Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 26 (2012) 127–148

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

From English pitch accent detection to Mandarin stress detection, where is the difference?[☆]

Chongjia Ni [a,b], Wenju Liu [a,*], Bo Xu [a]

[a] *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[b] *School of Statistics and Mathematics, Shandong University of Finance, Jinan 250014, China*

Received 5 August 2010; received in revised form 7 September 2011; accepted 13 September 2011
Available online 5 October 2011

Abstract

Although English pitch accent detection has been studied extensively, there relatively a few works explore Mandarin stress detection. Moreover, the comparison and analysis between Mandarin stress detection and English pitch accent detection have not been touched for such counterpart tasks. In this paper, we discuss Mandarin stress detection and compare it with English pitch accent detection. The contributions of the paper are two aspects: one is that we use classifier combination method to detect Mandarin stress and English pitch accent by using acoustic, lexical and syntactic evidence. Our proposed method achieves better performance on both the Mandarin prosodic annotation corpus—ASCCD and the English prosodic annotation corpus—Boston University Radio News Corpus (BURNC) when compared with the baseline system. We also verify our proposed method on other prosodic annotation corpus and continuous speech corpus. The other is the feature analysis. Duration, pitch, energy and intensity features are compared for Mandarin stress detection and English pitch accent detection. Based on the analysis of prosodic annotation corpora, we also verify some linguistic conclusions.

Keywords: Mandarin stress detection; Boosting classification and regression tree (CART); Conditional random fields (CRFs); Neural network (NN); Support vector machine (SVM)

1. Introduction

Prosody is a complex weave of physical, phonetic effects that is being employed to express attitude, assumptions, and attention as a parallel channel in our daily speech communication. The semantic content of a spoken or written message is referred to as its denotation, while the emotional and attentional effects intended by the speaker or inferred by a listener are part of the message's connotation. Prosody plays an important supporting role in guiding a listener's recovery of the basic messages (denotation) and a starring role in signaling connotation, or the speaker's attitude toward the message, toward the listener(s), and toward the whole communication event. From the listener's point of view, prosody consists of systematic perception and recovery of a speaker's intentions based on pauses, pitch, rate/relative duration, and loudness. Many speech applications can benefit from corpus annotated with prosodic information, such as

[☆] This paper has been recommended for acceptance by Shrikanth Narayanan.

[*] Corresponding author at: National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. Tel.: +86 10 826 1450 5.

E-mail addresses: cjni@nlpr.ia.ac.cn (C. Ni), LWJ@nlpr.ia.ac.cn (W. Liu), xubo@nlpr.ia.ac.cn (B. Xu).

0885-2308/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2011.09.002

speech understanding and speech synthesis, 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, the main prosodic event that we consider is stress (or prominence, highlighting). Stress refers to the greater perceived strength or emphasis of some syllables in a phrase. Many research studies have been done in this area at both the syllable and word level. Approaches typically combine lexical, syntactic features, such as part-of-speech, word identity and term frequency, with acoustic features derived from the speech waveform, such as duration, pitch, and energy. A variety of machine learning approaches have been used in order to model these acoustic, lexical and syntactic features.

Different classifiers combination method is often used to prosodic event detection. In this paper, we combine some methods, which have been used for English or other language pitch accent detection, to detect Mandarin stress and English pitch accent by using acoustic, lexical and syntactic evidence, and discuss the differences and the similarities between Mandarin stress detection and English pitch accent detection. We use classifier combination method, which is the combination of boosting classification and regression tree (CART) classifier and conditional random fields (CRFs) classifier, to detect Mandarin stress and English pitch accent, and verify our proposed method (an optimal classifier combination) through two different ways. One is on prosodic annotation speech corpora. On the Mandarin prosodic annotation speech corpus—ASCCD and English prosodic annotation speech corpus—Boston University Radio News Corpus (BURNC), our proposed method can achieve 81.1% Mandarin stress detection accuracy rate and 90.8% English pitch accent detection accuracy rate separately, and there are 1.8% and 3.5% improvements when compared with their baseline system, respectively. We also compare our proposed method with the previous counterpart work on the same training set and testing set, and verify our proposed method on another Mandarin prosodic annotation corpus—Coss-1 in which some of speeches data are labeled with prosody, and 52 situational dialogues between a man and a woman are contained. Our proposed method achieves better results. The other way is on the “863” continuous speech corpus, in which only a small quantity of speeches are labeled with prosody. We use our proposed automatic stress annotation method to label “863” continuous speeches. When compared with a small number of manual annotations, the concordance rate is 95.5%. In this paper, we also analyze the function of the duration, pitch, energy and intensity features in Mandarin stress detection and English pitch accent, and compare the differences and the similarities between Mandarin stress detection and English pitch accent detection. Based on the feature analysis on prosodic annotation corpus, we also verify some linguistic conclusions.

The paper is organized as follows. Next section will describe related work. In Section 3, we provide details about the corpora. In Section 4, the features used in Mandarin stress detection and English pitch accent detection are introduced, which include acoustic features, lexical and syntactic textual features. In Section 5, the stress 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 the similarities between Mandarin stress detection and English pitch accent detection. In Section 8, we discuss the differences between Mandarin stress detection and English pitch accent detection further. The final section gives a brief summary along with future research directions.

2. Related work

Many approaches have explored to pitch accent detection at the word, syllable and vowel level based on acoustic, lexical and syntactic information. These approaches can be divided into three categories according to information source utilized in pitch accent detection, namely: detecting stress or pitch accent only from the acoustic information, detecting stress or pitch accent only from text information, and detecting stress or pitch accent from both the acoustic and text information.

Detecting stress or pitch accent from the acoustic information is often used. Applying HMM and other short-frame based models to detect stress is one of these methods. Chen and Withgott utilized a supervised Hidden Markov Model to model smoothed pitch and intensity features in order to detect stress. This approach is the first of many approaches to utilize HMM and other short frame based models to detect stress [1]. Conkie used an HMM to detect pitch accent using speaker normalized pitch and energy values at 10 ms frames, with delta and delta deltas of values. The acoustic HMM achieved 82.8% accuracy rate [2]. Ananthakrishnan used a coupled HMM (CHMM) that modeled the asynchrony between different acoustic streams, to detect pitch accent. He used pitch, energy and duration features as inputs to train CHMM model. The pitch accent detection accuracy rate at word level and at syllable level, respectively, are 72.03% and

73.93% [3]. While a great number of approaches have utilized short frame acoustic features for pitch accent detection, others have extracted the acoustic information over vowels, syllables or words. And then, these features are applied to train classifiers by using supervised machine learning methods to detect whether the syllable or word is accented or not. Wightman and Ostendorf utilized decision trees to model acoustic evidence (such as pitch, energy and duration evidence) in order to detect binary pitch accent at the syllable level [4]. Ostendorf and Ross proposed a stochastic modeling framework to predict pitch accent at the syllable level. The pitch, duration and energy feature and segmental characteristic of syllable sequence were used as input of this structure [5]. Sun proposed ensemble learning methods to predict pitch accent at the syllable level. Using boosting with CART methods, he had acquired 89.90% pitch accent detection accuracy rate by using acoustic features on a single speaker from BURNC evaluation task [6]. Rosenberg and Hirschberg applied a two-stage classification technique which predicts pitch accent at rates close to human performance by using energy, pitch and duration related features. They could achieve 84.1% accuracy rate on the read portion of the Boston Directions Corpus (BDC) at the word level [7]. Chen built Gaussian mixture model (GMM) based on acoustic evidence at maximum likelihood framework for binary pitch accent detection. 77.34% accuracy rate could achieve on the leave-one-speaker-out evaluation task on BURNC at the syllable level [8]. Ananthakrishnan and Narayanan used the maximum a posteriori (MAP) framework for prosodic event detection at the syllable level. They utilized neural network (NN) to model acoustic evidence, and achieved 80.07% pitch accent detection accuracy rate when combined with a 4-g de-lexicalized prosodic language model [9]. Jeon showed that the neural network classifier achieved the best performance for modeling acoustic evidence, and support vector machines were more effective for lexical and syntactic evidence. The NN-based acoustic model yielded 83.53% pitch accent detection accuracy rate at the syllable level [10].

Predicting stress from text has been studied extensively in the past due to its critical role in text-to-speech system. In general, when approaches utilize only text features, it is for prosodic assignments, as opposed to prosodic analysis. Many approaches to prosodic assignment operate similarly. The common is to utilize some supervised statistical machine learning methods and some features derived from part-of-speech tags, syntactic chucks or syntactic parse trees to assign pitch accent locations. The differences come down to the machine learning method and the features. Hirschberg described a technique for pitch accent detection using part-of-speech information, complex nominal status and surface position information. He could achieve 76.5% accuracy rate on 3 speakers for BURNC evaluation task at the word level [11]. Ross and Ostendorf applied an HMM over decision tree posteriors to detect pitch accent from text. They utilized a multi-stage approach, first detected the pitch accent, then assigned pitch accent type, and finally assigned phrase boundary intonation. Based on the text from a single BURNC speaker (f2b), part-of-speech, prosodic phrase structure, given/new status, lexical stress information, paragraph structure were all extracted for training the decision tree model. They could predict pitch accent placement from text with 87.7% accuracy rate at the syllable level and 82.5% accuracy rate at the word level [12]. Gregory and Altun utilized conditional random fields (CRFs) to detect pitch accent only based on text related features. The features, such as part-of-speech, probabilistic variables, were applied to model pitch accent. When evaluating on Switchboard corpus, they could achieve 76.36% accuracy rate at the word level [13]. Nenkova identified a simple lexical attribute, which is remarkably successful in pitch accent detection. They defined a term named accent ratio, which is used to capture the accent rate of a given word. This feature could acquire 75.59% pitch accent detection accuracy on Switchboard corpus [14]. Fernandez and Ramabhadram explored applying conditional random fields to automatically label major and minor break indices and pitch accent by using a large set of fully automatically extracted acoustic and linguistic features. Their experimental results demonstrated the robustness of their used features have the function of reducing the amount of training data when used in a discriminative training framework. They also explored how to adapt the baseline system in an unsupervised fashion to target dataset for which no prosodic labels are available. F-measure was used to summarize performance. They could achieve 83.5% when classifying pitch accent on BURNC [15].

Combination of the acoustic information with the text information at the word level or syllable level for pitch accent or stress detection is studied extensively too. Conkie combined the acoustic prosodic model based on HMM with the syntactic prosodic model based on a stochastic finite state model for pitch accent detection, and 88.3% pitch accent accuracy rate could achieve [2]. Ananthakrishnan used a coupled HMM (CHMM) to model the multiple, asynchronous acoustic feature streams, and achieved 79.50% accuracy rate at word level and 74.84% accuracy rate at syllable level when combined with a syntactic language model [3]. Wightman and Ostendorf combined the acoustic prosodic model based on decision trees with a probabilistic model (bi-gram) for pitch accent detection. When evaluated on a single speaker from the BURNC, 81.51% accuracy rate could achieve [4]. At the stochastic modeling framework, Ostendorf

Table 1
Summary of different approaches pitch accent detection performance.

Paper	Features	Corpus	Domain	Model	Accuracy rate (%)
Conkie [2]	Acoustic	TTS and BN	Word	HMM	82.8
Conkie [2]	Acoustic, lexical and syntactic	TTS and BN	Word	HMM and stochastic model	88.3
Ananthakrishnan [3]	Acoustic	BURNC	Word	CHMM	72.03
Ananthakrishnan [3]	Acoustic, lexical and syntactic	BURNC	Word	CHMM and syntactic language model	79.5
Wightman et al. [4]	Acoustic, lexical and syntactic	BURNC	Syllable	Decision tree and HMM	81.51
Ostendorf et al. [5]	Acoustic, lexical and syntactic	BURNC	Syllable	Stochastic model	89
Sun [6]	Acoustic	BURNC	Syllable	AdaBoost	89.9
Sun [6]	Acoustic, lexical and syntactic	BURNC	Syllable	AdaBoost and bagged	92.78
Rosenberg [7]	Acoustic	BDC	Word	Ensemble machine learning	84.1
Chen [8]	Acoustic	BURNC	Syllable	GMM	77.34
Chen [8]	Acoustic, lexical and syntactic	BURNC	Syllable	GMM and NN	86.4
Ananthakrishnan et al. [9]	Acoustic	BURNC	Syllable	NN	80.07
Ananthakrishnan et al. [9]	Acoustic, lexical and syntactic	BURNC	Syllable	NN and *n*-gram	86.75
Jeon [10]	Acoustic	BURNC	Syllable	NN	83.53
Jeon [10]	Acoustic, lexical and syntactic	BURNC	Syllable	NN and SVM	89.8
Hirschberg [11]	Lexical and syntactic	BURNC	Word	CART	76.5
Ross et al. [12]	Lexical and syntactic	BURNC	Word	CART and HMM	82.5
Gregory et al. [13]	Lexical and syntactic	Switchboard	Word	CRFs	76.36
Nenkova [14]	Lexical and syntactic	Switchboard	Word	Decision tree	76.65

and Ross utilized the pitch, duration, energy feature and segmental characteristic of syllable sequence to detect pitch accent. When evaluated on a single speaker from the BURNC, Ostendorf and Ross were able to detect pitch accent with 89% accuracy rate at the syllable level [5]. Sun combined the acoustic model based on Adaboost with the lexical and syntactic model based bagged CART model for pitch accent detection, 92.78% accuracy rate could achieve on a single speaker from BURNC evaluation task at the syllable level [6]. After Chen combined the acoustic prosodic model based on Gaussian mixture model (GMM) with the syntactic prosodic model based on artificial neural network (ANN), he could achieve 86.4% pitch accent detection accuracy rate on the BURNC at the syllable level [8]. At the maximum a posteriori (MAP) framework, Ananthakrishnan and Narayanan utilized an *n*-gram structure for prosodic language model, and neural network (NN) for modeling acoustic evidence. 86.75% pitch accent detection could achieve when combing acoustic prosodic model based on NN with lexical and syntactic prosodic model based on *n*-gram at the syllable level [9]. Jeon listed the English pitch accent detection results based on various classifiers, compared different modeling methods based on Boston University Radio News Corpus, and finally drew conclusions that neural network was very efficient to model acoustic evidence, and SVM was better than other classifiers to model lexical and syntactic evidence. The combined model of acoustic and syntactic models achieved an accuracy of 89.8% in English pitch accent detection at the syllable level [10]. Table 1 lists the performance of approaches described above.

In contrast to English pitch accent detection, very few researches about Mandarin stress detection have been reported. Shao et al. [16] applied three stress prediction models (acoustic model, linguistic model and mixed model) based on artificial neural networks (ANNs) to predict Chinese Mandarin sentential stress. The result showed that the mixed model was better than the other two models, and achieved 84.3% accuracy rate. Hu et al. [17] first designed some questions to classify syllables into different categories according to its context, so each syllable has a class label. In order to overcome the problem of data sparsity and co-articulation of syllables in some classes, he used K-mean clustering to classify syllables in these classes, which contain few syllables, into other classes. Finally, for each class, he used Eq. (1) to model stress.

$$Y \times \$^{s} = CX \times \$^{s} + B \tag{1}$$

where $X \times \$^{s}$ is the feature of syllable; $Y \times \$^{s}$ is the type of syllable stress; C, B are the undetermined coefficients. His method could achieve around 81% accuracy rate. Massive progresses have been made based on acoustic, lexical

Table 2
Stress distribution in ASCCD.

Total	Unstressed	Stressed
87,586	53,656	33,930
100%	61.26%	38.74%

Table 3
The syllable distribution of different speakers in ASCCD.

Speaker	F001	F002	F003	F004	F005	M001	M002	M003	M004	M005	Common
Unstressed											
Untoned[a]	851	873	851	857	855	782	888	851	867	818	722
Normal	4322	4644	4833	4678	4610	4527	4889	3850	4698	4214	1174
Stressed	3589	3245	3073	3226	3293	3449	2988	4052	3199	3714	682

[a] Mandarin untoned syllable is the special voice variation because syllable read continuously. It not only relates to variation in segment, but also in supra-segment. It is mainly manifested in decrease in duration, narrow in pitch range, weakening in intensity. The normal syllable is normal pronunciation syllable. We believe the untoned syllable and normal syllable are unstressed syllable.

and syntactic information in the recent years. However, it is still not clear the differences and the similarities between those two applications. To address this issue, our work will analyze and compare the differences and the similarities between Mandarin stress detection and English pitch accent detection.

3. Corpora

Three corpora—ASCCD, Coss-1 and BURNC, annotated with prosody are used in our experiments. ASCCD is designed for TTS and labeled with prosody, is used in our research. 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 [18] to describe sound variation phenomena, such as centralization, reduction, and insertion. The break indices and stress were annotated based on C-ToBI system [19]. In the corpus, prosodic boundary was labeled by 0, 1, 2, 3, and 4, which stand for syllable boundary in prosodic word, prosodic word boundary, minor prosodic phrase boundary, major prosodic phrase boundary and intonation group boundary, respectively. Stress was 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 will classify the syllables into one of unstressed syllable and stressed syllable, and won't discriminate PW stress, MIP stress, and MAP stress further. Table 2 lists the distribution of stress in the corpus. Table 3 lists the syllable distribution of different speakers in the corpus.

Coss-1 is used for verifying our proposed method, in which some of speech data are labeled with prosody. The speeches with prosodic annotation are 52 situational dialogues between a man and a woman, which are about 8 min. The annotation information consists of Pinyin, mood, tone and intonation, stress and prosodic structure. Table 4 lists the distribution of stress in the corpus.

Boston University Radio News Corpus (BURNC) is used to verify our proposed method for English pitch accent detection [20]. BURNC is a database of broadcast news style read speech that contains the ToBI-style prosodic annotations for part of data. Data annotated with ToBI-style labels are available for six speakers (f1a, f2b, f3a, m1b,

Table 4
Stress distribution in Coss-1.

Total	Unstressed	Stressed
1662	1263	399
100%	75.99%	24.01%

Table 5
The statistics of Boston University Radio News Corpus.

	Female			Male		
	f1a	f2b	f3a	m1b	m2b	m3b
#Utterances	74	164	33	72	51	24
#Words	3993	12,607	2733	5059	3608	2093
#Syllables	6562	20,700	4422	8144	5904	3354
#Accents	2344	7061	1545	2786	2113	1094

m2b and m3b), which amounts to speeches of 3 h. The corpus is annotated with orthographic transcription, automatically generated and hand-corrected part-of-speech (POS) tags, pitch and automatic phone alignments information. Table 5 lists the statistics of BURNC.

In our experiment, we use Mandarin continuous speech corpus which is provided by China National Hi-Tech Project 863 for Mandarin large vocabulary continuous speech recognition (LVCSR) system development, to implement automatic stress annotations [21]. “863” continuous speech corpus contains 200 speakers (100 male, 100 female), and 520–625 sentences for each speaker. The texts are from People’s Daily, which contain 2185 continuous statements in total. This means that each sentence is repeated by multiple speakers. For long statements in all 2185 continuous statements, they are splitted into multiple sentences according to punctuation. For Mandarin large vocabulary continuous speech recognition (LVCSR) system development, each speech file includes a Chinese character text file and a Chinese Pinyin (pronunciation) text file. The dictionary contains 48,186 Chinese characters. In our experiments, we only use 83 male speakers’ data (48,373 sentences, 55.6 h) to annotate stress.

4. Features

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

4.1. Features used in Mandarin stress detection

4.1.1. Duration

The linguistic theories of prosodic stress (or prominence) tend to consider syllable duration as one of the fundamental acoustic parameters for detecting syllable stress. For every syllable, we extract the following duration features:

durSyl: the duration of current syllable (second);
normarlDurSyl: the normalized duration of current syllable;
durSilCurFol: the duration of the silence pause between the current syllable and the following syllable (second);
durSilCurPre: the duration of the silent pause between the current syllable and previous syllable (second);
silTypeCurFol: the type of the silence between the current syllable and the following syllable[1];
silTypeCurPre: the type of the silence between the current syllable and previous syllable;
ratioDurCurPre: the ratio between the duration of current syllable and the duration of previous syllable;
ratioDurCurFol: the ratio between the duration of current syllable and the duration of following syllable;
finalDur: the finals duration of the current syllable;
normalFinalDur: the normalized finals duration of the current syllable; and
ratioFinalCurPre: the ratio between the finals duration of current syllable and the finals duration of previous syllable.

[1] There are different types of silence in the SAMPA-C system. They are long silence, silence and voiced silence.

For those normalized duration related features, the *Z*-score[2] method is used. There are 11 duration related features in total.

4.1.2. Pitch

At first, we extract pitch by setting the time step to be 0.01 s, Pitch floor to be 50 Hz, Pitch ceiling to be 500 Hz to extract pitch contour with the help of Praat [22], and then in order to reduce the effect by both inter-speaker and intra-speaker variation, we use *Z*-score method to normalize pitch. For each syllable, we compute the minimum (pthMin), maximum (pthMax), range (maximum minus minimum, pthRange), mean (pthMean), root mean squared (pthRMS) and standard deviation of pitch (pthSdDev) as pitch related statistic features. We also compute pitch related dynamic features in the contextual window. The following methods are used to compute dynamic features.

Let us use P^C_{max} represent the maximum pitch in current syllable, P^C_{min} represent the minimum pitch in current syllable, P^C_{mean} represent the mean pitch in current syllable, P_{mean} represent the mean pitch in the contextual window, $P_{\text{std.dev}}$ represent the standard deviation in the contextual window, P_{max} represent the maximum pitch in the contextual window, and P_{min} represent the minimum pitch in the contextual window. So the dynamic features in contextual window can be computed using the formulas (2)–(7).

$$P^*_{\text{max}} = \frac{P^C_{\text{max}} - P_{\text{mean}}}{P_{\text{std.dev}}} \tag{2}$$

$$P^*_{\text{mean}} = \frac{P^C_{\text{mean}} - P_{\text{mean}}}{P_{\text{std.dev}}} \tag{3}$$

$$P^{**}_{\text{max}} = \frac{P^C_{\text{max}}}{P_{\text{max}} - P_{\text{min}}} \tag{4}$$

$$P^{**}_{\text{mean}} = \frac{P^C_{\text{mean}}}{P_{\text{max}} - P_{\text{min}}} \tag{5}$$

$$P^{**}_{\text{min}} = \frac{P^C_{\text{min}}}{P_{\text{max}} - P_{\text{min}}} \tag{6}$$

$$P^{**}_{\text{range}} = \frac{P^C_{\text{max}} - P^C_{\text{min}}}{P_{\text{max}} - P_{\text{min}}} \tag{7}$$

Most Chinese words are monosyllabic or disyllabic; and the previous syllable has more influence than the following syllable on stress. Therefore, we choose the previous two syllables and one following syllable of current syllable as the contextual window. There are 12 the pitch related features in total.

4.1.3. Energy

There is consensus that the energy of a word or syllable correlates with stress. How to use the energy information in the speech signal to predict stress has not yet been determined. Sluijter and van Heuven [23] showed that stress strongly correlates with the energy within a particular frequency, namely that greater than 500 Hz in Dutch. Heldner [24,25] and Fant et al. [26] examined the role of this “spectral emphasis” in read Swedish speech, and found that the relationship between the energy in a particular spectral region and the overall energy of the signal was an excellent predictor of pitch accent. Tamburini [27] showed that the energy components of the 500–2000 Hz frequency band were more predictive of prominence than those from either 0 to 500 Hz or above 2000 Hz. Rosenberg and Hirschberg [7] found that the frequency region most robust to speaker difference was between 2 and 20 bark, and using only energy features, he could predict stress in read speech with an accuracy of 81.9%. Frequency between 500 Hz and 2000 Hz is used to compute the energy related features. First, we take short time Fourier transforms (Hamming window, 0.02 s window length, 0.01 s frame shift) to speech, and then compute the energy of special frequency band which is between 500 Hz and 2000 Hz. The energy related features include the minimum (engMin), maximum (engMax), mean

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

(engMean), range (maximum minus minimum, engRange), standard deviation (engStdDev) and root mean squared (engRMS) of energy for current syllable. In order to capture the dynamic variety of the energy in the context of the current syllable, we also calculate the dynamic features of the current syllable in the contextual window. The method of computing energy dynamic features is the same as the method of computing pitch dynamic features. There are 12 energy related features in total.

4.1.4. Intensity

We also compute the intensity related features which are similar to compute pitch and energy related features. There are 12 intensity related features in total.

We acquire intensity of speech with the help of Praat [22]. In the processing, we set minimum pitch to be 50 Hz, time step to be 0.01 s.

The difference between the energy and the intensity related features is that we only compute the speech energy related features between 500 Hz and 2000 Hz, but there is not restriction when computing the intensity related features.

4.1.5. Lexical and syntactic features

Predicting stress from text has been studied extensively due to its critical role in the text-to-speech system. It has shown that many factors can affect stress placement. In this work, we first use Stanford Chinese word segmenter to segment Chinese word, then use Stanford postagger to get part-of-speech tags [28–30]. The lexical and syntactic related feature set mainly consists of the following:

sylID: the syllable with tone or the index of the syllable with tone;
T, T1, T2: the tone of the current, the previous and the following syllable;
bSeg: whether the current syllable is the boundary of word or not;
numSyl: number of syllables in the current word;
numSylFrmSta: number of syllables from the beginning of the current word boundary;
numSylFrmEnd: number of syllables to the end of the current word boundary;
ratioPstCurLen: the ratio between the position of the current syllable in the word and the length of the word; and
posTag, posTagPre, posTagFol: tag of the current, the previous and the following word.

We also compute the probabilistic variables according to training corpus, and regard these variables as features. These variables are unigram, bigram, reverse bigram, joint and reverse joint.

Let us suppose $w_i, \quad i = 1, 2 \cdots n$ represent ith syllable in the sentence (n is syllable number in the sentence). The probabilistic variables can be defined as following: The unigram variable is defined as $\log p(w_i)$, the bigram variable is defined as $\log p(w_i|w_{i-1})$, the reverse bigram variable is defined as $\log p(w_i|w_{i+1})$, the joint variable is defined as $\log p(w_{i-1}, w_i)$ and the reverse joint variable is defined as $\log p(w_i, w_{i+1})$.

There are 17 lexical and syntactic related features in total. In all of the lexical and syntactic related features, such features are categorical, including the types of the features, the syllable with tone or the index of the syllable with tone, the tone of the current, the previous and the following syllable, whether the current syllable is the boundary of word or not, and tag of the current, the previous and the following word. The others are numeric.

In order to use CRFs to model acoustic, lexical and syntactic features, each continuous feature is binned into 10 equal categories according to the range of the feature. We also tried more bins and got similar results, hence, only results binned by 10 will be reported.

Features obtained from acoustic cues, lexical and syntactic information are used for Mandarin stress detection. There are 64 the acoustic-related, lexical-related and syntactic-related features in total.

4.2. Features used in English pitch accent detection

In English pitch accent detection, we extract the following features at syllable level according to the Refs. [10,13,36]. In order to reduce the effect by inter-speaker and intra-speaker variation, both values of pitch and energy are normalized (*Z*-score) with utterance specific means and variances.

Pitch range (4 features): maximum pitch (pthMax), minimum pitch (pthMin), mean pitch (pthMean) and pitch range (difference between maximum and minimum pitch, pthRange).

Pitch contour (6 features): 6 coefficients of Legendre polynomial (pthCoef$_i$, i = 0, 1,. . .,5).

Energy range (4 features): maximum energy (engMax), minimum energy (engMin), mean energy (engMean) and energy range (difference between maximum and minimum energy, engRange).

Energy contour (6 features): 6 coefficients of Legendre polynomial (engCoef$_i$, i = 0, 1,. . .,5).

Duration (1 feature): duration of the syllable (durSyl).

Lexical and syntactic features (3 features): the syllable (sylID), lexical stress (exist or not) (bLexical) and POS tag (posTag).

For pitch contour and energy contour, we use 5-order Legendre polynomial expansion to get 6 coefficients of Legendre polynomial.

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

$$f(t) \approx \sum_{n=0}^{M} a_n P_n(t) \tag{8}$$

where $P_n(t) = \begin{cases} 1 & n = 0 \\ t & n = 1 \\ \dfrac{2n-1}{n} t P_{n-1}(t) - \dfrac{n-1}{n} P_{n-2}(t) & n \geq 2 \end{cases}$ is the n-th Legendre polynomial, a_n is the coefficient of expansion equation. Each coefficient in expansion Eq. (8) 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.

We also compute these lexical and syntactic features in the contextual window, which contains 2 previous and 2 next syllables of the current syllable. There are 15 lexical and syntactic features in total. So we can get 36 features 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 Ref. [10], Jeon listed Eqs. (9)–(12) that are often used for stress detection. So we cite directly and list these equations below.

The most likely sequence of stress $P^* = \{p_1^*, \; p_2^*, \cdots p_n^*\}$ is

$$P^* = \arg\max p(P|A, S) \tag{9}$$

$$P^* \approx \arg\max p(P|A)p(P|S) \tag{10}$$

$$P^* \approx \arg\max \prod_{i=1}^{n} p(p_i|a_i)^{\lambda} p(p_i|\phi(s_i)) \tag{11}$$

$$P^* \approx \arg\max \lambda \sum_{i=1}^{n} \log(p(p_i|a_i)) + \sum_{i=1}^{n} \log(p(p_i|\phi(s_i))) \tag{12}$$

where $A = \{a_1, a_2, \cdots, a_n\}$ is the sequence of acoustic feature, $a_i = (a_i^1, \, a_i^2, \cdots, a_i^t)$ is the acoustic feature vector corresponding to the syllable, $S = \{s_1, s_2, \cdots, s_n\}$ is the sequence of syntactic evidence, $\phi(s_i)$ is chosen such that it contains lexical and syntactic evidence from the contextual window of the current syllable, $\log(p(p_i|a_i))$ is the acoustic-stress model score, $\log(p(p_i|\phi(s_i)))$ is the syntactic-stress model score, and λ is a weighting between the acoustic-stress and syntactic-stress model. The acoustic-stress model and syntactic-stress model can be obtained by using machine learning methods. The statistical machine learning methods, such as classification and regression trees (CART), neural network (NN), support vector machine (SVM), can be used to model the acoustic related or lexical and syntactic related features, and then apply Eq. (12) to combine the acoustic-stress model and syntactic-stress model in order to form 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 explored a general framework for the Bayesian model combination in the context of classification. His framework models the relationship explicitly between each model's output and the unknown true label [31]. In fact, Eq. (12) is a specific case of classifier combination of two models.

Features extracted from the acoustic, lexical and syntactic are not fully independent. In order to reduce the computational complexity, $p(P|A, S)$ has been simplified to $p(P|A)p(P|S)$ in Eq. (10).

We can transform Eq. (9) into Eqs. (13)–(16).

$$P^* = \arg\max p(P|A, S)$$

$$P^* = \arg\max\, (\lambda p(P|A, S) + (1-\lambda)p(P|A, S)) \tag{13}$$

$$P^* = \arg\max\, (\lambda p_1(P|A, S) + (1-\lambda)p_2(P|A, S)) \tag{14}$$

$$P^* = \arg\max\, \left(\frac{\lambda}{(1-\lambda)}p_1(P|A, S) + p_2(P|A, S)\right) \tag{15}$$

$$P^* = \arg\max(wp_1(P|A, S) + p_2(P|A, S)) \tag{16}$$

where $\lambda/(1-\lambda)$ is equal to w. In Eq. (14), we suppose $0<\lambda<1$. We give $\lambda p(P|A, S)$ a new symbol $\lambda p_1(P|A, S)$, and $(1-\lambda)p(P|A, S)$ another new symbol $(1-\lambda)p_2(P|A, S)$. This is only a deformation of Eq. (9).

From Eqs. (13) to (16), we can find that (1) for each classifier p_1 or p_2, both the acoustic features and the lexical and 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, Eq. (16) or (14) is also the combination of different classifiers, and this combination method is two levels.

Eqs. (13)–(16) are only a deformation of Eq. (9). If we hold some hypothesis, Eqs. (13)–(16) 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 Eq. (16) is one type of methods, of which ensemble machine learning method is one [6]. If we don't 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, Eq. (16) can be written as Eq. (12).

The differences between our proposed method and the one proposed by Jeon 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 was the first algorithm that could adapt for the weak learners [32]. Conditional random fields (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 [33]. No matter what the word or syllable is or whether it is stressed 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 (CART) and CRFs methods to model p_1 and p_2, respectively.

6. Classification experiments

6.1. Experiments setup

In our experiments, WEKA implementation of C4.5 algorithm classifier (J48) is used to train decision tree model, and WEKA implementation of sequential minimal optimization (SMO) algorithm is used to train SVM model [34]. CRF++ 0.53 is used to train CRFs model [35]. We create two-layer multilayer perception network with a single hidden

Table 6
The performance of various acoustic-stress models on ASCCD and BURNC.

	Accuracy rate (%)	F-measure
Decision tree		
ASCCD		
Unstressed	76.7	0.795
Stressed	67.8	0.634
Mean	73.3	0.733
BURNC		
Unaccented	85.2	0.862
Accented	73.9	0.723
Mean	81.4	0.815
Neural network		
ASCCD		
Unstressed	78.3	0.786
Stressed	65.5	0.652
Mean	73.4	0.734
BURNC		
Unaccented	84.7	0.880
Accented	80.5	0.737
Mean	83.3	0.831

layer to train neural network (NN) model, in which the number of hidden unit in hidden layer is half of the number of input features, and the 2 output nodes are chosen corresponding to the stressed and the unstressed separately. The Boosting CART model that we used 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 Jeon's work [10], these classifiers, such as C4.5 decision tree, SVM, neural network, are utilized to model the acoustic or the lexical and syntactic features in English pitch accent detection on BURNC. So we decide to use these classifiers to model the acoustic or the lexical and syntactic features.

In all corpora, when constructing the training and testing sets, we guarantee that they are comprised not only of distinct speakers, but also of distinct lexical contents.

In ASCCD corpus, we randomly select some sections from each speaker to compose the training set Tr, and the others make up the testing set T. The ratio between the size of training and testing sets at sentence level is 5:1. The training set contains 72,798 syllables, and the testing set contains 14,788 syllables.

In Coss-1 corpus, we randomly select 42 dialogues for training, the other 10 dialogues for testing. We extract pitch, intensity and energy between 500 Hz and 2000 Hz from speech with the help of Praat. The setting of Praat is the same as in ASCCD corpus.

In BURNC, we use the pitch information, duration information and POS tag information coming from the annotation. The energy information is extracted by using Praat [22]. We randomly select some utterances to compose the training set and the testing set. The ratio between the size of training and testing sets at sentence level is 4:1. The training set contains 40,032 syllables, and the testing set contains 9054 syllables.

6.2. *Experimental results and analysis*

6.2.1. *The Mandarin stress detection and English pitch accent detection with acoustic-stress model*

First, we use decision tree and neural network to model the acoustic features. The experimental results are shown in Table 6.

From Table 6, we can find that: the performances of decision tree classifier and neural network classifier show no significant differences on ASCCD and BURNC. The performance of neural network (NN) is slightly better, but the performance of NN classifier in Mandarin stress detection is not better than in English pitch accent detection.

Table 7
The performance of various syntactic-stress models on ASCCD and BURNC.

	Accuracy rate (%)	F-measure
Decision tree		
ASCCD		
Unstressed	81.2	0.833
Stressed	74.4	0.711
Mean	78.6	0.786
BURNC		
Unaccented	89.6	0.880
Accented	75.3	0.778
Mean	84.7	0.845
SVM		
ASCCD		
Unstressed	83.5	0.797
Stressed	66.4	0.707
Mean	77.0	0.763
BURNC		
Unaccented	89.1	0.873
Accented	74.1	0.768
Mean	84.0	0.837
CRFs		
ASCCD		
Unstressed	81.8	0.819
Stressed	70.9	0.707
Mean	77.6	0.776
BURNC		
Unaccented	91.0	0.901
Accented	80.0	0.813
Mean	87.2	0.872

6.2.2. The Mandarin stress detection and English pitch accent detection with syntactic-stress model

For lexical and syntactic features, we utilize three different classifiers: decision tree, SVM and CRFs. Table 7 shows the performance of various syntactic-stress models on ASCCD and BURNC. From Table 7, we can find the performances of decision tree model, SVM model, and CRFs model on ASCCD are not as good as in English pitch accent detection on BURNC. The CRFs classifier achieves relatively better results than decision tree classifier and SVM classifier on BURNC.

6.2.3. The Mandarin stress detection and English pitch accent detection with combined model

Table 8 shows the performance of various combined models.

In Table 8, Boosting CART* classifier and CRFs* classifier are obtained by using the acoustic, lexical and syntactic features, and are not obtained by weighting combination through Eq. (12). 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. The “NN/SVM” and “NN/CRFs” are similar.

In Table 8, the value of λ in Eq. (12) ranging from 0.5 to 1.5 has a good effect, and can fuse the classification results of the acoustic-stress classifier and the syntactic-stress classifier. The λ of the combined models “NN/decsion tree”, “NN/SVM” “NN/CRFs” on ASCCD corpus are 0.5, 1.5 and 0.6. The λ of the combined models “NN/decsion tree”, “NN/SVM” “NN/CRFs” on BURNC corpus are 1.1, 1.5 and 0.7. The value of λ is tuned on the training set.

From Table 8, we can find that (1) the combination of different knowledge obtains better performance than each alone for all classifiers; (2) the Boosting CART classifier can provide the best classified efficiency on ASCCD, and the CRFs classifier can provide the best classified efficiency on BURNC; and (3) all classifiers used in Mandarin stress detection do not achieve the same as in English pitch accent detection.

Now, we can obtain a new classifier “Boosting CART* + CRFs*” by weighting combination of the Boosting CART* classifier and CRFs* classifier according to Eq. (16). The value of w in Eq. (16) is 1 on ASCCD and BURNC. This means that the weight in Eq. (14) is 0.5. We also find that the choice of weight is related to the performance of different

Table 8
The performance of various combined models on ASCCD and BURNC.

	Accuracy rate (%)	F-measure
NN/decsion tree		
ASCCD		
Unstressed	82.3	0.836
Stressed	74.3	0.724
Mean	79.3	0.793
BURNC		
Unaccented	88.6	0.907
Accented	84.9	0.806
Mean	87.3	0.874
NN/SVM		
ASCCD		
Unstressed	83.4	80.8
Stressed	68.3	71.4
Mean	77.6	77.3
BURNC		
Unaccented	88.9	0.902
Accented	82.7	0.802
Mean	86.8	0.869
NN/CRFs		
ASCCD		
Unstressed	82.3	0.836
Stressed	74.5	0.724
Mean	79.3	0.794
BURNC		
Unaccented	91.8	0.921
Accented	85.0	0.845
Mean	89.5	0.895
Boosting CART*		
ASCCD		
Unstressed	82.5	0.840
Stressed	75.3	0.729
Mean	79.7	0.798
BURNC		
Unaccented	92.1	0.913
Accented	82.4	0.836
Mean	88.8	0.887
CRFs*		
ASCCD		
Unstressed	83.2	0.835
Stressed	73.7	0.732
Mean	79.6	0.796
BURNC		
Unaccented	93.0	0.922
Accented	83.9	0.852
Mean	89.9	0.898

classifiers. If the performance of one classifier is better than the other classifier, the weight in Eq. (12) or (14) is greater than 0.5; if the performance of one classifier is equal to the other, the weight in Eq. (12) or (14) is about 0.5. When using “Boosting CART* + CRFs*” classifier to detect Mandarin stress and English pitch accent, this classifier yields 81.1% stress detection accuracy rate on ASCCD and 90.8% pitch accent detection accuracy rate on BURNC at syllable level. Table 9 lists the Mandarin stress and English pitch accent detection results.

From Table 9, we can find that the performance of the model by combining Boosting CART* model with the CRFs* model based on acoustic, lexical and syntactic features is better than Boosting CART* or CRFs* model alone. On ASCCD and BURNC, there are 1.8% and 3.5% improvements, respectively, when compared with their baseline system.

Table 9
The performance of our proposed classifier combination on ASCCD and BURNC.

	Accuracy rate (%)	F-measure
Baseline (NN/decision tree)		
ASCCD		
Unstressed	82.3	0.836
Stressed	74.3	0.724
Mean	79.3	0.793
BURNC		
Unaccented	88.6	0.907
Accented	84.9	0.806
Mean	87.3	0.874
Boosting CART* + CRFs*		
ASCCD		
Unstressed	86.3	0.849
Stressed	72.6	0.746
Mean	81.1	0.810
BURNC		
Unaccented	93.8	0.929
Accented	85.1	0.866
Mean	90.8	0.908

On ASCCD, we also compare the performances of ours with the previously correlative Hu's work under the same conditions (same training sets Tr1 and Tr2, same testing sets T1 and T2) [17]. Tr1 is made of the first 12 utterances from the speaker M001. There are 12 utterances in training set Tr1. Tr2 is made of the first 12 utterances from all speakers. There are 120 utterances in training set Tr2. T1 contains the last 6 utterances from the speaker M001. There are 6 utterances in testing set T1. T2 contains the last 6 utterances from all speakers. There are 60 utterances in testing set T2. In ASCCD corpus, each utterance contains 3–8 sentences. The experimental results are listed in Table 10. There are 2.16% and 0.95% absolute accuracy rate improvements to T1 and T2, respectively. In Table 10, we quote Hu's experimental results directly.

On BURNC, we also compare the performance of ours with the previous related work. When compared with Jeon's work [10], there is 1% absolute accuracy rate improvement; and when compared with Ananthakrishnan et al.'s work [9], there is 4.05% absolute accuracy rate improvement.

6.2.4. The further verification of our proposed classifier combination

In order to verify efficiency of our proposed method and performance of annotating the other speech further, we also conduct experiments on the other Mandarin corpus. First, we conduct experiments on other Mandarin corpus with prosodic annotation. Second, we conduct experiment on the Mandarin continuous speech corpus, which is not labeled with prosody.

6.2.4.1. On the Mandarin prosodic annotation corpus. On the Mandarin prosodic annotation corpus—Coss-1, we verify our proposed method. The features used in the experiment and the setting of the experiment are the same as in the experiment on the corpus—ASCCD. Table 11 lists the experiments on Coss-1 corpus.

From Table 11, we can find that our proposed classifier combination method achieves the best experimental results. When compared with the baseline system, our proposed method has 1.7% absolute accuracy rate improvement. Through the experiments on Coss-1 corpus, we can find that our proposed method is robust. In Table 11, the accuracy rate of our

Table 10
The correct rate using acoustic and text features in test set T1 and T2.

	T1	T2
Boosting CART* + CRFs*	86.36	81.95
Proposed method by Hu [17]	84.20	81.00

Table 11
The experimental results of various combined models on the Coss-1 corpus.

	Accuracy rate (%)	F-measure
Baseline (NN/decision tree)		
Unstressed	94.1	0.929
Stressed	72.6	0.756
Mean	89.1	0.889
NN/SVM		
Unstressed	93.6	0.905
Stressed	56.5	0.636
Mean	84.9	0.845
NN/CRFs		
Unstressed	94.1	0.923
Stressed	67.7	0.724
Mean	87.9	0.877
Boosting CART*		
Unstressed	95.1	0.935
Stressed	72.6	0.769
Mean	89.8	0.897
CRFs*		
Unstressed	94.6	0.923
Stressed	66.1	0.719
Mean	87.9	0.877
Boosting CART* + CRFs*		
Unstressed	94.1	0.94
Stressed	80.2	0.807
Mean	90.8	0.908

proposed method is 90.8%, which is higher than that of on ASCCD and BURNC. From the experiments on ASCCD and BURNC, we can find that the accuracy rate of our proposed method on ASCCD is lower than that of on BURNC. This also indicates that the corpus utilized in the evaluation has a significant impact on the Mandarin stress or English pitch accent detection rate in both languages.

6.2.4.2. On the Mandarin continuous speech corpus. We also verify our proposed method on Mandarin continuous speech corpus ("863" corpus). First, we extract pitch, intensity and energy between 500 Hz and 2000 Hz from the speeches with the help of Praat. The setting of Praat is the same as in ASCCD corpus. Second, with the help of automatic speech recognition system, all speeches are aligned in order to get time boundary of each syllable. With the help of Stanford Chinese word segmentation and Stanford postagger, we get POS tag information of each text file which contains Chinese characters. Third, we compute the acoustic, lexical and syntactic features listed in Section 3. Finally, we utilize the model acquired on ASCCD corpus using our proposed method to annotate the stress status of each syllable in each file in the "863" continuous speech corpus. We have annotated all sentences in the speech corpus. In order to verify our automatic labeling methods, we randomly select 120 sentences of 6 speakers to annotate manually. Each sentence is annotated by three persons. We consider the syllable is stressed, if it is annotated to stress by at least two persons. We consider the syllable is unstressed, if it is not annotated to stress by at least two persons. Suppose that the manual annotation is right, and then we can get the accuracy rate. Table 12 lists the experimental results on the 120 sentences at syllable level.

Table 12
The annotation results on part of Mandarin continuous speech corpus.

	Accuracy rate (%)	F-measure
Unstressed	94.4	0.957
Stressed	96.8	0.954
Mean	95.5	0.956

From Table 12, we can find that (1) our proposed annotation method is effective. (2) The experimental results on "863" corpus are better than on ASCCD corpus. I think that the following two reasons would lead to this phenomenon. The first one may lie in different speech style. "863" corpus is read-style speech corpus, and the speed of speech is moderate. Each of wave file in "863" corpus is a sentence. Though ASCCD corpus is read-style speech, there are lots of extra phenomena. The duration of speech wave file in ASCCD corpus is longer than that in "863" corpus, and there is long-standing silence between some syllables. Each utterance in ASCCD corpus consists of several sentences. The speed of speech in ASCCD corpus is faster than that in "863" corpus. The second reason may lie in the amount of manual annotation data. Because of time-consuming to annotate prosody manually, we only annotate seldom speech data, only 120 sentences. In "863" corpus, there are 48,373 sentences. The ratio between the number of manual annotation data and the number of all of speech data in "863" corpus is very low.

7. Feature analysis

Sluijter and van Heuven [23] believed that in English, duration was proved the most reliably correlative with stress. Overall intensity and vowel quality were the poorest cues. Spectral balance, however, turned out to be a reliable cue, close in strength to duration. Bolinger [37] hypothesized that features based on fundamental frequency (or pitch) provide higher discrimination as compared to features based on duration and energy. Kochanski et al. got the following conclusions through studying seven dialects of British and Irish English. Fundamental frequency played a minor role in distinguishing prominent syllables from the rest of the utterance. Instead, speakers primarily marked prominence with patterns of loudness and duration. All dialects and speaking styles studied in their studies shared a common definition of prominence [38]. Silipo and Greenberg also got similar conclusions in the spoken American English. Their conclusions were that fundamental frequency played a relatively minor role in the assignment of prosodic stress in casually spoken American English, and that amplitude and duration were primary acoustic parameters associated with the patterning of stress-relevant cues in spontaneous American English [39]. Rosenberg and Hirschberg detected pitch accent at word, syllable and vowel level, and their experimental results indicated that a word-based approach is superior to syllable- or vowel-based detection [40].

Chao [41] described that the formation of stress in Mandarin firstly involved expanding the tonal range and lengthening the duration and secondly raising the air stream. But Shen [42] pointed out that intonation was used to realize the sentence stress. Sentence stress and rhythmic stress would modify the top line and bottom line of tonal contours, respectively. It was pitch rather duration that contributed more to sentence stress. Wang studied the relationship between tone pattern and word stress in Mandarin. Her experimental results indicated that tone had something to do with the engendering of the unstressed syllables at lexical level and with word stress pattern in continuous speech. When appearing in the final position in a disyllable word, the syllable with the high level tone, for example tone 1, had the smallest possibility to be unstressed at lexical level, and also made itself most prominent in perception. The syllable with the low tone, for example tone 3, did the reverse [43].

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

7.1. *Different features group*

We first utilize duration, pitch, energy, intensity, lexical and syntactic related features, respectively, to train corresponding duration, pitch, energy, intensity, lexical and syntactic model on ASCCD and BURNC, and then use these models to detect stress and pitch accent on ASCCD testing set and BURNC testing set. Table 13 lists the experimental results.

From Table 13, we can find that (1) in Mandarin stress detection, the lexical and syntactic related features have better performance in detecting stress than the acoustic related features. In English pitch accent detection, the lexical and syntactic related features also have better performance in detecting pitch accent than that of the acoustic related features. (2) In Mandarin stress detection, for the acoustic related features, the duration related features prove the most reliable features, what follows are the intensity related features, the pitch related features, and the energy related features. For English pitch accent detection, the energy related features provide high discriminations, and the pitch related features do not provide higher discrimination than the energy related features. This phenomenon is also similar

Table 13
The pitch, duration, energy and intensity related information make contribution to stress and pitch accent detection.

	Accuracy rate (%)	F-measure
Duration		
ASCCD	72.7	0.730
BURNC	73.7	0.736
Intensity		
ASCCD	69.2	0.695
BURNC	–	–
Pitch		
ASCCD	69.0	0.693
BURNC	74.2	0.746
Energy		
ASCCD	67.2	0.673
BURNC	76.6	0.768
Lexical and syntactic		
ASCCD	78.8	0.789
BURNC	87.4	0.874

to the previous related work [38,39]. (3) In Mandarin stress detection, if we only use the duration or pitch or energy or intensity separately to detect syllable stress, the correct rate of stress detection is low. This also states that stress is a complicated language phenomenon. The single acoustic feature has little effect on Mandarin stress detection. In English pitch accent detection, the duration related features are important, and also provide better discrimination to pitch accent than other acoustic related features.

7.2. Single feature

In order to inspect the performance of single feature in stress or pitch accent detection more precisely, we first implement the detection tasks by using single feature, respectively, and then rank these features in descending order according to the detection accuracy rate. It is obvious that the more important in stress or pitch accent detection, the higher in the rank.

Table 14 lists the results of some features for Mandarin stress detection.

From Table 14, we find that the features, which are related with syllable position in the Chinese character, are important to Mandarin stress detection. The duration related features, such as the type of the silence between the

Table 14
The rank of features according to the correct rate in Mandarin stress detection.

Rank	Description of features	Accuracy (%)
1	Unigram probability	70.33
2	The ratio between the position of the current syllable in the word and the length of the word	70.03
3	The duration of the silent pause between the current syllable and previous syllable	69.81
4	Number of syllables to the end of the current word boundary	68.12
5	Whether the current syllable is the boundary of word or not	68.10
6	Reverse joint probability	67.37
7	Joint probability	67.28
8	Bi-gram probability	67.09
9	Reverse Bi-gram probability	66.54
10	The type of the silence between the current syllable and previous syllable	65.81
11	The syllable with tone or the index of the syllable with tone	65.60
12	The standard deviation of intensity in the syllable	64.56
13	The mean of pitch in the syllable	64.46
14	The maximum pitch in the syllable	64.28
15	The ratio between duration of the current syllable and duration of the following syllable	64.26

Table 15
The rank of features according to the correct rate in English pitch accent detection.

Rank	Description of features	Accuracy (%)
1	Whether the syllable has lexical stress or not	79.82
2	The duration of the syllable	73.70
3	The maximum energy of the syllable	71.88
4	The 3th coefficient of pitch contour Legendre polynomial expansion	71.04
5	The PosTag of the previous syllable of the current syllable	69.15
6	The energy range of the syllable	68.80
7	The 3th coefficient of energy contour Legendre polynomial expansion	68.04
8	The 1th coefficient of energy contour Legendre polynomial expansion	68.01
9	The following syllable	67.58
10	The 4th coefficient of pitch contour Legendre polynomial expansion	67.06
11	The syllable	66.96
12	The mean pitch of the syllable	66.95
13	The minimum pitch of the syllable	66.95
14	The 5th coefficient of pitch contour Legendre polynomial expansion	66.77
15	The 0th coefficient of energy contour Legendre polynomial expansion	66.70

current and previous syllable, are also important to this task. From Table 14, we still can find that the probability related features are important in detecting Mandarin stress.

Table 15 lists the result of some features for English pitch accent detection.

From Table 15, we can find that (1) in the listed 15 features, there are 11 acoustic related features and 4 lexical and syntactic related features. (2) Whether the syllable has lexical stress or not is also important in English pitch accent detection. (3) The duration of syllable provides very high discrimination in English pitch accent detection. (4) The maximum energy of the syllable is also important in English pitch accent detection.

Finally, we can summarize this section. From the feature analysis in detections of Mandarin stress and English pitch accent, the similarity between them is that the lexical and syntactic related features are important to them. For example, in Mandarin stress detection, the probability related features and the features which are related with syllable position in Chinese characters are important; and in English pitch accent detection, whether the syllable has lexical stress or not is important features. It is very reasonable that the lexical and syntactic features yield high accuracy rate on the BURNC and ASCCD because speech is constrained by the text contents. In Mandarin, most Chinese characters are monosyllable and disyllable, and the syllable position in Chinese characters has a greater relationship with Mandarin stress. Similarly, English pitch accent also has a greater relationship with the lexical stress, and the mutual information between the feature bLexical and the classification label is 0.243.

The difference between Mandarin stress detection and English pitch accent detection is that the acoustic related features in English pitch accent detection provide higher discrimination than that of in Mandarin stress detection. (1) The duration related features are important both to Mandarin stress detection and English pitch accent detection. The ratio between the duration of the finals of current syllable and the previous syllable and the duration of the silent pause between the current syllable and the previous syllable are important in Mandarin stress detection. In Mandarin, duration is often used to mark stress. According to the statistics on ASCCD, we find that there often exists silence before the stressed syllable, which can explain why the feature durSilCurPre (the duration of the silent pause between the current syllable and the previous syllable) is important. This also illustrates that the probability is big that the first syllable is stressed after silence. The feature ratioDurCurPre (the ratio between the duration of current syllable and the duration of previous syllable) indicates that Mandarin stress is realized mainly by some attribution comparison between the current syllable and the following or the previous syllable, such as the duration comparison and pitch comparison. For some syllables, the stress may be realized by the duration comparison, and for others, the stress may be realized by the pitch comparison. The duration of the syllable is important in English pitch accent detection, and the duration of the syllable in Mandarin stress detection provides minor discrimination when compared with the duration of the syllable in English pitch accent detection. According to the statistics on BURNC, the accented syllables often have the longer duration than the unaccented ones. (2) The pitch related features provide minor discrimination in both Mandarin stress detection and English pitch accent detection when compared with the duration related features in both Mandarin stress detection and English pitch accent detection. Of all pitch related features, the mean pitch and the maximum pitch of

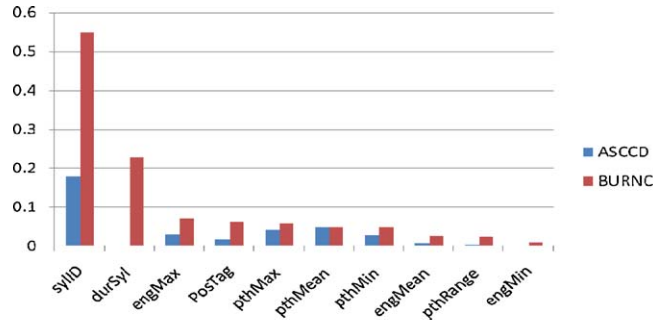

Fig. 1. The mutual informatin comparison in ASCCD and BURNC.

the syllable are relatively more important in Mandarin stress detection. Of all pitch related features, the mean pitch of the syllable and some coefficients of pitch contour Legendre polynomial expansion are relatively more important in English pitch accent detection. From these comparisons, we can find that the mean pitch of syllable, maximum pitch of syllable or some coefficients of pitch contour Legendre polynomial expansion are more important when compared with the other pitch related features. This also indicates that the stress or pitch accent is realized by increasing the pitch value of syllable and elevating the position of pitch curve. (3) In Mandarin stress detection, the energy related features in special frequency band do not provide higher discrimination than the duration, pitch and intensity related features. Of all energy related features, the maximum energy of the syllable and the some coefficients of the energy contour Legendre polynomial expansion are important in English pitch accent detection. In English pitch accent detection, the energy related features provide higher discrimination than the duration and pitch related features. For the accented syllables, their energy ranges are expanded, and their energy maximum are enhanced.

8. Discussion

Although we provide more features and higher model complexity in Mandarin stress detection than that in English pitch accent detection in the experiments, we can find that the accuracy rate of Mandarin stress detection is usually lower than that of English pitch accent detection. In Section 7, we have made comparison between Mandarin stress detection and English pitch accent detection based on feature analysis. Now, we make comparison between Mandarin stress detection and English pitch accent detection further. We will compare the differences and similarities between them from the following different aspects.

8.1. From the mutual information

In order to compare the differences and similarities across languages, we compute the mutual information between the feature and the classification label to measure the discrimination of features used in Mandarin stress and English pitch accent detection.

In Mandarin stress detection, the sylID (the syllable with tone) has the maximum mutual information value 0.178, while in English pitch accent detection, the sylID (the syllable) has the maximum mutual information value 0.549. The maximum mutual information value in Mandarin stress detection is much smaller than that in English pitch accent detection. Therefore, the features utilized in Mandarin stress detection have minor discrimination when compared with the features utilized in English pitch accent detection. Fig. 1 shows some features utilized both in Mandarin stress detection and English pitch accent detection.

In our experiments, the mutual information of most features, except very few, in ASCCD is lower than that in BURNC. From these analyses, we can find that the features utilized in Mandarin stress detection have minor discrimination when compared with those utilized in English pitch accent detection.

8.2. *From the corpus*

On Mandarin ASCCD corpus and English BURNC corpus, we have conducted the comparison experiments between Mandarin stress detection and English pitch accent detection systematically. Now, we analyze these two corpora.

Table 3 lists the stressed syllable distribution of different speakers in ASCCD corpus. From the table, we can find that: (1) the untoned syllable has more stability, which means that all speakers have almost the same pattern in terms of the untoned syllables; (2) whether the syllable is stressed or not is different from individual to individual. The size of the set, which consists of the unstressed normal syllable by all 10 speakers, is 1174, and the size of the set, which consists of the stressed syllable by all 10 speakers, is 682. Table 2 shows that there are 33,930 stressed syllables in ASCCD corpus. Therefore, for every speaker, there are about 3393 stressed syllables. That is, there are at least 2711 ($3393-682=2711$) syllables which can be stressed or unstressed for every speaker. From Table 2, we know that there are 87,586 syllables in the corpus. That is, every speaker speaks about 8758 syllables, so we can find that almost 30% (2711/8758) syllables can be stressed or unstressed for every speaker. Which syllables are stressed or unstressed is not strictly fixed in Mandarin. The ToBI system has the advantage that it can be used consistently by labelers for a variety of styles [46], and Ostendorf et al. had made study about labeler consistency on a set of three storeys containing 1002 words based on BURNC [20]. They found agreement on presence versus absence for 91% of the words. On those 487 words that were marked by both labeling groups with an accent, there was 60% agreement on accent type with most of the disagreements occurring for the difficult L + H* versus H* distinction. When the H* was grouped together with the L + H* as in [46], there was 81% agreement on pitch accent type.

8.3. *From the language*

There are two differences in stress between Mandarin and English.

One is that Mandarin Chinese is a tonal language, but English is not. Mandarin is a tonal language, in which tone plays an important phonemic role. Each syllable in Mandarin corresponds to a morpheme (ideographic character) and is associated with a pitch tone. The pitch curve in Mandarin embraces tone, stress and intonation information, and cannot be freely used to mark stress. The acoustic evidence of stress comes from duration, pitch, spectrum tilting and intensity related information. The relations of these features are rather complicated. English is not a tonal language, but a lexical-stress language. There is a strong correlation between whether the lexical-stress exists or not and whether the syllable is stressed or not.

The other is that most Chinese characters are monosyllable but most English words are polysyllabic. According to Wang [44], there are some 30,000 disyllabic compounds in a large dictionary of standard Chinese, and just about 2000 are clearly heavy–light (the rest are either heavy–heavy or optionally heavy–heavy) disyllabic structure. The stress is clear in disyllabic Chinese characters with heavy–light structure. In contrast to English, there are about 13% of the words to be monosyllabic (based on a basic lexicon of 52,447 words) [45], and so most cases are that English words have heavy–light structure (one pitch accent in a polysyllabic word). The cases that English words have heavy–heavy structure are far less common (heavy–heavy with two pitch accents).

9. Conclusion and future work

In this paper, we select a novel classifier combination method, which is the combination of boosting classification and regression tree (CART) classifier and conditional random fields (CRFs) classifier, to detect Mandarin stress and English pitch accent, verify efficiency of our proposed method on the prosodic annotation corpus and continuous speech corpus, and compare performances with several different kinds of classifier and classifier combination model. We also analyze the function of the duration, pitch, energy and intensity features in Mandarin stress detection and English pitch accent detection, and comprehensively compare the differences and the similarities between Mandarin stress detection and English pitch accent detection. We also verify some linguistic conclusions based on the analysis of large corpora. In the future, we will refine the features and models, and exploit more methods to model acoustic, lexical and syntactic features. We will utilize the prosodic annotation continuous speech corpus to train prosody aided phone model, and build prosody aided speech recognition system in order to integrate prosodic information to improve the performance of speech recognition system.

Acknowledgements

The authors are thankful to the anonymous reviewers for their valuable comments and corrections in an earlier version of our manuscript, which contributed to the significant improvement of quality of this article.

This work was supported in part by the China National Nature Science Foundation (No. 60675026, No. 60121302, No. 90820303 and No. 90820011), 863 China National High Technology Development Project (No. 20060101Z4073, No. 2006AA01Z194), Shandong Province Nature Science Foundation (No. ZR2011FQ024), and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105).

References

Chen, F., Withgott, M., 1992. The use of emphasis to automatically summarize a spoken discourse. 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 229–232.

Conkie, A., Riccardi, G., Rose, R., 1999. Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events. In: Proc. EUROSPEECH, pp. 523–526.

Ananthakrishnan, S., Narayanan, S., 2005. An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model. In: 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 269–272.

Wightman, C., Ostendorf, M., 1994. Automatic labeling of prosodic patterns. IEEE Transactions on Speech and Audio Processing 2 (4), 469–481.

Ostendorf, M., Ross, K., 1997. A multi-level model for recognition of intonation labels. In: Sagisaka, Y., Campbell, N., Higuchi, H. (Eds.), Computing Prosody. Springer, Berlin.

Sun, X., 2002. Pitch accent prediction using ensemble machine learning. In: Proc. ICSLP, pp. 953–956.

Rosenberg, A., Hirschberg, J., 2007. Detecting pitch accent using pitch-corrected energy-based predictors. In: Proc. Interspeech, pp. 2777–2780.

Chen, K., Hasegawa-Johnson, M., Cohen, A., 2004. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic prosodic model. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 509–512.

Ananthakrishnan, S., Narayanan, S., 2008. Automatic prosodic even detection using acoustic, lexical and syntactic evidence. IEEE Transactions on Audio, Speech, and Language Processing 16 (1), 216–228.

Jeon, J.H., Yang Liu, 2009. Automatic prosodic events detection using syllable-based acoustic and syntactic features. In: 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4565–4568.

Hirschberg, J., 1993. Pitch accent in context: predicting into national prominence from text. Artificial Intelligence 63 (1–2), 305–340.

Ross, K., Ostendorf, M., 1996. Prediction of abstract prosodic labels for speech synthesis. Computer Speech and Language 10 (3), 155–185.

Gregory, M.L., Altun, Y., 2004. Using conditional random fields to predict pitch accent in conversational speech. In: 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), ACL, East Stroudsburg, PA, USA, pp. 677–684.

Nenkova, A., Brenier, J., et al., 2007. To memorize or to predict: prominence labeling in conversational speech. In: Proc. NAACL-HLT, pp. 9–16.

Fernandez, R., Ramabhadran, B., 2010. Discriminative training and unsupervised adaptation for labeling prosodic events with limited training data. In: Proc. Interspeech, pp. 1429–1432.

Shao, Y.Q., Han, J.Q., Liu, T., Zhao, Y.Z., 2006. Study on automatic prediction of sentential stress with natural style. Acta Acustica 31 (3), 203–210 (in Chinese).

Hu, W., Dong, H., Tao, J., Huang, T., 2005. Study on stress perception in Chinese speech. Journal of Chinese Information Processing 19 (6), 78–83 (in Chinese).

Chen, X., Li, A., Guo hua, S., Wu, H., Zhigang, Y., 2000. An application of SAMPA-C in standard Chinese. In: Proc. ICSLP, Beijing, vol. 4, pp. 652–655.

Li, A., 2002. Chinese prosody and prosodic labeling of spontaneous speech. In: Proc. Speech Prosody, Aix-en-Provence, France, pp. 39–46.

Ostendorf, M., Price, P.J., Shattuck-Hufnagel, S., 1995. The Boston University Radio News Corpus: Linguistic Data Consortium.

Xu, B., 1997. Research and integration of speaker independent Chinese dictation system. Ph.D thesis, Institue of Automation, Chinese Academy of Sciences.

Boersma, P., weenik, D., 2009. Praat: doing phonetics by computer. Available: http://www.praat.org.

Sluijter, A.M.C., van Heuven, V.J., 1996. Spectral balance as an acoustic correlate of linguistic stress. Journal of the Acoustical Society of America 100 (4), 2471–2485.

Heldner, M., 2001. Spectral emphasis as an additional source of information in accent detection. In: Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding, pp. 57–60.

Heldner, M., Strangert, E., Deschamps, T., 1999. A focus detector using overall intensity and high frequency emphasis. In: Proc. ICPhS, pp. 1491–1494.

Fant, G., Kruckenberg, A., Liljencrants, J., 2000. Acoustic–phonetic analysis of prominence in Swedish. In: Biotins, A. (Ed.), Intonation, Analysis, Modeling and Technology. Kluwer, pp. 55–86.

Tamburini, F., 2005. Automatic prominence identification and prosodic typology. In: Proc. Interspeech, pp. 1813–1816.

Tseng, H., Chang, P., Andrew, G., et al., 2005. A conditional random field word segmenter for Sighan bakeoff 2005. In: Proc. Fourth SICHAN Workshop on Chinese Language Processing.

Chang, P.-C., Galley, M., Manning, C., 2008. Optimizing Chinese word segmentation for machine translation performance. In: Proc. ACL Third Workshop on Statistical Machine Translation.

The Stanford Postagger. Available: http://nlp.stanford.edu/software/tagger.shtml.

Ghahramani, Z., Kim, H.-c., 2003. Bayesian Classifier Combination. Gatsby Computational Neuroscience Unit Tech. Report.

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

Lafferty, J.D., McCallum, A., Pereira, F.C.N., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proc. The Eighteenth International Conference on Machine Learning (ICML 2001), pp. 282–289.

Hall, M., Frank, E., Holmes, G., et al., 2009. The WEKA data mining software: an update. SIGKDD Explorations 11 (1), 10–18.

CRF++: yet another CRF toolkit. Available: http://crfpp.sourceforge.net/.

Jeon, J.H., Liu, Y., 2009. Automatic accent detection: effect of base units and boundary information. In: Proc. Interspeech, pp. 180–183.

Bolinger, D.L., 1958. A theory of pitch accent in English. Word 14 (2–3), 109–149.

Kochanski, G., Grabe, E., Coleman, J., Rosner, B., 2005. Loudness predicts prominence: fundamental frequency lends little. Journal of the Acoustical Society of America 118 (2), 1038–1054.

Silipo, R., Greenberg, S., 2000. Prosodic stress revisited: reassessing the role of fundamental frequency. In: Proc. NIST Speech Transcription Workshop.

Rosenberg, A., Hirschberg, J., 2009. Detecting pitch accent at the word, syllable and vowel level. In: Proc. NAACL-HLT, pp. 1784–1787.

Chao, Y.R., 1980. A Grammar of Spoken Chinese. University of California Press.

Shen, J., 1994. The study for the Chinese sentence stress. Linguistic Research vol.3 (in Chinese).

Wang, Y., 1994. Tone Pattern and Word Stress in Mandarin. International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, 2004.

Wang, L., 2008. Chinese Phonology. Zhonghua Shuju, Beijing.

Baayen, R. et al., 1995. The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium (LDC). University of Pennsylvania.

Pitrelli, J.F., Beckman, M., Hirschberg, J., 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In: Proc. ICSLP, Yokohama, vol. 2, pp. 123–126.