AUTOMATIC PROSODIC EVENTS DETECTION USING SYLLABLE-BASED ACOUSTIC AND SYNTACTIC FEATURES

Je Hun Jeon and Yang Liu

Computer Science Department
The University of Texas at Dallas, Richardson, TX, USA
{jhjeon, yangl}@hlt.utdallas.edu

ABSTRACT

Automatic prosodic event detection is important for both speech understanding and natural speech synthesis since prosody provides additional information over the short-term segmental features and lexical representation of an utterance. Similar to previous work, this paper focuses on automatic detection of coarse level representation of pitch accents, intonational phrase boundaries (IPB), and break indices. We exploit various classifiers and identify effective feature sets to improve performance of prosodic event detection according to acoustic, lexical, and syntactic evidence. Our experiments on the Boston University Radio News Corpus show that the neural network classifier achieves the best performance for modeling acoustic evidence, and that support vector machines are more effective for the lexical and syntactic evidence. The combination of the acoustic and the syntactic models yields 89.8% accent detection accuracy, 93.3% IPB detection accuracy, and 91.1% break index detection accuracy. Compared with previous work, the IPB performance is similar, whereas the results for accent and break index detection are significantly better.

Index Terms— Prosodic event detection, accent, intonational phrase boundary, break index, ToBI annotation

1. INTRODUCTION

The tonal and rhythmic aspects of speech are generally called prosody. Since it normally extends over more than one phoneme segment, prosody is said to be supra-segmental. Speakers use prosody to convey emphasis, intent, attitude, and emotion. These are important cues for interpretation of speech. Prosody can play an important role in speech understanding (such as speech act detection and word disambiguation) and natural speech synthesis, because it includes aspects of higher level information that is not completely revealed by segmental acoustics. Many speech applications can benefit from corpora annotated with prosodic information, but it is very expensive and time-consuming to perform prosody labeling manually, therefore, an automatic prosodic labeling algorithm will be very useful for building spoken language understanding systems.

The main prosodic events that we are concerned to detect automatically in this paper are phrasing and accent (or prominence). Prosodic phrasing refers to the perceived grouping of words in an utterance, and prominence refers to the greater perceived strength or emphasis of some syllables in a phrase. These tasks have been evaluated in several prior studies [2-5].

Prosody in spoken utterances correlates with acoustic and syntactic evidence. Acoustic cues such as duration, intensity, and pitch have a very close relationship with prominence and phrasal tones. Lexical and syntactic cues such as part-of-speech, syllable identity, and syllable stress also exhibit a high degree of correlation with prosodic events. Most of the previous work used these acoustic, lexical, and syntactic features, with some differences in the detailed feature representation and implementation. To model these evidence, a variety of machine learning approaches have been used previously, such as decision tree [2], neural network [3,5], maximum entropy model [4], Gaussian mixture model [3], and n-gram model [5]. The task setup was also different in prior studies, for example, syllable-based [2,5] versus word-based pitch accent detection [3,4]. The results of previous work on prosodic event detection are summarized in Table 1.

In this paper, we attempt to combine different sources of evidence to improve the performance of three prosodic event detection tasks: 1) accent, 2) intonational phrasal boundary, and 3) break index. We develop two models according to the source of evidence: 1) an acoustic-prosodic model based on the acoustic features, and 2) a syntactic-prosodic model using lexical and syntactic features. We exploit various classifiers for each model, propose a proper feature set, and evaluate the combination of the two models. Our experimental results have shown that our performance on these prosodic event detection tasks compares favorably to those in previous work.

In the next section, we provide details on the corpus and the prosodic event detection tasks. In Section 3, we describe our approach, including the acoustic and syntactic prosodic models, and the features used. Section 4 presents our experiments and results. The final section gives a brief summary along with future directions.

TABLE 1. Summary of previous work on prosodic event detection. All of these studies used Boston University Radio News Corpus [6]. The results shown are the detection accuracy. The last row summarizes our experimental results.

	Algorithm	Level	Pitch Accent	IP Boundary	Break Index
Wightman and Ostendorf [2]	HMM/CART	syllable	84.0%	71.0%	84.0%
Hasegawa-Johnson et al. [3]	Neural network/GMM	word	84.2%	93.1%	-
Sridhar et al. [4]	Maximum entropy	word	86.0%	93.1%	84.0%
Ananthakrishnan et al. [5]	Neural network/n-gram	syllable	86.8%	-	86.9%
Our approach in this paper	Neural network/SVM	syllable	89.8%	93.3%	91.1%

2. CORPUS AND TASKS

Automatic detection of phrasing and prominence requires appropriate representation schemes that can characterize prosody in a standardized manner. One of the most popular prosodic event labeling schemes is the ToBI framework [1]. The pitch accent tones (*) are marked at every accented syllable and have five types according to pitch contour: H*, L*, L*+H, L+H*, H+!H*. The phrase boundary tones are marked at every intermediate phrase boundary (L-, H-) or intonational phrase boundary (L-L%, L-H%, H-H%, H-L%). The break indices range in value from 0 through 4, where 4 means intonational phrase boundary, 3 means intermediate phrase boundary, and a value less than 3 means phrase-medial word boundary.

In this paper, our experiments are carried out using the Boston University Radio News Corpus (BU) [6], which consists of broadcast news style read speech and has ToBI-style prosodic annotations for part of the data. The corpus is annotated with orthographic transcription, automatically generated and hand-corrected part-of-speech (POS) tags, and automatic phone alignments.

The detailed representation of prosodic events in the ToBI framework creates a serious sparse data problem for automatic prosody detection. Take accented syllables as an example. In the BU corpus, the most frequent accent tone H* (which is 45% of the total accent tones) is only 16% among all the syllables, and most of the other accent types are under 5% in the entire corpus. This problem can be alleviated by grouping ToBI labels into coarse categories, such as presence or absence of pitch accents and phrasal tones. This also significantly reduces ambiguity of the task. In this paper, we thus use coarse representation (presence versus absence) for three prosodic event detection tasks:

- Pitch accents: accent mark * means presence.
- Intonational phrase boundaries (IPB): all of the IPB tones (%) are grouped into one category.
- Break indices: value 3 and 4 are grouped together to represent there is a break.

These three tasks are binary classification problems. Similar setup has also been used in previous work, therefore we can compare our results to theirs. Table 2 shows the coarse level statistics of the BU corpus we used in this study.

TABLE 2. Coarse level statistics of BU Corpus

	Female			Male		
	f1a	f2b	f3a	m1b	m2b	m3b
# Words	4386	12607	1732	5059	3399	1936
# Syllables	7158	20697	2855	8169	5636	3115
# Accents	2513	7063	1005	2792	1996	1094
# IPBs	808	2801	282	784	636	299
# Breaks	1215	3912	472	1254	935	446

3. PROSODIC EVENT DETECTION

We model the prosodic event detection problem as a classification task. We separately develop acoustic-prosodic and syntactic-prosodic models, and then combine the two models. For each model, we employ various classifiers to hypothesize the prosody labels from either acoustic or syntactic evidence. Note that our feature extraction is performed at the syllable level, similar to [2,5]. This is straightforward for accent detection since stress is defined to occur on syllables. In the case of IPB and break index detection, we use only the features from the final syllable of a word since these two events are associated with word boundaries.

3.1. The acoustic-prosodic model

The most likely sequence of prosodic events $P^* = \{p^*_1, \ldots, p^*_n\}$ given the sequence of acoustic evidence $A = \{a_1, \ldots, a_n\}$ can be found as following:

$$P^* = \arg\max_p p(P \mid A) \approx \arg\max_p \prod_{i=1}^{n} p(p_i \mid a_i) \quad (1)$$

where $a_i = (a_i^1 \ldots, a_i^t)$ is the acoustic feature vector corresponding to the syllable. Note that this assumes that the prosodic events are independent and they are only dependent on the acoustic observations in the corresponding locations, but we try to capture some of the sequential information in the features.

The acoustic cues used for prosodic events detection represent pitch, energy, and duration. We used the Momel algorithm [6] to estimate the pitch values and pitch contour. This algorithm uses a spline function to code pitch curve instead of a simple linear approximation. In order to reduce the effect by both inter-speaker and intra-speaker variation, both pitch and energy values were normalized (z-value) with utterance specific means and variances. The acoustic features used in our experiments are listed below. Again, all of the features are computed for a syllable.

- Pitch range (4 features): maximum pitch, minimum pitch, mean pitch, and pitch range (difference between maximum and minimum pitch).
- Pitch slope (5 features): first pitch slope, last pitch slope, maximum plus pitch slope, maximum minus pitch slope, and the number of changes in the pitch slope patterns.

- Energy range (4 features): maximum energy, minimum energy, mean energy, and energy range (difference between maximum and minimum energy).
- Duration (3 features): normalized vowel nucleus duration, pause duration after the word final syllable, and the ratio of vowel nucleus durations between this syllable and the next syllable.

Among the duration features, the pause duration and the ratio of vowel nucleus durations are only used to detect IPB and break index, not for accent detection.

3.2. The syntactic-prosodic model

The prosodic events P^* given the sequence of syntactic evidence $S = \{s_1, \ldots, s_n\}$ can be found as following:

$$P^* = \arg\max_P p(P \mid S) \approx \arg\max_P \prod_{i=1}^{n} p(p_i \mid \phi(s_i)) \quad (2)$$

where $\phi(s_i)$ is chosen such that it contains lexical and syntactic evidence from a fixed window of syllables surrounding location *i*.

There is a very strong correlation between the prosodic events in an utterance and its lexical and syntactic structure. For pitch accent detection, [5] showed that the lexical features such as the canonical stress patterns from the pronunciation dictionary perform better than the syntactic features, while for IP boundary and break index detection, the syntactic features such as POS tag work better than the lexical features. We use different feature types for each task, as listed below.

- Accent detection: syllable identity, lexical stress (exist or not), word boundary information (boundary or not), and POS tag. We also include syllable identity, lexical stress, and word boundary features from the previous and the following context window.
- IPB and Break index detection: POS tag, the frequency of syntactic phrase that the word initiates and terminates. All of these features from the previous and the following context windows are also included.

3.3. The combined model

The two models above can be combined as a maximum likelihood classifier. If we assume that the acoustic observations are conditionally independent of the syntactic features given the prosody label, the task of prosodic event detection is to find the optimal sequence P^* as follows:

$$P^* = \arg\max_P p(P \mid A, S) \approx \arg\max_P p(P \mid A) p(P \mid S) \approx \arg\max_P \prod_{i=1}^{n} p(p_i \mid a_i)^{\lambda} p(p_i \mid \phi(s_i)) \quad (3)$$

where λ is a variable that can be used to adjust the weighting between syntactic and the acoustic models.

4. EXPERIMENTS AND RESULTS

In all our experiments, we randomly split the utterances in the corpus and performed 5-fold cross validation for the three prosodic event detection tasks. The final result is the average of the 5-fold cross validation results. Performance is measured using the classification accuracy.

4.1 Acoustic-prosodic event detection

As described in Section 3.1, there are four sets of acoustic features: pitch range (PR), energy range (ER), pitch slope (PS), and duration (Dur). We examine the effectiveness of different feature sets. For pitch processing, we also evaluate the contribution of using the Momel algorithm [6], which has never been used in these prosodic event detection tasks. We use "M" together with the pitch feature sets to represent using the Momel algorithm, and "L" for not using it, that is, using the raw pitch values for pitch range and linear approximation for pitch slope. The base features we used are L-PR+ER+Dur, i.e., pitch range based on the raw pitch values, energy range, and duration features. These are similar to the features used in [5].

First we evaluated various classifiers, including decision trees [2], GMM [3,5], maximum entropy [4], and neural network (NN) [5], based on the base features. Among these classifiers, NN outperformed the others, especially for accent detection. Therefore we used NN-based acoustic-prosodic model to investigate contribution of different features. For NN, we used the default setting of the Weka toolkit, and the size of the hidden layers is half of the number of input features. Results are shown in Table 3. Duration features are used in all the test settings since our focus is mainly on pitch related features. For accent detection, using an estimated pitch curve with the Momel algorithm is much better than using the raw values, and the pitch slope features we introduced also contribute to performance improvement. On the IPB and break index detection tasks, the contribution of the estimated pitch curve and slope features is not as significant as for accent detection. Compared to the results reported in previous work, (as shown in the last row of Table 3), the performance of IPB and break detection tasks with our feature set is better; however, the accent detection accuracy is slightly worse than [2]. This is possibly because [2] used not only speaker dependent test set but also lexical stress as an acoustic feature in their model.

TABLE 3. NN-based acoustic model for prosodic event detection

	Accent	IPB	Break
Chance performance	65.43	80.74	71.73
Dur+L-PR+ER	76.03	91.06	84.16
Dur+M-PR+ER	79.20	91.00	84.27
Dur+M-PR+M-PS	81.01	91.03	**84.89**
Dur+M-PR+ER+M-PS	**83.53**	**91.19**	84.82
Referenced performance	84.0 [2]	84.1 [4]	84.61 [5]

4.2 Syntactic-prosodic event detection

For syntactic features, we employed four different classifiers: decision trees, NN [3], maximum entropy [4],

TABLE 4. SVM-based syntactic model for accent detection

	Accent
Syl +2 prev cxt	85.41
Syl+stress+2 prev cxt	85.90
Syl+stress+bnd+2 prev cxt	87.78
Syl+stress+bnd+POS+2 prev cxt	87.79
Syl+stress+bnd+POS+2 prev cxt+ 2 next cxt	**87.92**
Referenced performance [5]	85.7

TABLE 5. SVM-based syntactic model for IPB and Break detection

	IPB	Break
POS+3 prev cxt+3 next cxt	89.74	87.73
POS+POC+3 prev cxt+3 next cxt	91.28	89.61
POS+POC+3 prev cxt+2 next cxt	**91.36**	**89.76**
POS+POC+4 prev cxt+3 next cxt	91.23	89.46
Referenced performance [4]	91.5	85.0

and SVMs. We observed that SVMs with polynomial kernel achieved the best performance on all the three tasks, and thus used SVM-based syntactic-prosodic model to evaluate different feature sets.

Table 4 shows the performance of accent detection according to various feature sets. The addition of boundary feature (bnd) yields more performance gain than any other additional features. The POS feature is not a significant factor, which is similar to the results in [5]. The accuracy of our syntactic-prosodic model is higher than that in [5], which used a factored 3-gram with syllable identity and lexical stress features.

Table 5 shows the results for IPB and break index detection. The best performance gain of IPB and break detection is achieved by adding the phrase opening and closing (POC) information of a word, which is similar to the results in [3]. Even though the number of features in [4] is more than ours, our break index detection performance is much better than that of [4].

4.3 Combined prosodic event detection

Table 6 summarizes the performance of the combined model using Equation (3), along with the results using acoustic and syntactic models alone. The values of λ in Equation (3) were dynamically chosen by tuning on the training set. For each test fold, λ ranges from 0.90 to 1.33 for accent detection, 0.62 to 1.33 for IPB detection, and 0.65 to 1.27 for break detection. In order to account for the different class distributions for the three tasks, we also include the F-measure results in the table (inside the parenthesis). This allows us to compare across tasks. As shown in Table 6, the combination of the two knowledge sources yields better performance than each alone for all the tasks. We also notice that for accent and break index detection, the syntactic model outperforms the acoustic one, whereas for IPB detection, the two models have comparable results. Compared to previous work, even though there is some difference in experimental setup, it seems that the performance of our proposed models and feature sets to detect accent and break indices is significantly better than previous work.

TABLE 6. Combined model for prosodic event detection. Results shown are detection accuracy and F-measure (inside parenthesis).

	Accent	IPB	Break
Acoustic	83.53 (0.75)	91.19 (0.75)	84.89 (0.68)
Syntactic	87.92 (0.83)	91.36 (0.76)	89.76 (0.81)
Combined	**89.84** (0.85)	**93.31** (0.81)	**91.06** (0.83)
Referenced	86.75 [5]	93.09 [4]	86.91 [5]

5. CONCLUSIONS

In this paper, we exploit various classifiers and identify effective feature sets to improve performance of three prosodic event detection tasks according to acoustic, lexical, and syntactic evidence. Our experiments on Boston University Radio News corpus show that neural network is very efficient to model acoustic evidence, and SVM is better than other classifiers to model syntactic evidence. The combined model of acoustic and syntactic models achieves an accuracy of 89.8% for accent detection, 93.3% for IPB detection, and 91.1% for break index detection, a result significantly better than previous work.

We used similar features for IPB and break index detection. As shown in our experiment results, the difference between these two tasks is evident (in particular, the different contributions from the two information sources). In the future, we plan to exploit further the difference of these two tasks and refine features and models.

6. ACKNOWLEDGMENT

This work is partly supported by DARPA under Contract No. HR0011-06-C-0023. Distribution is unlimited. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

7. REFERENCES

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," *in Proc. of ICSLP*, Canada, pp. 867–870, 1992.

[2] C.W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *in IEEE Transactions on Audio and Speech Processing*, vol. 2, pp. 469-481, 1994.

[3] K. 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. of ICASSP*, USA, pp. 509-512, 2004.

[4] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *in IEEE Transactions on Audio, Speech, and Language processing*, vol. 16, pp. 797-811, 2008.

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

[6] D. Hirst and R. Espresser, "Automatic modeling of fundamental frequency using a quadratic spline function," *in Travaux de l'Institut de Phonetique d'Aix*, vol. 16, pp. 75-85, 1993.

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