodic Break and IPB Detection

Type	Feature Name	Feature Description
Duration	dSilDur	Silence duration after the syllable
	durSyl	Duration of the syllable
	dDurRatio	Ratio between the following and the current syllable
	durVowel	Vowel duration of the syllable
Pitch Statistical	pthMax	Maximum of the syllable pitch
	pthMin	Minimum of the syllable pitch
	pthRange	Difference between Pth_Max and Pth_Min
	pthMean	Mean of the syllable pitch
Pitch Contour	pthCoef0, pthCoef1, pthCoef2, pthCoef3, pthCoef4, pthCoef5	Coefficient of 5-order Legendre polynomial expansion
Energy Statistical	engMax	Maximum of the syllable energy
	engMin	Minimum of the syllable energy
	engRange	Difference between Eng_Max and Eng_Min
	engMean	Mean of the syllable energy
Energy Contour	engCoef0, engCoef1, engCoef2 engCoef3, engCoef4, engCoef5	Coefficient of 5-order Legendre polynomial expansion
Lexical and syntactic	Pos	POS tag of the syllable
	WordInit	Number of syntactic phrases the word initiates
	WordTerm	Number of syntactic phrases the word terminates

to the syllable, $\phi(\boldsymbol{s}_i)$ is chosen such that it contains lexical and syntactic evidence from the contextual window of the current syllable, $\log(p(p_i|\boldsymbol{a}_i))$ is the acoustic-prosodic model score, $\log(p(p_i|\phi(\boldsymbol{s}_i)))$ is the syntactic-prosodic model score, and λ is a weighting between the acoustic-prosodic and syntactic-prosodic models. The acoustic-prosodic model and syntactic-prosodic model can be obtained by machine learning methods. The statistical machine learning methods, such as classification and regression tree, neural network, support vector machine, can be used to model the acoustic-related or lexical and syntactic related features, and then apply (5) to combine the acoustic-prosodic model and syntactic-prosodic models in order to acquire the final model. When modeling the acoustic-related or lexical and syntactic related features, the same method or different methods can be utilized to model different kinds of features. About the combination of different classifiers, Ghahramani and Kim explored a general framework for the Bayesian model combination in the context of classification. Their framework models the relationship explicitly between each model's output and the unknown true label[19]. In fact, (5) is a specific case of classifier combination of two models.

Features extracted from acoustic, lexical and syntactic evidence are not fully independent. In order to represent the complex relationship among features coming from different evidences, we utilize the Bayesian network to represent these relationships. Fig.1 lists the Bayesian network.

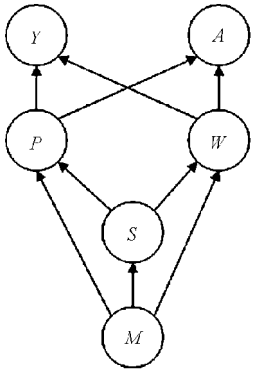

Fig.1. Bayesian network representing the complex relationship among different features.

In Fig.1, P denotes a sequence of prosody labels associated with each word in the word sequence W, describing the prosodic status of each word. S represents a sequence of labels describing the prosodic and syntactical role of each word. M denotes the meaning of the utterance which may affect the distribution of S, W and P. A is the acoustic observation sequence sampled at either frame or segmental level, and Y is the prosodic observation sequence sampled at syllable or word level. From the figure, we can find 1) A is dependent on both W and P. 2) Y is also dependent on both W and P. 3) P and W are mutually dependent. 4) W, P and S are all dependent on M.

In order to reduce the computational complexity, $p(P|A,S)$ has been simplified to $p(P|A)p(P|S)$ in (3).

We can transform (2) into (6)∼(9).

$$\begin{aligned} P^* &= \arg\max p(P|A,S) \\ &= \arg\max(\lambda \times p(P|A,S)+ \\ &\quad (1-\lambda) \times p(P|A,S)) \qquad (6) \\ &= \arg\max(\lambda \times p_1(P|A,S)+ \\ &\quad (1-\lambda) \times p_2(P|A,S)) \qquad (7) \\ &= \arg\max\Big(\frac{\lambda}{(1-\lambda)} \times p_1(P|A,S)+ \\ &\quad p_2(P|A,S)\Big) \qquad (8) \end{aligned}$$

$$= \arg\max(w \times p_1(P|A,S) + p_2(P|A,S)), \quad (9)$$

where $\frac{\lambda}{(1-\lambda)}$ is equal to w. In (8), we suppose $0 < \lambda < 1$. We give $\lambda \times p(P|A,S)$ a new symbol $\lambda \times p_1(P|A,S)$ and $(1-\lambda) \times p(P|A,S)$ another new symbol $(1-\lambda) \times p_2(P|A,S)$. This is only a deformation of (2).

From (6)∼(9), we can find that: 1) For each classifier p_1 or p_2, both the acoustic features and lexical and the syntactic features are utilized to model. 2) After modeling both the acoustic features and the lexical and syntactic features, two different classifiers are combined linearly. 3) In fact, (9) or (7) is also the combination of different classifiers, and this combination method is two levels.

(6)∼(9) are only a deformation of (2). If we hold some hypothesis, (6)∼(9) can turn out to be other methods. For example, if the same method is used to model p_1 and p_2, the method used in (9) is one type of methods, of which ensemble machine learning method is one[20]. If we do not use the same method to model p_1 and p_2, and hold the hypothesis that the acoustic features and the syntactic features are independent, (9) can be written as (5).

The differences between our proposed method and the one proposed by Jeon and Liu[10] are that: 1) Our proposed classifier combination method does not adopt the independent assumption between the acoustic features and the lexical and syntactic features; 2) Our proposed classifier combination method first models all features, including the acoustic and the lexical and syntactic features, and then combines these models by classifier combination method, while the Jeon's method first models the acoustic or lexical and syntactic information separately, and then combines these models by classifier combination method.

"Boosting" is a general method for improving the performance of the learning algorithm. It is a method for finding a highly accurate classifier on the training set, by combining "weak hypotheses", each of which needs only to be moderately accurate on the training set. It has been applied with great success to several benchmark machine learning problems by using decision trees mainly as base classifiers. AdaBoost is very popular and perhaps the most significantly historical milestone as it is the first algorithm that could be adapted for the weak learners[21]. CRFs are undirected graphical models that encode a conditional probability distribution with a given set of features. CRFs are often used for labeling or parsing sequential data, such as natural language text[22]. No matter what the word or syllable is or whether it is a prosodic break or not, it may depend on not only the current word or syllable features, but also the previous and following word or syllable features. Boosting methods can make use of the current word or syllable features greatly. CRFs methods can model the previous and following word or syllable features. We use Boosting classification and regression tree and CRFs methods to model p_1 and p_2 respectively.

6 Experiments

6.1 Experiments Setup

In our experiments, Weka implementation of C4.5 algorithm classifier (J48) is used to train decision tree model[23]. LibSVM is used to train SVM model (we choose RBF as the SVM kernel)[24]. CRF++ 0.53 is used to train CRF model③. We create 3-layer feed forward back propagation network to train multi-layer perception (MLP) model, and set the size of the hidden layer to be half of the number of input features. The Boosting CART classifier that we use in our experiments is obtained by using Weka classifier MultiBoostAB as the strong classifier, and select C4.5 decision tree (J48) as the weak classifier.

In ASCCD, we randomly select 50 sections from each speaker (totally 10 speakers, 75 sections for each speaker) to compose the training set TR, and the others make up the test set T. The ratio between the sizes of training set and testing set at sentence level is 2:1. The training set contains 58 949 syllables and the testing set contains 28 637 syllables.

In BURNC, we use the pitch information, duration information and POS tag information coming from the annotation. The energy information is extracted using Praat. The method of getting energy information is the same as in Section 3. We randomly split the utterances coming from all speakers in the corpus and perform 5-fold cross validation for binary intonational phrase boundary and binary break detection tasks. The final result is the average of the 5-fold cross validation results. In the following, we use asterisk to denote classifiers trained using all acoustic, lexical and syntactic features.

6.2 Experimental Results and Analysis

6.2.1 Acoustic Prosodic Model

First, we use decision tree and neural network to model the acoustic features. The testing results are shown in Table 8.

From Table 8, we can find that there are certain differences between the performances of the decision tree

③CRF++: Yet another CRF toolkit, http://crfpp.sourceforge.net/

Table 8. Performance of Various Acoustic Prosodic Models on ASCCD

Classifier	Category	Precision (%)	Recall (%)	*F*-Measure
Decision tree	Non-break	79.19	93.73	0.858 5
	Break	84.77	58.61	0.693 0
	Mean	81.27	80.63	0.809 5
Neural network	Non-break	80.99	93.34	0.867 3
	Break	84.95	63.18	0.724 6
	Mean	82.46	82.09	0.822 7

classifier and neural network classifier. The performance of neural network classifier is slightly better.

6.2.2 Lexical and Syntactic Prosodic Model

For lexical and syntactic features, we use three different classifiers: decision tree, SVM and CRFs. Table 9 shows the performance of various lexical and syntactic prosodic models on testing set. From Table 9, we can find that the CRFs classifier and SVM classifier achieve relatively good results. The recall rate of the prosodic break detection on ASCCD corpus obviously is better than break and IPB detection on BURNC.

Table 9. Performance of Various Lexical and Syntactic Prosodic Models on ASCCD

Classifier	Category	Precision (%)	Recall (%)	*F*-Measure
Decision tree	Non-break	90.60	77.72	0.836 7
	Break	69.78	86.45	0.772 2
	Mean	82.83	80.98	0.818 9
SVM	Non-break	92.80	90.90	0.918 0
	Break	85.20	88.10	0.866 0
	Mean	89.90	89.80	0.899 0
CRFs	Non-break	91.44	87.50	0.894 3
	Break	80.41	86.24	0.832 2
	Mean	87.33	87.03	0.871 8

From the comparison between Table 8 and Table 9, we can find the performance of the lexical and syntactic prosodic model is better than that of the acoustic prosodic model.

6.2.3 Combined Model

Table 10 shows the performance of various combined models on testing set. The value of λ is a constant, but in different classification combinations, the value may not be the same. In our experiments, we find that the value of λ in (5) ranging from 0.4 to 0.9 has good effect, and can fuse the classification results of the acoustic prosodic classifier and the syntactic-prosodic classifier. From Table 10, we can find that: 1) the combination of different knowledge obtains better performance when compared with each one alone; 2) the Boosting CART classifier can provide better classified efficiency.

Table 10. Performance of Various Combined Models on ASCCD

Classifer	Category	Precision (%)	Recall (%)	*F*-Measure
NN/Decision tree	Non-break	90.44	82.00	0.860 1
	Break	73.85	85.43	0.792 2
	Mean	84.25	83.28	0.837 6
NN/SVM	Non-break	92.76	90.89	0.918 2
	Break	85.20	88.07	0.866 1
	Mean	89.94	89.84	0.898 9
NN/CRFs	Non-break	93.18	90.28	0.917 1
	Break	84.48	88.90	0.866 3
	Mean	89.93	89.77	0.898 5
Boosting CART*	Non-break	91.35	90.10	0.907 2
	Break	83.74	85.66	0.846 9
	Mean	88.51	88.45	0.8848
CRFs*	Non-break	92.15	90.26	0.912 0
	Break	84.17	87.08	0.856 0
	Mean	89.18	89.07	0.891 2

In Table 10, Boosting CART* classifier and CRFs* classifier are obtained by using acoustic, lexical and syntactic features, and are not obtained by weighting combination through (5). In Table 10, the combined model "NN/Decision tree" means that the acoustic-based features are modeled by NN, and the lexical-based and syntactic-based features are modeled by decision tree. "NN/SVM" and "NN/CRFs" are similar.

Now, we can obtain a new classifier by weighting the combination of the Boosting CART* classifier and CRFs* classifier according to (9). The value of w in (9) is 1. This means that the weight in (7) is 0.5. We find that the choice of weight is related to the performance of different classifiers. If the performance of one classifier is better than the other classifier, the weight in (5) and (7) is greater than 0.5; if the performance of one classifier is equal to the other, the weight in (5) and (7) is about 0.5. When using this classifier to detect Mandarin prosodic break, the classifier yields 90.34% prosodic break detection precision rate and 6.09% improvement when compared with the baseline. Table 11 lists the prosodic break detection results on ASCCD.

From Table 11, we can find that the performance of the classifier classification is better than Boosting CART* or CRFs* model alone. Because CRFs* model provides the context information and is a mutual complementarity to Boosting CART* model, the performance of "Boosting CART* + CRFs*" model improves.

Table 11. Performance of Our Proposed Model on ASCCD

Classifier	Category	Precision (%)	Recall (%)	F-Measure
Baseline	Non-break	90.44	82.00	0.860 1
(NN/Decision	Break	73.85	85.43	0.792 2
tree)	Mean	84.25	83.28	0.837 6
Boosting	Non-break	93.20	91.05	0.921 1
CART*	Break	85.53	88.83	0.871 5
+ CRFs*	Mean	90.34	90.23	0.902 8

6.2.4 *Further Verification of Our Proposed Classifier Combination*

On English Prosodic Annotation Corpus – BURNC. In order to verify our proposed method and compare Mandarin and English prosodic break detection, we train and test break and IPB model on BURNC. Table 12 and Table 13 list our experimental results about break and IPB detection on BURNC respectively.

Table 12. Performance of Different Models on BURNC for Break Detection

Classifier	Category	Precision (%)	Recall (%)	F-Measure
NN/Decision	Non-break	92.80	96.98	0.948 4
tree	Break	81.53	63.81	0.715 3
(Baseline)	Mean	90.86	91.27	0.910 6
NN/SVM	Non-break	90.02	97.77	0.937 4
	Break	81.78	47.95	0.604 4
	Mean	88.61	89.19	0.889 0
NN/CRFs	Non-break	92.67	97.68	0.951 1
	Break	84.96	62.86	0.722 4
	Mean	91.34	91.68	0.915 2
Boosting	Non-break	95.04	96.71	0.958 7
CART*	Break	82.72	75.74	0.790 7
	Mean	92.92	93.10	0.930 1
CRFs*	Non-break	94.69	96.47	0.955 7
	Break	81.33	74.19	0.774 7
	Mean	92.39	92.60	0.925 0
Boosting	Non-break	95.47	97.33	0.963 9
CART* +	Break	85.79	77.76	0.815 8
CRFs*	Mean	93.81	93.96	0.938 8

From Tables 12 and 13, we can find that: 1) Our proposed method can obtain better effect. The recall rate of our proposed method on break or IPB detection improves a lot compared with the method which combines the acoustic prosodic model with lexical and syntactic model by utilizing (5), such as NN/decision tree, NN/SVM. 2) The recall rates of the combined models, such as NN/decision tree, NN/SVM, NN/CRFs, are low although the precision rates of these models were not low. From this side, we can get the conclusion that if we only model one type of the acoustic-related features, the recall rate may be very low. Our proposed methods which combine all features can avoid this. This also indicates that our proposed method is effective.

Table 13. Performance of Different Models on BURNC for IPB Detection

Classifier	Category	Precision (%)	Recall (%)	F-Measure
NN/	Non-IPB	93.66	98.84	0.961 8
Decision tree	IPB	84.86	49.22	0.622 9
(Baseline)	Mean	92.64	93.07	0.928 6
NN/SVM	Non-IPB	93.19	98.72	0.958 8
	IPB	82.28	45.19	0.582 8
	Mean	91.92	92.49	0.922 0
NN/CRFs	Non-IPB	94.45	98.83	0.965 9
	IPB	86.31	55.86	0.678 2
	Mean	93.50	93.83	0.936 7
Boosting	Non-IPB	95.41	98.59	0.969 7
CART*	IPB	85.60	63.98	0.732 2
	Mean	94.27	94.56	0.944 1
CRFs*	Non-IPB	96.21	97.76	0.969 8
	IPB	80.63	70.78	0.753 7
	Mean	94.40	94.62	0.945 1
Boosting	Non-IPB	96.04	98.62	0.973 1
CART* +	IPB	86.84	69.14	0.769 9
CRFs*	Mean	94.97	95.19	0.950 8

When compared the previous work by Jeon and Liu[10], our proposed method has 2.71% improvement for break detection and 1.67% improvement for IPB detection.

On the Mandarin Continuous Speech Corpus. We also verify our proposed method in Mandarin continuous speech corpus ("863" corpus). We have annotated all sentences in the speech corpus. In order to verify our automatic labeling methods, we select 200 sentences to annotate manually from 10 speakers randomly. Each sentence is annotated by three persons. If the syllable is annotated as a break by at least two persons, we think the syllable is a break. If the syllable is annotated as a non-break by at least two persons, we think the syllable is a non-break. We suppose that the annotation labeled by people is right. Table 14 lists the experimental results on the 200 sentences at the syllable level.

Table 14. Annotation Results on Part of Mandarin Continuous Speech Corpus

Category	Precision (%)	Recall (%)	F-Measure
Non-break	91.91	95.17	0.935 1
Break	92.64	87.89	0.902 0
Mean	92.21	92.19	0.922 0

From Table 14, we can find that our proposed method can get better effect when labeling the speech. Through labeling the prosody, we can construct the prosodic dependent large-scale continuous speech corpus. This lays foundation for the further application, such as prosody dependent speech recognition.

7 Feature Analysis

In this section, we first analyze the function of duration, pitch, energy and intensity related features in Mandarin and English prosodic break detection comprehensively, and then the importance of the single feature in Mandarin and English prosodic detection is examined one by one.

7.1 Different Feature Groups

We utilize duration, pitch, energy, lexical and syntactic related features separately to detect prosodic break on ASCCD or break and IPB on BURNC. Table 15 lists the experimental results.

From Table 15, we can find that: 1) The lexical and syntactic features have good effect for the prosodic break (or break, IPB) detection not only on ASCCD but also on BURNC. 2) For the acoustic-related features, the duration-related features are proved the most reliable, and the following are energy-related features and pitch-related features.

7.2 Single Feature

We analyze the importance of the individual type of acoustic features, such as duration, pitch, energy and intensity, and lexical and syntactic features in the prosodic break or IPB detection. We compute the difference in means between the two classes using a 2-sample t-test with unequal variance, and rank the features. When making t-test at 5% significance level on ASCCD, we find the p-value of all features is less than 5%. That is, all features in the prosodic break detection on ASCCD are important. For the break detection on BURNC, when making t-test at 5% significance level, we find that the p-value of some features is greater than 5%, including PPP_WordInit, PPP_WordTerm, PP_Pos, PP_WordInit, PP_WordTerm, F_Pos, FF_Pos. That is, these features are not so important as the other features for break detection on BURNC. For the IPB detection on BURNC, when making t-test at 5% significance level, we find that the p-value of some features is greater than 5%, including pthCoef4, pthCoef2, PPP_Pos, PPP_WordInit, PP_WordInit, PP_WordTerm, P_Pos, F_Pos, FF_Pos. Table 16 lists the top 20 features used in the prosodic break (or IPB) detection on ASCCD (or on BURNC).

Table 15. Contribution of the Different Feature Groups for Prosodic Break and IPB Detection

Corpus	Features	Type	Precision (%)	Recall (%)	F-Measure
ASCCD	Duration	Non-break	79.58	94.86	0.865 5
		Break	87.24	59.08	0.704 5
		Mean	82.44	81.51	0.819 7
	Pitch	Non-break	76.73	90.69	0.831 3
		Break	77.46	53.79	0.634 9
		Mean	77.00	76.92	0.769 6
	Energy	Non-break	76.64	92.60	0.838 7
		Break	80.88	52.57	0.637 2
		Mean	78.22	77.67	0.779 4
	Lexical	Non-break	93.49	89.56	0.914 9
	and	Break	83.62	89.53	0.864 7
	syntactic	Mean	89.81	89.55	0.896 8
BURNC	Duration	Non-break	89.20	98.05	0.934 2
		Break	80.61	40.59	0.539 9
		Mean	87.77	88.48	0.881 2
	Pitch	Non-break	85.02	99.17	0.915 5
		Break	75.10	12.51	0.214 4
		Mean	83.36	84.74	0.840 5
	Energy	Non-break	85.37	99.06	0.917 0
		Break	76.05	15.00	0.250 5
		Mean	83.82	85.06	0.844 3
	Lexical	Non-break	93.14	96.16	0.946 2
	and	Break	77.06	64.52	0.702 3
	syntactic	Mean	90.46	90.89	0.906 8
BURNC	Duration	Non-IPB	93.26	98.86	0.959 8
		IPB	82.91	43.57	0.571 3
		Mean	92.10	92.65	0.923 7
	Pitch	Non-IPB	89.97	99.50	0.944 9
		IPB	75.72	12.38	0.212 8
		Mean	88.36	89.70	0.890 3
	Energy	Non-IPB	90.26	99.28	0.945 6
		IPB	73.09	15.41	0.254 4
		Mean	88.33	89.85	0.890 8
	Lexical	Non-IPB	93.15	97.85	0.954 4
	and	IPB	71.74	43.19	0.539 2
	syntactic	Mean	90.74	91.70	0.912 2

From Table 16, we can find that the top 20 features are almost the same in BURNC break detection and IPB detection. The difference is that the feature F_WordTerm only exists in the top 20 features for BURNC break detection; the feature pthCoef3 only exists in the top 20 features for BURNC IPB detection. The order of the top 20 features is not the same in BURNC break detection and IPB detection, but the order of the top 5 features is the same. In the top 20 features, there are 4 duration-related features, 7 energy-related features, 5 pitch-related features and 4 lexical and syntactic related features. For the prosodic break detection on ASCCD, there are 3 duration-related features, 2 energy-related features, 7 pitch related features and 8 lexical and syntactic related features in the top 20

features. In the top 5 features, there are 4 lexical and syntactic related features. In the top 10 features, there are 6 lexical and syntactic related features, 2 duration-related features and 2 pitch-related features. The lexical and syntactic features are important for ASCCD prosodic break detection.

Table 16. Contribution of Different Features for Break or IPB Detection

Rank	ASCCD Prosodic Break	BURNC Break	BURNC IPB
1	BSeg	durSyl	durSyl
2	Tdis	durVowel	durVowel
3	F_Hdis	WordInit	WordInit
4	F_BSeg	engMin	engMin
5	PDur	engRange	engRange
6	SilD	dDurRatio	pthMean
7	BPDlt	engCoef4	pthMin
8	F_Tdis	dSilDur	dSilDur
9	PMDlt	pthRange	engCoef4
10	P_BSeg	pthMin	pthRange
11	EngCoef1	pthMean	dDurRatio
12	Hdis	engMean	engMean
13	PDlt	engCoef5	engCoef5
14	P_Tdis	pthCoef5	pthCoef5
15	PthMin	F_WordInit	WordTerm
16	PRatio	WordTerm	engCoef3
17	PthMean	engCoef3	F_WordInit
18	TPDlt	pthMax	engCoef0
19	EngRange	F_WordTerm	pthMax
20	SylDur	engCoef0	pthCoef3

For Mandarin prosodic break detection, the features, which are related to Chinese characters, such as BSeg, Tdis, F_Hdis, F_BSeg F_Tdis, are important. This also indicates that there is an overlapping between the prosodic break and Chinese characters boundary. For BURNC break or IPB detection, the first feature is durSyl, which is the duration of the syllable. This phenomenon is also found in previous work[10]. The context POS related features, such as P_Pos, F_Pos, FF_Pos, are not important on BURNC break or IPB detection. For ASCCD prosodic break detection, the duration-related features and the relation of adjacent syllable in pitch value side have good performance. The tone-related features do not appear in the top 20 features.

8 Discussion

There are lots of researches about how the syllables are organized into groups and the relationship between intonational phrasing and syntactic structure in language production[25-27]. Frazier *et al.* believed that prosodic phrasing is central to language comprehension, and they speculated that prosody might supply the basic skeleton that allows us to hold an auditory linguistic sequence in memory while the brain processes it[25]. Watson and Gibson evaluated several theories of how syntactic/semantic structure influences the placement of intonation boundaries in language production. They presented evidence that the intonational phrasing of a sentence is partly a function of the size of upcoming and recently processed syntactic constituents, modulated by the semantic relationships among the constituents' syntactic heads[26]. Xu and Wang investigated grouping-related F0 patterns in Mandarin by examining the effect of syllable position in a group while controlling for tone, speaking mode, number of syllables in a group, and group position in a sentence[27].

From the feature analysis in Mandarin and English prosodic break detection, similarity between Mandarin and English prosodic break detection is that the lexical and syntactic related features are important in both Mandarin and English prosodic break detection. The features, which are related with position in Chinese character, are important to Mandarin prosodic break detection. In Mandarin and English prosodic detection, the lexical and syntactic features coming from the following syllable of the current syllable are important. This means that the prosodic break mainly relates with the features coming from the following syllables of the current syllable.

Although the acoustic related features are important both in Mandarin and English prosodic break detection, the difference between Mandarin and English prosodic break detection is that the acoustic-related features in English prosodic break detection provide higher discrimination than the acoustic-related features in Mandarin prosodic break detection. Now we analyze these differences between duration, pitch and energy. 1) The duration-related features, such as "SilDur" and "PDur", are important to Mandarin and English prosodic break detection. This also indicates that the syllables grouped by two prosodic breaks have the most consistent grouped-related patterns in the syllable duration aspect. The two features "SilDur" and "PDur" are especially important to discriminate prosodic break. 2) The pitch-related features provide minor discrimination in Mandarin and English prosodic break detection when compared with the duration-related features in Mandarin and English prosodic break detection. The three features "PRatio" "PDlt" and "BPDlt" are relatively more important in Mandarin prosodic break detection. The feature "PRatio" is also F0 displacement at the prosodic break position. Of the pitch-related features, the two features "BPDlt", "PMDlt" are relatively more important in English prosodic break detection. These features, which are related to pitch contour, are not important to Mandarin and English prosodic

break detection. 3) The energy-related features provide higher discrimination both in Mandarin and English prosodic break detection when compared with the pitch-related features in both Mandarin and English prosodic break detection.

9 Conclusions and Future Work

In this paper, we developed the classifier combination method to detect Mandarin prosodic break by using acoustic, lexical and syntactic evidence. This method has the following advantages: 1) We do not adopt the independent assumption between the acoustic features and the lexical and syntactic features, and do not increase the complexity of model training at the same time; 2) The method models not only the features of the current syllable but also the contextual features of the current syllable at model level, and realizes the complementarities by taking the advantages of each model; 3) The method not only can improve the precision rate when detecting prosodic break or intonational phrase boundary but also can avoid the decrease of the recall rate on prosodic annotation corpus, especially on English prosodic annotation corpus. This method achieves the complementarities by taking the advantages of each model, and yields 90.34% prosodic break detection accuracy rate. We verified our proposed method on BURNC, and also utilized our trained model to annotate the actual Chinese sentences based on continuous speech corpus. Our proposed method got better effect in all these comparisons. In this paper, we also analyzed the features used in our experiment, and got some significant conclusions, which will be helpful for the prosodic break detection. In the future, we will refine our models and features, and exploit other methods to model acoustic, lexical and syntactic features. We will utilize the prosodic annotation continuous speech corpus to train prosody dependent phone model, and build prosody dependent speech recognition system in order to integrate prosodic information to improve the performance of the speech recognition system.

References

[1] Huang X. Acero A, Hon H W. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall, 2001.

[2] Pitrelli J, Beckman M, Hirschberg J. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proc. ICSLP*, September 1994, pp.123-126.

[3] Chen X, Li A, Sun G, Wu H *et al.* An application of SAMPA-c in standard Chinese. In *Proc. ICSLP*, Oct. 2000, pp.652-655.

[4] Li A. Chinese prosody and prosodic labeling of spontaneous speech. In *Proc. Speech Prosody,* April 2002, pp.39-46.

[5] Ostendorf M, Price P J, Shattuck-Hufnagel S. The Boston university radio news corpus. Technical Report No. ECS-95-001, Boston University, March 1995.

[6] Wightman C, Ostendorf M. Automatic labeling of prosodic patterns. *IEEE Trans. Speech and Audio Processing*, 1994, 2(4): 469-481.

[7] Ross K, Ostendorf M. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Language*, 1996, 10(3): 155-185.

[8] Chen K, Hasegawa-Johnson M, Cohen A. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic prosodic model. In *Proc. ICASSP*, May 2004, Vol.1, pp.509-512.

[9] Ananthakrishnan S, Narayanan S. Automatic prosodic even detection using acoustic, lexical and syntactic evidence. *IEEE Trans. Audio, Speech, and Language Processing*, 2008, 16(1): 216-228.

[10] Jeon J H, Liu Y. Automatic prosodic events detection using syllable-based acoustic and syntactic features. In *Proc. ICASSP,* April 2009, pp. 4565-4568.

[11] Srihar V K R, Bangalore S, Narayanan S S. Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. *IEEE Trans. Audio Speech and Language Processing*, 2008, 16(4): 797-811.

[12] Chou Y, Chiang C, Wang Y *et al.* Prosody labeling and modeling for Mandarin spontaneous speech. In *Proc. Speech Prosody*, May 2010.

[13] Hu W. Study on prosody modeling in Chinese [Ph.D. Thesis]. Institute of Automation, Chinese Academic of Sciences, 2007.

[14] Ni C, Liu W, Xu B. Automatic prosody boundary labeling of Mandarin using text and acoustic information. In *Proc. the 6th ISCSLP*, December 2008, pp.1-4.

[15] Packard J L. The Morphology of Chinese: A Linguistic and Cognitive Approach. Cambridge University Press, 2000.

[16] Tseng H, Chang P, Andrew G *et al.* A conditional random field word segmenter for sighan bakeoff 2005. In *Proc. the 4th SIGHAN Workshop on Chinese Language Processing*, October 2005, pp.168-171.

[17] Chang P, Galley M, Manning C. Optimizing Chinese word segmentation for machine translation performance. In *Proc. the 3rd Workshop on Statistical Machine Translation*, June, 2008, pp.224-232.

[18] Toutanova K, Klein D, Manning C, Singer Y. Feature rich part-of-speech tagging with a cyclic dependency network. In*Proc. HLT-NAACL*, May 2003, pp.173-180.

[19] Kim H, Ghahramani Z. Bayesian classifier combination. In *Proc. the 15th Int. Conf. Artificial Intelligence and Statistics*, April 2012, pp.619-627.

[20] Sun X. Pitch accent prediction using ensemble machine learning. In *Proc. the 2nd ICSLP*, September 2002, pp.953-956.

[21] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997, 55(1): 119-139.

[22] Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. the 8th International Conference on Machine Learning*, June 2001, pp.282-289.

[23] Hall M, Frank E, Holmes G *et al.* The WEKA data mining software: An update. *SIGKDD Explorations Newsletter*, 2009, 11(1): 10-18.

[24] Chang C, Lin C. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011, 2(3), Article No.27.

[25] Frazier L, Carlson K, Clifton C Jr. Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences*, 2006, 10(6): 244-249.

[26] Watson D, Gibson E. The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, 2004, 19(6): 713-755.

[27] Xu Y, Wang M. Organizing syllables into groups: Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics*, 2009, 37(4): 502-520.

Chong-Jia Ni is a lecturer at Shandong University of Finance and Economics. He got his Ph.D. degree in engineering from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences at 2011. His research interest covers machine learning, speech recognition, and speech synthesis.

Ai-Ying Zhang is a lecturer at Shandong University of Finance and Economics. She is a Ph.D. candidate at Institute of Remote Sensing Application, Chinese Academy of Sciences. Her research interest covers machine learning and digital signal processing.

Wen-Ju Liu is a professor at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interest covers speech recognition, speech synthesis and computational auditory science analysis.

Bo Xu is a professor at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interest covers multiple media content management, speech recognition, speech synthesis and statistical machine translation.