ACOUSTIC-BASED PITCH-ACCENT DETECTION IN SPEECH: DEPENDENCE ON WORD IDENTITY AND INSENSITIVITY TO VARIATIONS IN WORD USAGE

Anna Margolis and Mari Ostendorf

University of Washington
Department of Electrical Engineering, Seattle, WA

ABSTRACT

Past work has produced fairly accurate automatic pitch-accent detectors, but it has often been noted that the accent class of a word is highly dependent on word identity, with some words and word types usually being accented and others not. We argue that a good accent detector should not only have high overall accuracy, but also be able to distinguish between accented and unaccented variants of the same word. We report on experiments with several classifiers trained on a hand-labeled corpus, using a large set of acoustic features. Results show that while the classifiers have a high overall accuracy, they perform disappointingly on words with atypical accent status or whose prior accent status is more uncertain. We further report on attempts to improve the performance on these sub-tasks via feature selection and engineering of the training set.

Index Terms— Speech analysis, Speech understanding, Prosody

1. INTRODUCTION

Since the invention of prosodic annotation standards such as Tones and Break Indices (ToBI) [1] and the availability of annotated corpora, there has been much interest in building classifiers to automatically recognize symbolic prosody classes. This is potentially useful for many applications areas, including both spoken language processing and speech annotation for data-driven speech synthesis. A variety of feature types are used in these classifiers: "acoustic" features computed from the speech waveform, based on pitch, energy, or duration, are often combined with "lexical" (or textual) features based on syllable or word identity, part-of-speech, or syntax [2, 3, 4, 5]. It has been noted [6, 7, 8, 9] that word identity is often an effective predictor of word-level pitch accent label: rare words, multi-syllable and content words are very likely to be accented, for instance, while function words are likely to be unaccented. Unfortunately, a prosodic classifier based on such a simple heuristic is of limited use for many applications. When the identity of the word is already known, we would like to be able to use acoustic features to derive additional information, which may be useful for distinguishing multiple meanings or understanding intent. For instance, acoustic features should be able to distinguish between "Are you from *Seattle?*" [or somewhere else?] vs. "Are you *from* Seattle?" [or just visiting?].

Here we explore this idea with several experiments on the Boston Radio News Corpus (BU-RNC) [10], focusing on the binary accent vs. no-accent classification problem. While this is a read corpus, the general approach of our analysis will be useful for conversational speech as well. We train classifiers to detect accented words, using a large set of acoustic features, then analyze how well the classifiers perform on specific subsets of the test set, including matched pairs of accented and unaccented versions of word tokens and words with atypical accent status (unaccented versions of words that are usually accented and visa versa). Experiments using three types of classifiers show that even when different learning strategies have roughly the same performance on a general set, they may give very different results on the more difficult cases. Analysis of features used in training with general data vs. matched pairs shows that even with acoustic features alone, word identity effectively plays a major role when training on general data via the word duration feature. In training with matched pairs, fundamental frequency cues are more important. However, using these findings to improve the sensitivity of the classifiers is mostly unsuccessful—the accuracy on the difficult subsets remains low, suggesting the need for improvements in feature extraction techniques.

2. RELATED WORK

Many researchers have investigated automatic detection of pitch accent, prominence, or stress in speech at both the syllable and word level.[1] 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 pitch, energy, and duration. Early work used acoustic and a few lexical features in a decision tree with a Markov sequence model to detect combinations of accent and boundary tones [11]. Subsequent work improved performance through better feature sets and more sophisticated models; for instance, [3, 4, 2] achieved 84-88% accuracy at the word or syllable level on the Boston Radio News Corpus.

Certain words and word classes tend to be emphasized more than others, so features tied to word identity are very useful for predicting prominence. Several studies have shown the relationship between part-of-speech (POS) and pitch accent; it is well-known that POS classes associated with function words are less likely to be accented. The relationship between word frequency and ToBI accent status is explored in [7], showing that on average, unaccented words tend to be frequent terms, but it is not uncommon for frequent terms to be accented. In [12], sophisticated linguistic features like "contrast", "animacy" and "information status" were compared with simpler lexical features like POS and word frequency, demonstrating the surprising result that the simpler features work quite well on their own and are not helped by the more sophisticated ones. Word identity itself may be even more predictive, however, as has been suggested in [8]. That work used a feature called "accent ratio", defined as the fraction of times the word was accented in the training corpus if the fraction is significantly different from 0.5, and 0.5 otherwise. They analyzed several other lexical features as well, including stop-word status, un-

[1] In keeping with previous work, and since we deal only with a word-level, binary accent detection problem, we use the term "accent" as synonymous with "prominence", "phrasal stress", or "emphasis."

igram and bigram probability, TF-IDF, number of characters in the word, position in the utterance, POS, and dialogue act of the utterance. Accent ratio was found to be the best, and by itself was able to classify 75% of words correctly; adding other features improved accuracy only slightly. However, like most authors, they reported results only in terms of overall accuracy on the full test set, and did not analyze the types of errors made by the classifiers. A brief error analysis was conducted in [12], noting that one common mistake of their classifier was missing accents in function words, which are not usually accented. They also noted that the variations in possible accent placement in some phrases presents a limit on how well actual accent placement can be predicted from textual features.

Acoustic features have the potential to correct these deficiencies and give hints where the text is ambiguous. However, past work showed that classifiers use acoustic features to effectively learn lexical information, and suggested that this lexical information was used heavily in decisions. In [6], the authors demonstrated that, for predicting word-level accent, un-normalized duration features are more effective than their counterparts normalized by number of syllables or expected word duration, due to the fact that un-normalized duration features encode information about word identity that is removed from the normalized versions. In our study, we also find that un-normalized word duration is more useful than any other feature.

Our work involves a feature analysis that compares usefulness of features for distinguishing accented/unaccented word classes vs. for distinguishing accented/unaccented versions of word tokens. This differs from related work such as [13], which compared the usefulness of features for detecting accent vs. for detecting contrast, and [14], which analyzed features for predicting focus and accent.

3. METHODS

The Boston Radio News Corpus is commonly used in prosodic classification research. It consists of news stories read by professional radio announcers, and is partially annotated with accents, boundary tones, and boundary break indices, based on the ToBI prosodic labeling conventions for American English [1]. In this paper we focus on the accent annotations, collapsing all word accent tone types (excluding boundary tones) to a single "accent" class, leading to a binary accent-vs-non-accent classification problem as has been treated in other works. Our experiments are based on the prosodically-annotated portion from speaker 'f2b', which is the speaker most extensively annotated. This portion contains paragraphs from 32 radio stories with a total of 9091 annotated words, of which about 55% bear accents. We use heldout data from the radio section for testing (rather than the standard labnews subset), since this work aims at analysis to improve system design. Because of the small amount of data we perform 6-fold cross-validation testing on the radio set; each story is used in exactly one test set, and there is no overlap in stories between a train set and its corresponding test set. The average size of a test set is 1515 words, with a max of 1728 and a min of 1208 words. In the experiments that follow, we report the mean and standard deviation of the results over the 6 folds.

To analyze how well the classifiers are able to distinguish accented and unaccented variants of words, we constructed a "matched" test set (a subset of the main test set for each fold) composed of pairs of identical word tokens pulled from the accented and unaccented classes. For example, if the word "school" appears three times accented and once unaccented, we use the unaccented instance and one (randomly selected) of the three accented instances. If a classifier learned only concepts related to how certain words tokens are *usually* classified, it would not perform well on this test set. The mean size of this set (over the 6 folds) is 207 words.

The matched test set is exactly 50% words from the accented class, whereas the frequency in training is about 55%. A classifier trained on the usual distribution will tend to slightly err on the side of the accented class. To eliminate this as an issue in our analysis, we use a training set that is exactly 50% accented. This is achieved by using all the unaccented words in the full training division, and a same-size subset of the accented words. We refer to this as the "full" training set, to distinguish it from the engineered training sets described later; its mean size is 7576 words.

The features, based on those described in [15], include a variety of pitch, energy and duration features derived from the speech waveform and phone-level forced alignment of the transcriptions. Although each feature is associated with the word, some features are based on context, e.g., the next word or a window after the word boundary. Our classification experiments treat each word and its features independently. The duration features include word duration, time into the current story, pause duration (before and after), and several duration features using phone alignments: last rhyme, stressed rhyme, last vowel, stressed vowel, max and average vowel and phone duration, where the stressed syllable was derived from a stress dictionary. These duration features have multiple versions corresponding to different normalizations using the within-story phone duration statistics or a phone-duration statistics table derived from a separate broadcast news corpus. Word duration is included both raw and normalized by the sum of the mean durations of phones in the word. We used the n-normalized versions of features, rather than the z-normalized ones, in order to limit the size of the feature vector and for clarity of the results. (In preliminary analyses, the two versions seemed to be mostly redundant, with the n-version more often ranked above the z-version in utility.) We use a total of 69 numeric features, plus two categorical features specifying the rising/falling pitch and energy pattern at the boundary after the word.

We used three different classifiers: BoosTexter[2], a boosting algorithm based on single-feature decisions; decision trees, as implemented in IND[3]; and a Gaussian linear classifier, as implemented in MATLABArsenal[4]. BoosTexter was trained with 200 rounds in all cases. We used the cross-validation pruning option in IND with the recommended 10 folds. All 71 features are used in the BoosTexter and IND classifiers; only the 69 numeric features are used in the Gaussian classifier.

4. EXPERIMENTS AND RESULTS

4.1. Error Analysis

Table 1, columns 1 and 2, shows the results of the three classifiers trained on the full training set, and tested on the full test set and matched test set. The classifiers all do roughly equally on the general test set, but performance on the matched test set is at least 10 points worse in all cases, and the Gaussian does especially badly.

We also look at performance on words with "unusual usage"—unaccented words that are usually accented and visa versa. We use the concept of "accent ratio", the fraction of occurrences (in the training set) that a word appears accented, first introduced in [9] for analysis of accent usage and used in [8] as a feature. Since our training set is small, and in order to minimize noise, we consider only words that occur more than 5 times in the training set. We define

[2]http://www.cs.princeton.edu/~schapire/boostexter.html
[3]http://opensource.arc.nasa.gov/software/ind/
[4]http://www.informedia.cs.cmu.edu/..
/yanrong/MATLABArsenal/MATLABArsenal.htm

"unusual usage" to mean either (a) words with accent ratio above 0.8 that are unaccented, or (b) words with accent ratio below 0.2 that are accented. Performance on this set is shown in Table 1, column 3. The set is small (average of 36 words), so the results have high variance. The classifiers all do much worse than average on this set, but still are able to detect unusual usage about 60% of the time, which means that they are not simply classifying all short function words as unaccented and all long content words as accented. The best classifier on the full test set (BoosTexter) is also the best on the matched and unusual usage tests.

Table 1. Accuracy of classifiers trained on full training set, tested on full test set, matched test subset, and "unusual usage" subset

	general test	matched	unusual usage
BoosTexter	88.0 ± 1.8	74.1 ± 4.7	61.9 ± 12.4
IND	86.3 ± 1.7	69.2 ± 3.5	57.8 ± 13.0
Gaussian	87.1 ± 1.4	64.2 ± 4.3	61.6 ± 9.2

We next analyze the results in the full test set to see how the classifiers do on different types of words. We then bin the word instances in the test set by their accent ratio using uniform intervals and plot the accuracy in each bin, e.g. the data point at 0.3 represents the accuracy in the bin [0.2,0.4). Figure 1 shows the results. The classifiers get about 95% accuracy on words with accent ratio less than 0.2, which are usually unaccented, and worse than average on those in the middle range (0.2 to 0.8). Note that we are using only acoustic features, so it is not possible for the classifier to directly identify the accent ratio of a word. However, almost all of the low accent-ratio words are function words ("the", "an", "in", "was", etc.). As has been pointed out elsewhere [6], the classifier learns to recognize this class of words by the un-normalized word duration feature (and, we should note, possibly other features as well).

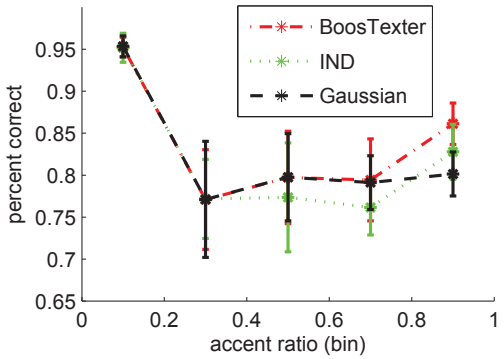

Fig. 1. Results by accent ratio, trained on the general training set. The error bars represent one standard deviation in each direction.

4.2. Feature Analysis

As we and others have noted, un-normalized duration features are related not just to the actual accent status, but also to the propensity of the word to be accented, since very short words are mainly function words. We performed an analysis to determine which features differ most between accented and unaccented versions of word tokens. We collected matched pairs of accented/unaccented word tokens from the training set (just like the matched-pairs test set used above), and computed the difference of each feature for each pair. We then computed the t-statistic for the feature differences—this is basically a matched t-test, where the samples are matched by word identity. (We did this for only one of the training sets; the set contained 747 matched pairs). When ranked by the magnitude of their t-statistic, the top 6 features are pitch features. The top 20 include word duration (both normalized and raw) and several segmental duration features (stressed rhyme, last, average and max vowel duration, average and max phone duration). Two energy features are in the top 20, which measure differences in the max and average energy between this word and the next.

For comparison, we looked also at feature differences between the accented and unaccented classes in the full training set. We computed the difference in means between the two classes using a two-sample t-test with unequal variance. When the features were ranked, pitch features were less important overall: the top feature (by far) was un-normalized word duration. Only two of the top 10 features were pitch features, whereas in the matched test, 7 of the top 10 were pitch features. On the other hand, there was general overlap in the features found in the top 20.

We also looked at feature distributions for the general accented and unaccented classes, and for accented and unaccented classes in the matched set. For many features, there was a clear difference between the matched and general accented and unaccented classes. The most extreme example of this is un-normalized word duration (word-dur), shown in Figure 2 (a). This feature appears to be very useful in the general case, with a dramatic difference between the accented and unaccented distributions, but is much less so in the matched case. Its discriminating power in the general training set is largely, but not entirely due to the difference in kinds of words appearing in the two sets. By contrast, Figure 2 (b) shows the distributions of "f0k-maxk-mode-n", which was the best pitch feature in both the general and matched t-tests (and the best feature overall in the matched case). The distributions are more similar; this feature appears more indicative of actual accent status. For many other features, the distribution differences were simply less pronounced between the matched classes than in the general case.

4.3. Can we do better?

We hypothesized that the classifiers trained on the full set might tend to learn a model of word classes that are usually emphasized, and that through engineering of the training set we might be able to force them to focus on cues related to actual emphasis. We first tried training on the "matched" training set. The average size of this training set is 1503 words, so it is much smaller than the full training set. This reduction in size is a cost of the matched training approach.

Table 2 shows the results of these matched-train classifiers tested on the full test set, the matched test subset and the "unusual" usage test subset. We first note that the accuracy on the matched test set remains significantly below the accuracy achieved on the general test set in Table 1; furthermore, results on the general test set are now much worse than before, yet still better than those on the matched test set. There is some improvement on unusual usage and matched test subsets: all three classifiers appear to perform better on unusual usage, although for IND and BoosTexter the difference is within the standard deviation of the old results. The most dramatic improvement is the Gaussian classifier, which had originally

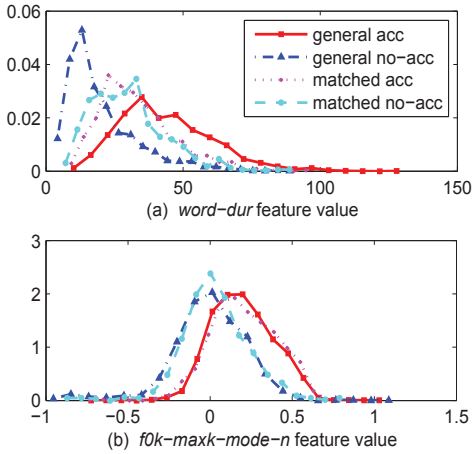

Fig. 2. Empirical distributions of two features in the general accented and unaccented classes, vs. the matched accented and unaccented classes

performed the worst on the matched test set. It now performs best on both the matched and unusual usage sets. By contrast, IND and BoosTexter improved little or not at all on the matched test set. We believe this is due to the fact that these classifiers have the ability to effectively select or weight features differently for different word types; the Gaussian treats the features independently, and so cannot produce a different decision rule for short vs. long words. Features like word duration that appear highly discriminative as a whole will dominate the decision. The matched training set did not generally help performance in the middle range of accent ratios–with a few exceptions, accuracy was worse in all the accent ratio bins, although it remained higher in the lowest bin than in the others.

We also did a few feature selection experiments with BoosTexter, using just pitch and energy features or everything except unnormalized word duration. This only slightly worsened results on the general set and did not significantly improve results on the matched set. This suggests that it might simply be harder to classify the words in the matched set; another possibility is that lexical information is learned from pitch and energy features as well, an issue that should be studied further. However, general overlap in the important features from the matched and general t-tests suggests that feature selection is not a promising method to reduce the disparity in accuracy.

Table 2. Accuracy trained on matched training set, tested on full test set, matched test subset and "unusual usage" subset

	general test	matched	unusual usage
BoosTexter	79.7 ± 1.5	75.9 ± 3.9	72.6 ± 10.7
IND	74.9 ± 1.9	71.2 ± 3.8	68.9 ± 6.9
Gaussian	79.5 ± 2.1	78.0 ± 4.2	80.2 ± 12.4

5. CONCLUSION

Our results suggest that for some classifiers such as linear Gaussian, training with an engineered training set may improve the ability to discriminate accented and unaccented versions of words and correctly classify words with unusual accent status. However, there appears to be a ceiling on how well the classifiers can do on these subtasks that is well below the overall accuracy rate, as typically reported. In future work on accent detection, researchers might consider comparing classifiers not just on the basis of overall accuracy but also on things like matched word pairs and unusual usage words, since classifiers that appear to perform the same on overall accuracy can be very different on these important subtasks. Our results also suggest the importance of pitch features. One avenue of future work might be to develop features that better capture the difference between accented and unaccented versions of word tokens. For instance, vowel-normalization of pitch features might allow greater sensitivity to differences when the same vowel appears accented and unaccented.

6. REFERENCES

[1] K. Silverman, M. Beckman, et al., "ToBI: A standard for labeling English prosody," in *Proc. ICSLP*, 1992, pp. 867–870.

[2] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 216–228, 2008.

[3] X. Sun, "Pitch accent prediction using ensemble machine learning," in *Proc. ICSLP*, 2002.

[4] 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. ICASSP*, 2004.

[5] Michelle L. Gregory and Yasemin Altun, "Using conditional random fields to predict pitch accents in conversational speech," in *Proc. ACL*, 2004.

[6] A. Batliner, E. Nöth, et al., "Duration features in prosodic classification: Why normalization comes second, and what they really encode," in *Proc. of the ISCA Tutorial and Research Wksp on Speech Recognition and Understanding*, 2001.

[7] J. F. Pitrelli, "ToBI prosodic analysis of a professional speaker of American English," in *Proc. ICSA–Speech Prosody 2004*.

[8] A. Nenkova, J. Brenier, et al., "To memorize or to predict: Prominence labeling in conversational speech," in *Proc. HLT-NAACL*, 2007.

[9] J. Yuan, J. M. Brenier, and D. Jurafsky, "Pitch accent prediction: Effects of genre and speaker," in *Proc. Interspeech*, 2005.

[10] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Tech. Rep., Boston University, March 1995.

[11] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 469–481, 1994.

[12] Jason M. Brenier, Ani Nenkova, et al., "The (non)utility of linguistic features for predicting prominence in spontaneous speech," in *Proc. IEEE/ACL 2006 Wksp on Spoken Language Technology*, 2006.

[13] A. Nenkova and D. Jurafsky, "Automatic detection of contrastive elements in spontaneous speech," in *Proc. ASRU*, 2007.

[14] Sasha Calhoun, "Predicting focus through prominence structure," in *Proc. Interspeech*, 2007.

[15] E. Shriberg, A. Stolcke, et al., "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.