The Detection of Emphatic Words Using Acoustic and Lexical Features

Jason M. Brenier, Daniel M. Cer†, Daniel Jurafsky‡*

Department of Linguistics*, Computer Science†, University of Colorado, Boulder, USA
Department of Linguistics‡, Stanford University, Stanford, USA
jbrenier@colorado.edu, Daniel.Cer@cs.colorado.edu, jurafsky@stanford.edu

10.21437/Interspeech.2005-576

Abstract

In this study, we describe an automatic detector for prosodically salient or emphasized words in speech. Knowledge of whether a word is emphatic or not could improve Text-to-Speech synthesis as well as spoken language summarization. Previous work on emphasis detection has focused on the automatic recognition of pitch accents. Our model extends earlier research by automatically identifying emphatic pitch accents, a subset of pitch accents that mark special discourse functions with extreme degrees of salience. The overall best performance achieved by our system was 87.8% correct, 8.0% above baseline performance. The results of a feature selection algorithm show that the top-performing features in our models are primarily acoustic measures. Our work identifies important cues for emphasis in speech and shows that it is possible for an automated system to distinguish between two levels of perceived prominence in pitch accents with a high degree of accuracy.

1. Introduction

Speakers can prosodically emphasize a word in an utterance in order to make it stand out with respect to surrounding words. The most common model of such emphasis is the presence of a pitch accent on a word [1]. Pitch accented words are points of intonational prominence in speech that are realized acoustically through increased duration and intensity and more extreme fundamental frequency ($f0$) minima and maxima.

Pitch accents, and prosodic prominence in general, reflect various aspects of discourse-pragmatic structure including information status and contrast. Prosodic prominence has also been found to correlate with incredulity and uncertainty readings of a text, question type, adverbial focus, anaphoric links, topic structure, correction, and turn-taking cues.

Research has shown that listeners can consistently perceive significant differences in prominence among pitch accented syllables [5], [6]. For the purposes of this study, we are primarily concerned with detecting a subset of pitch accents that have been shown to be categorically interpreted as distinct from neutral pitch accents [8]. The accents in this special class are perceived as having extreme degrees of salience when compared to other pitch accents within a particular intonational context and fit into the highest range of a listener's accent prominence scale. These accents may be described as conveying an acute degree of emphasis and will hereafter be referred to as *emphatic pitch accents*.

Some examples of emphatic pitch accents taken from the corpus used for our experiments are given in (1-3) (SMALL CAPS indicate pitch accents, **bold** words indicate emphatic pitch accents). These examples suggest that emphatic pitch accents mark contrast, negation, and new terms with highly significant information content. Figure 1 shows an additional example with annotation details.

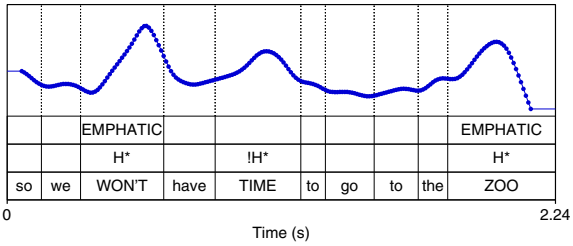

Figure 1: *$f0$ contour for a sentence from our corpus containing emphatic pitch accents.*

(1) But the **snake** is **dark** and the **grass** is **light** green.

(2) "**That's** when we use **peripheral** vision," added SAM.

(3) TOM would **never** forget to feed his FISH.

The automatic detection of emphasis has a number of applications in human language technology systems. These include the generation of improved prosody contours in unit-selection speech synthesis systems, content spotting in spoken language summarization systems, the identification of focal elements in speech understanding systems, and improved facial animation generation for interactive tutors.

Some of these tasks involve text input, while others utilize acoustic input. For this reason, we will describe experiments in which our system tags emphatic words given the following: acoustic input (*emphasis detection*), text input (*emphasis prediction*), and both acoustic and text features. In order to provide a useful comparison, we also train a detector of pitch accents that will allow us to understand which features play a role specifically for emphasis, above and beyond their role in pitch accent detection.

The paper is organized as follows. First, we describe the labeling conventions, corpus and features used for the development of our classifiers. Second, we explore the machine learning methods used for training and discuss our experimental design. Third, we provide results for our classification tasks. In the final section, we draw conclusions from our experiments and describe future directions for our research.

2. Previous Work

While there has been little previous work specifically on the detection of emphatic pitch accents, several automatic prominence classifiers have been constructed, most of which have focused on the prediction and detection of pitch accents in general. These classifiers have been trained on various acoustic and

lexical features and have been implemented using a variety of machine learning techniques.

Text-based features such as a word's part of speech and discourse status have been used to train decision trees and have achieved accent prediction accuracy of 80-98% [9]. A combination of prosodic and syntactic features have been used with a GMM-ANN hybrid model to achieve 84% accuracy [10] and with HMMs to achieve 88% accuracy [11]. A dynamical system model has shown performance around 89% [12] and ensemble learning techniques with decision trees have shown 87.2% [13]. Recently, a CRF sequence model for conversational speech data has been shown to achieve 76.4% accuracy [15], and a Bayesian Quadratic Forest classifier trained exclusively on acoustic features has achieved best performance of 61.4% on a large database of multiple speech styles [16].

One study which has aimed at detecting relative degrees of accent prominence is that of [17]. In this study, the most prominent pitch accented words in a phrase were identified with a rule-based classifier that used maximum values for acoustic features such as overall intensity and high frequency emphasis (spectral tilt). The best performance achieved for this classifier was 25% for read speech and 66% for spontaneous speech.

3. Data

A corpus of short child-directed stories read by one female native speaker of American English was used for all experiments in this study. The corpus contains a total of 4 stories with 2906 words and approximately 800 intonational phrases. Since our corpus contains data from only one speaker, our models are speaker-dependent. Unlike some of the prior pitch accent studies, we are, therefore, unable to report how well our model would perform on the speech of unseen subjects.

4. Annotation

The corpus was annotated by the first author with labels indicating the presence or absence of pitch accent and major intonational breaks, following the general ToBI standards [20]. Pitch accents that were unambiguously perceived as being emphatic when compared to other pitch accents in the surrounding intonational context were labeled as emphatic pitch accents.

Each sentence in the corpus was automatically aligned with its corresponding audio segment at the word and phone levels using the SONIC continuous speech recognizer [21]. The phone sequences for each word were then syllabified using the NIST syllabification package.

The distribution of the prominence classes in our corpus is shown in Table 1. Note that the set of emphatic words was entirely contained within the set of all pitch accented words; words labeled as emphatic were also labeled as being pitch accented. Non-emphatic pitch accented words are approximately 2.5 times more frequent than emphatic pitch accented words. Because of the genre of this corpus (emotive child-directed speech), it contains a large number of both emphatic and neutrally accented words. This makes the corpus useful for studies of emphasis, but may limit the generalizability of our results to other genres.

5. Input Feature Vectors

Following annotation, we extracted both acoustic and text-based features from our corpus. The acoustic features were extracted for each word in the corpus using the Praat sound analysis package [22]. The text-based features were generated using only the orthographic transcript of the stories in the corpus as a main source.

Table 1: *Distribution of accent prominence labels*

Prominence Class	Frequency	Percentage
unaccented words	1389	47.8%
accented words	1517	52.2%
Total Number of Words	**2906**	
non-emphatic words	2319	79.8%
emphatic words	587	20.2%
Total Number of Words	**2906**	

5.1. Acoustic Features

Prior research on the perception and detection of prosody has shown that acoustic features based on $f0$, duration, and intensity are all indicators of prosodic prominence (see a review in [23]). In addition, changes in pitch range have been found to signal a distinction between normal and emphatic accents [8]. We relied on these findings in selecting the acoustic features for our system.

We extracted a variety of acoustic features for each stressed syllable, word and intonational phrase in our corpus. The 12 base acoustic features used in our experiments are provided in Table 2. Variants of each of these features were added to the feature set after normalization. We also included values for the raw and normalized features of the immediately neighboring words. Our final acoustic input vector contained 486 features; 18 basic features (12 plus an extra 6 including each $f0$ feature in Hz and ERB) times 3 words (current, previous, following), times 3 normalizations (unnormalized, normalized by word, and normalized by IP), times 3 (once for each word, once for each words' main syllable, and once for the surrounding intonational phrase).

5.2. Text Features

Table 2 presents the text-based features used in our experiments. These features were extracted from the text of the stories in the corpus. The word informativeness measures of negative log frequency and TF*IDF were included based on successful experiments which implemented these features in a pitch accent prediction model [26]. TF*IDF was defined as the frequency of a word in a particular document (Term Frequency) multiplied by the logarithm of the ratio of the total number of documents in the corpus to the total number containing that word (Inverse Document Frequency). Word class information was captured using both broad categories following [9] (open class, closed cliticized, closed unaccented, closed accented) as well as the full set of Penn Treebank tags. Both sets of word class tags were assigned using the Maximum Entropy Part-of-speech Tagger [27]. Each word was also labeled as forming part of an expression of negation or exclamation. The final two features in Table 2 rely on intonation phrase boundary information; for the current experiments we used hand-labeled boundaries.

6. Experiments

We used a Maximum Entropy classifier for all experiments. This technique requires a binary input vector for training, so we performed feature discretization on all continuous features using 2.5% CDF intervals as bins.

6.1. Feature Selection

The maximum entropy framework supports feature selection, which automatically chooses a subset of features that are most

Table 2: *Principal feature set*

Acoustic Features	Text Features
Duration	Word, syllable identity
$f0$ min, max, mean	Total number of words in ip
$f0$ excursion, slope, stdv	Word position value in ip
Intensity min, max, mean	TF*IDF
Intensity excursion, stdv	Negative log frequency
	Broad/specific word class
	Exclamation, negation

useful for a given task. Feature selection is done iteratively, at each pass selecting the single feature which most improves the model. The order in which the features are selected by the algorithm can be used as a rank ordering of the features in terms of their importance to the classification task.

6.2. Feature Sets

In our first set of experiments, we trained models using the full set of acoustic and text-based features. These experiments were meant to model prominence detection tasks in which speech is provided and text can be derived through automatic speech recognition procedures. Two main experiments were conducted using the joint feature set. The first specifically measured performance on emphatic pitch accent detection based on a binary classification of a word as emphatic/non-emphatic. The second measured performance on overall pitch accent detection based on a binary classification of a word as accented/unaccented.

In our other experiments, we also performed these two classification tasks, but our system was trained on either acoustic features alone or text features alone. The text-only experiments were meant to emulate prominence prediction tasks that might be necessary in TTS applications.

7. Results and Discussion

All results reported here are based on a 4-fold cross-validation of each system in which training was performed on three stories and testing was performed on the remaining unseen fourth story. This method was chosen to avoid unfair testing on text from a story that was also used for training. The comparison baseline for the emphatic pitch accent tasks is defined as the performance of a system that assigns the majority class label *non-emphatic* to all test tokens. The baseline for the pitch accent tasks is defined as the performance of a system that assigns the majority class label *accented* to all test tokens. Tables 3 and 4 summarize performance in the six separate experiments that were conducted.

Table 3: *Overall correct classification rate*

	Baseline	Joint	Acoustic	Text
Emphasis	79.8%	87.8%	87.1%	79.7%
Accent	52.2%	84.4%	78.2%	78.4%

Overall, the emphatic pitch accent systems performed reasonably well. When trained on both acoustic and text features, our system performed 8.0% above baseline. With only acoustic features, performance was only slightly degraded. When trained only on text features, the system performed roughly at baseline, considerably worse than either the full feature set or the acoustic features alone. This is not surprising, considering claims as to the importance of acoustic cues in the perception of prominence.

Our pitch accent system performed within range of other state-of-the-art pitch accent classification systems. The system trained on all features achieved performance that was 32.2% above baseline. Unlike in the emphasis task, the text-based features alone performed slightly better than the acoustic features alone, although the difference was minimal. These results suggest that it was the combination of both acoustic and text-based features that boosted performance for this task. Moreover, it appears that the text-based features that we chose were appropriate for pitch accent classification but less so for emphasis classification. This demonstrates that there are indeed differences in the nature of emphasis and accent classification. We suspect that more appropriate text-based features for emphasis detection might be semantically or pragmatically oriented.

Given that all words with emphatic pitch accents also fall into the class of words containing a pitch accented syllable, we also wanted to consider how knowledge of whether a word contained a pitch accented syllable would affect emphatic pitch accent detection performance. A post-hoc oracle experiment in which the hand-labeled presence or absence of accent in a word was included with all other acoustic and text-based features showed that performance for the emphasis detection task only increased to 88.6%.

Table 4: *Emphasis: Performance results for each feature set*

P=precision, R=recall

		Predicted Class	
JOINT, P=39.4% R=28.6%		**Non-emphatic**	**Emphatic**
Actual Class	**Non-emphatic**	2058	258
	Emphatic	419	168
ACOUSTIC, P=33.3% R=31.5%		**Non-emphatic**	**Emphatic**
Actual Class	**Non-emphatic**	1946	370
	Emphatic	402	185
TEXT, P=42.4% R=25.6%		**Non-emphatic**	**Emphatic**
Actual Class	**Non-emphatic**	1963	353
	Emphatic	437	150

8. Feature Selection Results

The Maximum Entropy feature selection method described above was performed on the full feature sets and text-only feature sets for both the emphasis and accent tasks. The top contributing features for the emphasis task are shown in Table 5.

Table 5: *Top 10 Ranked Features: Emphasis*

Joint Features	Text Features
Max intensity (word)	Word ident
$f0$ excursion (Hertz) (word)	Total words ip
Syll dur	Neg log freq word
Syll mean $f0$ (Hertz)	Ident prev word
ip-normed mean $f0$ (ERB) (word)	TF*IDF
Word ident	TF*IDF prev word
ip-normed max intensity (word)	Exclamation
Syll normed duration	Neg log freq of prev word
ip-normed $f0$ excursion (Hertz)	Ident foll word
$f0$ mean (ERB) (IP)	Ident stressed syll prev word

The top five features in the emphasis detection task that integrated the joint feature set were responsible for over 99% of the entire system performance. As expected, measures of intensity, $f0$, and duration were highly effective, taking the top three positions. In fact, acoustic features consistently outranked the text-based features (with the exception of word identity), suggesting that the text-based features that we chose were inappropriate for the task. Both syllable and word-based acoustic features were valuable, and features normalized within the domain of the intonational phrase made a reasonable contribution.

Fewer normalized features ranked in the top ten for the accent detection task. Pitch accent detection may be more dependent upon raw feature values in our system due to the greater differences between accented and unaccented words. Classifying a word according to its relative accent prominence may require a more fine-tuned metric for comparison to other words in the intonational phrase.

The preceding and following syllable information was not effectively used by this system, most likely because syllable information was gathered only for the canonically stressed syllable of a word. This method may not allow the model to make generalizations regarding rhythm and the interaction of adjacent syllables. This could be due to the fact that the stressed syllable of the preceding or following word wasnt always adjacent to the stressed syllable of the target word. In this case, information about intervening unstressed syllables is unavailable to the system. A solution to this problem would be a restructuring of the task so that classification is performed for syllable units rather than word units.

Word identity was fairly highly ranked as a feature in the system. This could be attributed to the fact that words occurring two or more times in the corpus carrying an emphatic pitch accent accounted for 61.7% of emphatic tokens in the corpus. For the text-only emphasis experiment, word identity was the top-ranked feature.

Both measures of information content were also important features in emphasis prediction. Indeed, the majority of emphasized words in the corpus tended to be semantically rich words, so it makes sense that measures of the information content of a word would aid in classification.

Although word class information proved to be helpful in prior pitch accent prediction studies, it did not play a major role in the emphasis detection task. This may be due to the fact that words realized with emphatic pitch accents make up only a small subset of the pitch accented words, and emphatic words may only account for a small percentage of any given word class. Moreover, we have seen evidence in our corpus that a large percentage of the words that are given special emphasis are words that typically may not have received a standard pitch accent. This seems to be mostly motivated by the special functions that emphasis achieves, such as contrast. In our pitch accent prediction task, two measures of word class rank among the top ten contributing features: the specific word class of the target word, and the broad word class of the following word.

9. Conclusions

Overall, both our emphatic pitch accent detector and our pitch accent detector performed well when using full feature sets. Our feature analyses showed that normalized intensity, duration, and $f0$ features were effective for emphasis detection. The text-based features we chose were not substantial contributors in our emphasis experiments, but they were in our pitch accent experiments. Due to fundamental differences in the emphasis and accent tasks, it is our belief that text features that are more closely tied to discourse function and semantic content could improve emphasis prediction.

Our work has shown that it is possible for an automated system to distinguish between two levels of perceived prominence in pitch accents and that this system can label words as emphatic or non-emphatic based on this distinction with a high degree of accuracy. We hope that the prominence processing system developed in our experiments can be used to improve other natural language processing systems such as speech synthesizers.

10. Acknowledgements

This research was partly supported by the Edinburgh-Stanford LINK and the NSF via IIS-0325399.

11. References

[1] Bolinger D. “A theory of pitch accent in English.” Word, 33:3-20, 1958.

[5] Terken, J. “Fundamental frequency and perceived prominence of syllables.” JASA 89:1768-1776, 1991.

[6] Gussenhoven, C. and Rietveld, T. “Fundamental frequency declination in Dutch: Testing three hypotheses.” Journal of Phonetics, 16, 355-369, 1988.

[8] Ladd, D. R. and Morton, R. “The perception of intonation emphasis: Continuous or categorical?” Journal of Phonetics, 25: 313-342, 1997.

[9] Hirschberg, Julia. “Pitch accent in context: Predicting intonational prominence from text.” Artificial Intelligence, 63:305-340, 1993.

[10] Chen, K. and Hasegawa-Johnson, M. “An Automatic Prosody Labeling System Using ANN-Based Syntactic-Prosodic Model and GMM-Based Acoustic-Prosodic Model.” Proc. of ICASSP, 2004.

[11] Conkie, A., Riccardi, G., and Rose, R. C. “Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events.” Proc. of Eurospeech, Budapest, Hungary, pp. 523-526, 1999.

[12] Ross, K. and Ostendorf, M. “A dynamical system model for recognizing intonation patterns.” Proc. Eurospeech, Madrid, pp. 993-996, 1995.

[13] Sun, X. “Pitch Accent Prediction Using Ensemble Machine Learning.” Proc. of ICSLP, 2002.

[15] Gregory, M. and Altun, Y. “Using Conditional Random Fields to Predict Pitch Accents in Conversational Speech.” Proc. of ACL, Barcelona, pp. 677-683, 2004.

[16] Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. “Loudness Predicts Prominence; Fundamental Frequency Lends Little.” JASA, (submitted)

[17] Heldner, M., Strangert, E., and Deschamps, T. “A focus detector using overall intensity and high frequency emphasis.” Proc. of ICPhS, San Fransisco, 1999.

[20] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. “ToBI: a standard for labeling English prosody.” Proc. of ICSLP, vol. 2, pp. 867-870, 1992.

[21] Pellom, B. “Sonic: The University of Colorado Continuous Speech Recognizer, Technical Report TR-CSLR-2001-01.” CSLR, University of Colorado, 2001.

[22] Boersma, P., and Weenink, D. “Praat: doing phonetics by computer (Version 4.3.01)” [Computer program]. Retrieved from http://www.praat.org/, 2005.

[23] Terken, J. and Hermes, D. “The perception of prosodic prominence.” In M. Horne (ed.): Prosody: Theory and experiment. Studies presented to Gösta Bruce. Dordrecht: Kluwer, pp. 89-127, 2000.

[26] Pan, S. and McKeown, K. “Word Informativeness and Automatic Pitch Accent Modeling.” Proc. of EMNLP/VLC, College Park, Maryland, 1999.

[27] Rhatnaparkhi, A. “A Maximum Entropy Part-Of-Speech Tagger.” Proceedings of EMNLP, May 17-18, 1996.