Automatic Prosodic Events Detection by Using Syllable-based Acoustic, Lexical and Syntactic Features

Chong-Jia Ni，Wen-Ju Liu and Bo Xu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

{cjni,LWJ,xubo}@nlpr.ia.ac.cn

10.21437/Interspeech.2011-530

Abstract

Automatic prosodic events detection and annotation are important for both speech understanding and natural speech synthesis. In this paper, the complementary model method is proposed to detect prosodic events. This method discards the independent assumption between the acoustic features and the lexical and syntactic features, models not only the features of the current syllable but also the contextual features of the current syllable at the model level, and realizes the complementarities by taking the advantages of each model. The experiments on Boston University Radio News Corpus show that the complementary model can yield 91.40% pitch accent detection accuracy rate, 95.19% intonational phrase boundaries (IPB) detection accuracy rate and 93.96% break index detection accuracy rate. When compared with the previous work, the results for pitch accent, IPB and break index detection are significantly better.

Index Terms: complementary model, boosting classification and regression tree (CART), conditional random fields (CRFs)

1. Introduction

Prosody is generally used to describe aspects of a spoken utterance's pronunciation which are not adequately explained by segmental acoustic correlates of sound units (phones). The prosodic information associated with a unit of speech, say, syllable, word, phrase, or clause, influences all the segments of the unit in an utterance. They are also referred to as supra-segment that transcends the properties of local phonetic context. Many speech applications, such as automatic speech recognition, speech understanding and speech synthesis, can benefit from corpus annotated with prosodic information, but it is very expensive and time consuming to do such work manually, so an automatic prosodic annotation algorithm will be very useful for building prosodic annotation corpus.

Many researches have been made in automatic prosodic events detection at both the syllable and the 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, intensity. A variety of machine learning approaches have been used in order to model these acoustic, lexical and syntactic features. Wightman utilized decision tree to model acoustic evidence (such as pitch, intensity and duration evidence), and combined with a probabilistic model (bi-gram) to detect binary accent, IPB and break index. Their method achieved 71% for IPB detection, 84% for accent detection and 84% for break index detection at syllable level on Boston University Radio News corpus [1]. Ostendorf used a multilevel 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 ToBI break index tier, could achieve 66.9% accuracy rate [2]. Chen built Gaussian mixture model (GMM) based on acoustic evidence and artificial neural network (ANN) model based on syntactic evidence in maximum likelihood framework to binary intonational phrase boundary and pitch accent detection, and achieved 90% accuracy rate for IPB detection and 84% accuracy rate for pitch accent detection at word level [3]. Ananthakrishnan and Narayanan used a maximum a posteriori (MAP) framework for prosodic events detection. They used an n-gram structure for prosodic language model, and utilized neural network (NN) to model acoustic evidence. When combined 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 and 87% accent accuracy rate at syllable level [4]. Jeon showed that the neural network classifier achieved the best performance in modeling acoustic evidence, and support vector machines were more effective for lexical and syntactic evidence. The combination of the acoustic and syntactic models yielded 93.3% intonational phrase boundary (IPB) detection accuracy, 89.8% pitch accent detection accuracy, and 91.1% break index detection accuracy [5].

In this paper, we propose the complementary model, which is the combination of boosting classification and regression tree (CART) model and conditional random fields (CRFs) model, to detect prosodic events, which are intonational phrase boundary, break and pitch accent. This method discards the independent assumption between the acoustic features and the lexical and syntactic features. This method models not only the features of the current syllable but also the contextual features of the current syllable at the model level, and realizes the complementarities by taking the advantages of each model.

The rest of this paper is organized as follows. In section 2, we provide details of the corpus. In section 3, the features used in prosodic events detection are introduced, including acoustic features, lexical and syntactic features. In section 4, the prosodic events detection algorithm is presented. The experimental results and analysis are introduced in section 5. The final section gives a brief summary along with future research directions.

2. The corpus

Boston University Radio News Corpus (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, m2b and m3b), which amounts to speeches of 3 hours. The corpus is annotated with pitch accent tones (*), break index (0~4), orthographic transcription, automatically generated and hand-corrected part-of-speech (POS) tags, pitch and automatic phone alignments information [6]. In BURNC corpus, we take binary accent, break and IPB detection. If a syllable is marked with accent mark *, we believe the pitch

accent is present, which means that the syllable is stressed. The break index 3 and 4 are grouped together to represent that there is a break. All of the IPB tones are grouped into one category. Table 1 lists the statistics of BURNC.

Table 1. *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	12607	2733	5059	3608	2093
#Syllables	6562	20700	4422	8144	5904	3354
#Accents	2344	7061	1545	2786	2113	1094
#Breaks	1116	3914	744	1247	986	459
#IPBs	748	2801	437	784	657	292

3. The features

In this section, we introduce the acoustic, lexical and syntactic features utilized in prosodic events detection. In order to eliminate the natural variations among different speakers, the pitch related and intensity related features are normalized by utilizing z-score method.

For every syllable, we compute the following features.

Pitch range (4 features): maximum pitch (pthmax), minimum pitch (pthmin), mean pitch (pthmean) and pitch range (difference between maximum and minimum pitch) (pthrng).

Pitch contour (6 features): The coefficients of 5-order Legendre polynomial expansion (pthCoefi,i=0,1,…,5).

Intensity range (4 features): maximum intensity (intmax), minimum intensity (intmin), mean intensity (intmean) and intensity range (difference between maximum and minimum intensity) (intrng).

Intensity contour (6 features): The coefficients of 5-order Legendre polynomial expansion (intCoefi, i=0,1,..,5).

Duration (4 features): The duration of the syllable (syldur); the silence duration after the syllable (sildur); the ratio between the following syllable and the current syllable (durratio); the vowel duration of the syllable (voweldur).

The lexical and syntactic related features (6 features): The pos tag of the syllable (pos); the number of syntactic phrases the word initiates (wrdInit); the number of syntactic phrases the word terminates (wrdTrm); the syllable identification (ID); lexical stress (exist or not) (blexstress); and word boundary information (boundary or not) (bwrdbnd).

For the pitch or intensity contour, we utilize the following method to compute expansion coefficients.

Let us suppose $f(t)$ to be the pitch or intensity 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 $P_n(t) = \begin{cases} 1 & n=0 \\ t & n=1 \\ \frac{2n-1}{n} t P_{n-1}(t) - \frac{n-1}{n} P_{n-2}(t) & n \geq 2 \end{cases}$ is the n-th Legendre polynomial, a_n is the coefficient of expansion equation. Each coefficient in expansion equation (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.

In the lexical and syntactic related features, these features—pos, wrdInit and wrdTrm, are used for break and IPB detection. These features—pos, ID, blexstress and bwrdbnd, are used for pitch accent detection. In order to capture the context influence in prosodic events detection, we also compute the lexical and syntactic related features in the contextual windows. For the break and IPB detection, we choose 3 previous and 2 next syllables of the current syllable as the contextual window. For pitch accent detection, we choose 2 previous and 2 next 2 syllables of the current syllable as the contextual window. So we can get 42 features for the break and IPB detection, and 41 features for the pitch accent detection.

4. Prosodic events detection based on complementary model

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 reference [5], J. H. Jeon listed the equations (2) ~ (5) that are often used for prosodic events detection. So we cite directly and list these equations below.

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

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

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

$$\approx \arg\max \prod_{i=1}^{n} p(p_i|a_i)^{\lambda}\, p(p_i|\phi(s_i)) \tag{4}$$

$$\approx \arg\max \lambda \sum_{i=1}^{n} \log\left(p(p_i|a_i)\right) + \sum_{i=1}^{n} \log\left(p(p_i|\phi(s_i))\right) \tag{5}$$

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, $\log\left(p(p_i|a_i)\right)$ is the acoustic prosodic model score, $\log\left(p(p_i|\phi(s_i))\right)$ is the lexical and syntactic prosodic model score, and λ is a weighting between the acoustic prosodic and the lexical and syntactic-prosodic model. The acoustic prosodic model and the lexical and syntactic prosodic model can be obtained by using machine learning methods. The traditional modeling techniques, 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 equation (5) to combine the acoustic prosodic model and the lexical and syntactic prosodic 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, Z. 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 [7]. In fact, the equation (5) is a specific case of classifier combination of two models.

We all know that the acoustic-based features and the lexical-based and syntactic features-based are not independent. In order to reduce the computational complexity, $p(P|A,S)$ has

been simplified to $p(P|A)p(P|S)$ in equation (3). In fact, it is not necessary to do so.

We can transform equation (2) into equation (8).

$$P^* = \arg\max p(P|A,S) \quad (6)$$

$$= \arg\max\left(\lambda \cdot p(P|A,S) + (1-\lambda)\cdot p(P|A,S)\right) \quad (7)$$

$$= \arg\max\left(\lambda \cdot p_1(P|A,S) + (1-\lambda)\cdot p_2(P|A,S)\right) \quad (8)$$

This is only a deformation of equation (2). We give $\lambda \cdot p(P|A,S)$ a new symbol $\lambda \cdot p_1(P|A,S)$, and $(1-\lambda)\cdot p(P|A,S)$ another new symbol $(1-\lambda)\cdot p_2(P|A,S)$. If the same method is used to model p_1 and p_2, the method used in equation (8) is one type of methods, of which ensemble machine learning method is one [8]. 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 lexical and syntactic features are independent, the equation (8) can be written as equation (5). If we take different methods to model p_1 and p_2, and quit the hypothesis that the acoustic features and the lexical and syntactic features are independent, we can get another prosodic events detection method. We call this method as the complementary method, and p_1 and p_2 as a pair of complementary model. In fact, this method is also the combination of different classifiers. The differences between this method and the one proposed by Jeon are that (1) This classifier combination method does not adopt the independent assumption between the acoustic features and the lexical and syntactic features; (2) This 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.

Let us suppose $\lambda \neq 1$ in equation (8). We can transform equation (8) to equation (13):

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

$$= \arg\max\left(\lambda \cdot p(P|A,S) + (1-\lambda)\cdot p(P|A,S)\right) \quad (10)$$

$$= \arg\max\left(\lambda \cdot p_1(P|A,S) + (1-\lambda)\cdot p_2(P|A,S)\right) \quad (11)$$

$$= \arg\max\left(\frac{\lambda}{(1-\lambda)}\cdot p_1(P|A,S) + p_2(P|A,S)\right) \quad (12)$$

$$= \arg\max\left(w\cdot p_1(P|A,S) + p_2(P|A,S)\right) \quad (13)$$

where $\frac{\lambda}{(1-\lambda)}$ is equal to w, $p_1(P|A,S)$ is equal to $p(P|A,S)$, and $p_2(P|A,S)$ also substitutes $p(P|A,S)$.

"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" [9], 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. 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 [10]. We all know that no matter what the word or syllable is, 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 syllable features greatly. CRFs methods can model the previous and following syllable features. We use Boosting classification and regression tree (CART) and CRFs methods to model p_1 and p_2 separately.

5. Experiments

5.1. Experiments setup

In our experiments, WEKA implementation of C4.5 algorithm classifier (J48) is used to train decision tree model [11]. LibSVM is used to train SVM model [12]. CRF++ 0.53 is used to train CRFs model [13]. We create 2-layer multilayer perception network with a single hidden layer, in which the number of hidden unit in hidden layer is half of the number of input features, to train neural network (NN) model. 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 BURNC, we use the pitch information, duration information and POS tag information coming from the annotation. The intensity information is extracted by using the Praat [14]. We randomly split the utterances coming from all speakers in the corpus and perform 5-fold cross validation for prosodic events detection tasks. The final result is the average of the 5-fold cross validation results.

5.2. Experimental results and analysis

5.2.1. The acoustic prosodic model

We use decision tree and neural network to model the acoustic features. The experimental results are shown in Table 2.

Table 2. *The performance of various acoustic-prosodic models.*

		Accuracy Rate (%)	F-Measure
NN	Pitch accent	83.95	83.90
	Break	89.28	88.96
	IPB	93.31	93.07
Decision Tree	Pitch accent	81.45	81.38
	Break	88.84	88.44
	IPB	92.73	92.43

From Table 2, we can find that: There are certain differences between the performances of decision tree model and neural network model. The performance of neural network model is slightly better.

5.2.2. The lexical and syntactic prosodic model

Table 3. *The performance of various lexical and syntactic-prosodic models.*

		Accuracy Rate (%)	F-Measure
Decision Tree	Pitch accent	86.34	86.52
	Break	90.85	90.76
	IPB	91.83	91.40
SVM	Pitch accent	84.28	84.56
	Break	87.02	84.72
	IPB	90.66	88.80
CRFs	Pitch accent	88.54	88.63
	Break	90.83	90.64
	IPB	92.18	91.91

For lexical and syntactic features, we employ three different classifiers: Decision tree, SVM and CRFs.

Table 3 shows the performance of various lexical and syntactic prosodic models. From Table 3, we can find that: The CRFs model achieves relatively good results. From the comparison between Table 2 and Table 3, we can find the performance of the lexical and syntactic prosodic model is better than the acoustic prosodic model.

5.2.3. The combined model

Table 4 shows the performance of various combined models. The value of λ in equation (5) ranging from 0.4 to 0.9 has good effect, and can fuse the classification results of the acoustic prosodic classifier and the lexical and syntactic prosodic classifier. The value of λ is tuned on the training set. From Table 4, we can find that: (1) The combination of different knowledge obtains better performance than each alone for all classifiers. (2) The Boosting CART classifier and CRFs classifier can provide better classified efficiency.

Table 4. *The performance of various combined models and complementary model.*

		Accuracy Rate (%)	F-Measure
NN/Decision tree (Baseline)	Pitch accent	87.70	87.80
	Break	91.27	91.06
	IPB	93.07	92.86
NN/SVM	Pitch accent	84.30	84.60
	Break	89.19	88.90
	IPB	92.49	92.20
NN/CRFs	Pitch accent	89.55	89.61
	Break	91.68	91.52
	IPB	93.83	93.67
Boosting CART*	Pitch accent	89.50	89.52
	Break	93.10	93.01
	IPB	94.56	94.41
CRFs*	Pitch accent	90.40	90.45
	Break	92.60	92.50
	IPB	94.62	94.51
Boosting CART* + CRFs*	Pitch accent	91.40	91.39
	Break	93.96	93.88
	IPB	95.19	95.08

In Table 4, Boosting CART* classifier and CRFs* classifier are obtained by using the acoustic, lexical and syntactic features, and are not obtained by weighting combination through equation (5). "NN/Decision tree" is the combination of the acoustic prosodic model based on NN and the lexical and syntactic prosodic model based on Decision tree by using equation (5). The "NN/SVM" and "NN/CRFs" are similar.

Now, we can obtain a new classifier "Boosting CART* + CRFs" by weighting combination of the Boosting CART* classifier and CRFs* classifier according to equation (13). The value of λ in equation (13) is 1. This means that the weight in equation (8) is 0.5. The experimental results are also listed in Table 4.

From Table 4, we can find: (1) The complementary model achieve 91.40% pitch accent detection accuracy rate, 95.19% intonational phrase boundaries (IPB) detection accuracy rate and 93.96% break index detection accuracy rate. (2) Our proposed method can obtain better effect when compared with the baseline system. There are 3.7% improvement on pitch accent detection task, 2.69% improvement on break detection task and 2.12% improvement on intonational phrase boundaries detection task. (3) The model generated by using acoustic, lexical and syntactic features has better performance than the models generated by using individual type of acoustic or syntactic features.

Although the experimental setting, features and data used in our experiment are not exactly the same as the previous Jeon's work on BURNC [5], a direct comparison of results may not be very exact. We still compare our experimental results with Jeon's. When compared with Jeon's experimental results, our proposed methods have a better performance in prosodic events detection. For pitch accent, IPB and break detection, there are 1.6%, 1.89% and 2.86% improvements separately.

6. Conclusion and discussion

In this paper, we develop complementary model method to detect English prosodic events by using acoustic, lexical and syntactic evidence. This method discards the independent assumption between the acoustic features and the lexical and syntactic features, models not only the features of the current syllable but also the contextual features of the current syllable at the model level, and realizes the complementarities by taking the advantages of each model. When compared with the baseline system and the previous Jeon's work, the complementary model can achieve better performance. In the future, we will refine our models and features, and exploit other methods to model acoustic, lexical and syntactic features.

7. Acknowledgements

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

8. References

[1] C. Wightman and M. Ostendorf, "Automatic labelling of prosodic patterns," IEEE Trans. Speech Audio Process., vol. 2, no. 4, pp.469-481, Oct. 1994.

[2] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," Computer Speech Lanauge, vol. 10, pp. 155-185, 1996.

[3] Ken Chen, M. Hasegawa-Johnson and A. Cohen, "An Automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic prosodic model," in Proc. ICASSP, pp.509-512, 2004.

[4] S. Ananthakrishnan and S. Narayanan, "Automatic prosodic even detection using acoustic, lexical and syntactic evidence," in IEEE Trans. on Audio, Speech, and Language Processing, vol.16, pp.216-228, 2008.

[5] J. H. Jeon and Yang Liu, "Automatic Prosodic Events Detection Using Syllable-based Acoustic and Syntactic Features," in Proc. ICASSP, pp.4565-4568, 2009.

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

[7] Z. Ghahramani, Z. Ghahramani and Hyun-chul Kim, "Bayesian Classifier Combination", 2003.

[8] Xuejing Sun, "Pitch Accent Prediction Using Ensemble Machine Learning," in Proc. ICSLP, 2002, pp.953-956.

[9] Y. Freund and R. E. Schapire, "A Decision- Theoretic Generalization of On-Line Learning and an Application to Boosting", Journal of Computer and System Sciences, 55(1), 119-139, 1997.

[10] Lafferty J D, McCallum A, Pereira F C N. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proc. ICML'01, pp.282-289, 2001.

[11] S., Garner, "Weka: the waikato environment for knowledge analysis," in Proc. the New Zealand Computer Science Research Students Conf., 1995.

[12] LibSVM—A Library for Support Vector Machines, Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[13] CRF++: Yet Another CRF toolkit, http://crfpp.sourceforge.net/

[14] P. Boersma, D. weenik, "Praat: doing phonetics by computer," Available: http://www.praat.org