INTERSPEECH
2009 BRIGHTON

Automatic Accent Detection: Effect of Base Units and Boundary Information

Je Hun Jeon and Yang Liu

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

10.21437/Interspeech.2009-70

Abstract

Automatic prominence or pitch accent detection is important as it can perform automatic prosodic annotation of speech corpora, as well as provide additional features in other tasks such as keyword detection. In this paper, we evaluate how accent detection performance changes according to different base units and what kind of boundary information is available. We compare word, syllable, and vowel-based units when their boundaries are provided. We also automatically estimate syllable boundaries using energy contours when phone-level alignment is available. In addition, we utilize a sliding window with fixed length under the condition of unknown boundaries. Our experiments show that when boundary information is available, using longer base unit achieves better performance. In the case of no boundary information, using a moving window with a fixed size achieves similar performance to using syllable information on word-level evaluation, suggesting that accent detection can be performed without relying on a speech recognizer to generate boundaries.

Index Terms: prosody, prominence, accent detection

1. Introduction

In English, an prominence or accent refers to an intonational prominence associated with a metrically strong syllable or word at the sentence level, which is often marked by pitch movements, increased energy, or prolonged duration. The task of automatically detecting pitch accent has received a considerable amount of research attention. It plays an important role not only in automatic annotation of speech corpora but also in providing useful features in other tasks such as speaker identification and keyword detection.

A wide range of supervised machine learning techniques have been applied to this problem using acoustic evidence. Wightman and Ostendorf [1] used a decision tree to detect accented syllables obtaining 84.0% accuracy on Boston University Radio News Corpus (BU) [6]. Hasegawa-Johnson et al. [2] achieved 77.3% accuracy on BU with Gaussian mixture model and Sridhar et al. [3] reported 80.1% accuracy using maximum entropy model. Neural networks have also been applied to this task with some success, yielding an accuracy of 74.1% in Ananthakrishnan et al. [4], and 83.5% in Jeon and Liu [5]. These results in previous work are not comparable because of different task setups. For example, the features and test sets are different. Furthermore, [1,4,5] used syllable-based accent detection, whereas [2,3] are word-based ones.

The perception of accent is commonly aligned with a syllable. However, the semantic and pragmatic implications of accenting lie at the word, or even phrase level. Proponents of syllable-based accent detection make the claim that it is easy to translate from syllable to word prominence and not vice versa. Word-based accent detection advocates argue that acoustic realizations of accent are rarely confined to a single syllable, and that mapping word to syllable prominence is not difficult since accent is realized on the lexically stressed syllable in a word.

Accent detection at the word or syllable level has different performance. In addition, defining the word and syllable boundaries manually or automatically has an impact on the performance of the detector. In this paper, we explore how accent detection performance changes, 1) when using different units (word, syllable, or vowel), and 2) when there is no boundary information available and estimated boundaries are used. Our experiment shows that when we have boundary information, using longer unit yields better performance. In the case of no boundary information, there is a performance drop; however, without any boundary information using a moving window/frame with a fixed size achieves similar performance to using syllable information on word-level evaluation.

In the next section, we describe our accent detection method including the acoustic features used. In Section 3, we provide detailed configuration of different base units and boundary estimation methods. Section 4 presents our experiments and results. The final section gives a brief summary along with future directions.

2. Accent detection method

We model the accent detection problem as a binary classification task, that is, a classifier is used to determine whether the base unit is accented or not. In this study, we only use acoustic/prosodic information. This will allow us to examine whether it is possible to perform accent detection without the need to run speech recognition for any given speech signal. The most likely sequence of accented events $A^*=\{a^*_1,\ldots,\ a^*_n\}$ given the sequence of prosodic features $P=\{p_1,...,p_n\}$ can be found as following:

$$A^* = \arg\max_a p(A \mid P)$$

$$\approx \arg\max_a \prod_{i=1}^{n} p(a_i \mid p_i)$$

where $p_i=\{p^1_i,..., p^t_i\}$ is the prosodic feature vector corresponding to the base unit (which will be explained in more details in Section 3). Note that this assumes that the prosodic events are independent and they are only dependent on the acoustic observations in the corresponding locations. We try to capture some of the dependency information in the features. In this paper, we use the neural network classifier in the Weka toolkit [11]. The choice of this classifier is based on our previous studies of different modeling approaches using prosodic information for accent detection [5]. We used 19 input nodes, 10 hidden layers, and 1 output node in the neural network, and the default options of Weka for training the model.

The prosodic features we use represent pitch, energy, and duration. Pitch and energy values are computed using Praat [7]. 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. We also carried out an approximation of the pitch and energy contour by taking 5 leading terms in the Legendre polynomial expansion. That is, each contour $f(t)$ (where t represents time) is approximated as

$$f(t) = \sum_{i=0}^{M} a_i L(t)_i$$

where $L(t)_i$ is the i^{th} Legendre polynomial. Each coefficient (a_i) models a particular aspect of the contour. For example, a_0 stands for the mean of the segment, a_1 is interpreted as the slope, a_2 gives information about the curvature, and a_3, a_4, a_5 model the fine details. In order for these coefficients to be comparable across different base units, it is important to carry out time normalization. All the segments must be scaled and mapped to the same interval. This approximation technique of the prosodic contours has been successfully applied in quantitative phonetics [8] and in engineering applications [9].

In total, we used 19 features in our experiments, listed below.

- Pitch range (3 features): maximum pitch, minimum pitch, and pitch range (difference between maximum and minimum pitch).
- Pitch contour (6 features): 6 coefficients of Legendre polynomial.
- Energy range (3 features): maximum energy, minimum energy, and energy range (difference between maximum and minimum energy).
- Energy contour (6 features): 6 coefficients of Legendre polynomial.
- Duration (1 feature): duration of the segment.

3. Base unit and boundary estimation

In this paper, we focus on two main factors that affect accent detection performance. The first one is the base unit used for accent detection. The second factor is boundary information. We will evaluate three different units, namely, word, syllable, and vowel, when the boundary information is available for different levels. The final decision will be at either syllable- or word-level. Therefore, the accent hypotheses from smaller units need to be aggregated to obtain the final decision.

During testing, often boundary information is not available. Attempts have been made to estimate boundary, such as [10]. In our study, in order to evaluate the impact of boundary information, we used two methods. First, we assume that each phone and word boundary is available. This is often the case from the current speech recognition systems. We estimate the syllable boundary by finding the lowest value (valley) in the energy contour between two vowel phones. In the second approach, boundary information in not available for any segments, thus we do not need to rely on having speech recognition output for accent detection. We use a moving window/frame with a fixed size. For this case, the maximum likelihood among all the frames within a syllable/word is used as the likelihood for that syllable/word for the final accent hypothesis. These settings allow us to examine the effect of boundary for different conditions — when it is fully available, partially available, or not available.

Table 1 summarizes the evaluation conditions we investigate in this paper. Figure 1 illustrates different base units and their boundaries using some example

Base units	Is boundary info. available?
I: word-based	Yes
II: syllable-based	Yes
III: vowel-based	Yes
IV: estimated syllable	Partially, phone duration is available, no syllable info.
V: fixed frame/window	No

Table 1. Five types of base units of accent detection, along with the boundary information used in them.

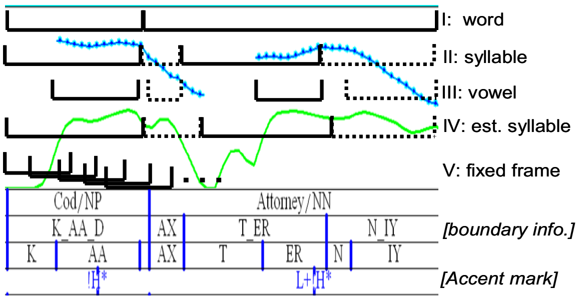

Figure 1. Example of base units and their boundaries.

words. In Figure 1, the lower part provides the boundary information of phone-, syllable-, and word-level, and accented marks. The upper part shows different base units, where a solid line is used to represent an accented segment, and a dotted line indicates non-accented. These are decided based on whether a segment includes the accented vowel or not.

4. Experiments and results

4.1. Data

In this paper, our experiments are carried out using the Boston University Radio News Corpus [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 tags, and automatic phone alignments. Among the ToBI annotated data, we use 263 utterances without any duplicated sentences. 161 utterances from speaker *f2b*, *f3b*, and *m2b* are used for training, and 102 utterances from speaker *f1a* and *m1b* are used for testing. The test utterances consist of 5,448 words and 8,962 syllables. Among them, there are 2,646 and 2,856 accented words and syllables respectively.

4.2. Results

In all our experiments, we evaluate the performance of the five types of base units as described in Section 3. The performance of each experiment is evaluated using accuracy and F-measure value for both syllable- and word-level decisions.

4.2.1. Using different units with known boundaries

First we compare the performance of different base units when boundary information is available. Three different units are used in this experiment: word, syllable, and vowels, corresponding to type I, II, and III in Table 1. For these different units, we used a matched training and testing setup, that is, the same type of unit is used in both training and testing. To obtain the final syllable- and word-level hypothesis, we perform some post-processing. For syllable level, in a vowel-based system (type III), we assume that if there is an accent hypothesis within the real syllable segment, then the syllable is accented. Note that we do not use word-based units for syllable level performance. For the word level accuracy, when using type II and III (syllable- and vowel-based units), if a word includes at least one accented unit, it is an accented word.

	Syllable		Word	
	Acc	F	Acc	F
Chance performance	0.69	-	0.52	-
I: word-based	-	-	0.84	0.85
II: syllable-based	0.84	0.73	0.79	0.78
III: vowel-based	0.78	0.61	0.73	0.71

Table 2. Accent detection performance using different base units (word-, syllable-, and vowel-based) when boundary information is available. Results are shown using accuracy (Acc) and F-measure (F) for both syllable and word level decisions.

The results are shown in Table 2. The chance performance shown in the table is by hypothesizing all the instances using the majority class. The system performance for different units is significantly better than chance performance. We observe that when the boundary information is available, using a longer unit generally yields better results. For the word level performance, type I (word-base unit) achieves the best performance. For the syllable-level results, a syllable-based setup is superior to the vowel based model.

4.2.2. Impact of boundary information

Our next experiment evaluates how system performance varies with the availability of boundary information using two settings, type IV and V in Table 1. For type IV, we use phone level boundaries and energy contour to estimate syllable boundaries. Once the estimated boundaries are available, we can use duration information in the feature set for accent detection. We use the same method to estimate syllable units during training. For type V, there is no boundary information and we use a fixed length window as the unit in the task. For this setup, we cannot use duration features since all the units have the same length. We experimented with different ways to generate training units. The best approach we found for training is to use the same window length as the test units and create only one sample per syllable centered around the middle point of the vowel within the syllable. The label for this sample is based on whether the syllable is accented or not.

The experimental results are shown in Table 3. For the fixed frame units, we show results for different window lengths. The length of the frame has some impact on the performance. The best performance is achieved when using 150ms, which is similar to the average length of the syllable in the labeled training data. We varied the step size of the moving window in our experiments and found that using 20 ms yielded the best performance. The results shown in the table correspond to this setup.

We can see from Table 3 that the performance of Type IV using estimated syllable boundary information is similar to the results of Type II in Table 2. Without using any boundary information (Type V), the results are similar to the type II results on word-level. We also investigated using estimated syllables from the energy

		Syllable		Word	
		Acc	F	Acc	F
IV: estimated syllable		0.84	0.72	0.79	0.78
V: moving windows with fixed length	250 ms	0.78	0.65	0.76	0.76
	200 ms	0.79	0.67	0.77	0.78
	150 ms	**0.80**	**0.69**	**0.79**	**0.79**
	100 ms	0.80	0.68	0.78	0.78

Table 3. Accent detection performance using different units (estimated syllable and fixed length windows) when boundary information is partially or not available.

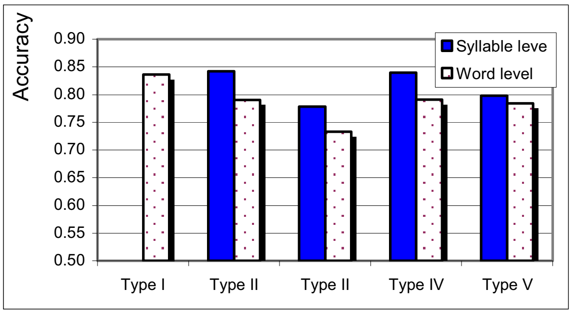

Figure 2. Comparison of 5 types of base units for accent detection. The notations for different types are the same as in Table 1. Results shown are the accuracy for syllable and word level decision.

contour when no boundary information is available, and obtained syllable- and word level accuracies of 56% and 61% respectively. These are worse than using a fixed length window for this testing condition, indicating that more sophisticated algorithms for syllable boundary estimation may be needed for better performance. The results in Table 3 are encouraging as they suggest we can still achieve reasonable performance even without the use of speech recognition to obtain word or syllable level boundary information.

In all of the above experiments, we found that the most important feature is duration when it is used. The next one is pitch range for Type I and energy maximum for Type II, III, and IV. For Type V, duration feature is not available, and we noticed that energy contour features are more important than other features.

A comparison of all the setups is shown in Figure 2 for accent detection. When there is enough boundary information, using longer segment unit achieves better performance. If there is no boundary information, we can use a moving window/frame with a fixed size without much performance loss compared to using known boundaries.

5. CONCLUSIONS

In this paper, we investigated the performance change of accent detection according to using different units and existence of their boundary information. In our experiments, we evaluated five different configurations. The results show that when boundary information is available, using word-level unit achieves better performance than syllable or vowel-based units. As expected, lack of boundary information results in performance degradation; however, using a moving window/frame with a fixed size achieves performance close to using syllable information on word-level evaluation. Currently the features used in our experiments are the same for different levels of units. We believe that some segment dependent features can improve performance and will investigate this in our future work.

6. Acknowledgment

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

[2] Chen, K., Hasegawa-Johnson, M., and Cohen, A., “An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic prosodic model”, *in Proc. of ICASSP*, pp. 509-512, 2004.

[3] Sridhar, V.K.R., Bangalore, S., and Narayanan, S., “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.

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

[5] Jeon, J.H., and Liu, Y., “Automatic prosodic events detection using syllable-based acoustic and syntactic features”, *in Proc. of ICASSP*, 2009.

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

[7] Boersma, P., “Praat, a system for doing phonetics by computer” Glot International, 5(9-10):341–345, 2001.

[8] Grabe, E., Kochanski, G., and Coleman, J., “Quantitative modeling of intonational variation”, *in Proc. of SASRTLM*, pp. 45–57, 2003.

[9] Dehak, N. Dumouchel, P., and Kenny, P., “Modeling prosodic features with joint factor analysis for speaker verification”, *in IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15: 2095-2130, 2007.

[10] Dashiell, A., Hutchinson, B., Margolis, A., and Ostendorf, M., “Non-segmental duration feature extraction for prosodic classification*”, in Proc. of Interspeech*, pp. 1092-1095, 2008.

[11] Witten, I., Frank, E., Trigg, L., Hall, M., Holmes, G., and Cunningham, S., “Weka: Practical machine learning tools and techniques with java implementation”, *in ICONIP/ANZIIS/ANNES International Workshop*, pp. 192–196, 1999.