Detecting Pitch Accents at the Word, Syllable and Vowel Level

Andrew Rosenberg
Columbia University
Department of Computer Science
amaxwell@cs.columbia.edu

Julia Hirschberg
Columbia University
Department of Computer Science
julia@cs.columbia.edu

Abstract

The automatic identification of prosodic events such as *pitch accent* in English has long been a topic of interest to speech researchers, with applications to a variety of spoken language processing tasks. However, much remains to be understood about the best methods for obtaining high accuracy detection. We describe experiments examining the optimal **domain** for accent analysis. Specifically, we compare pitch accent identification at the syllable, vowel or word level as domains for analysis of acoustic indicators of accent. Our results indicate that a word-based approach is superior to syllable- or vowel-based detection, achieving an accuracy of 84.2%.

1 Introduction

Prosody in a language like Standard American English can be used by speakers to convey semantic, pragmatic and paralinguistic information. Words are made intonationall prominent, or *accented* to convey information such as contrast, focus, topic, and information status. The communicative implications of accenting influence the interpretation of a word or phrase. However, the acoustic excursions associated with accent are typically aligned with the lexically stressed syllable of the accented word. This disparity between the domains of acoustic properties and communicative impact has led to different approaches to pitch accent detection, and to the use of different domains of analysis.

In this paper, we compare automatic pitch accent detection at the vowel, syllable, and word level to determine which approach is optimal. While lexical and syntactic information has been shown to contribute to the detection of pitch accent, we only explore acoustic features. This decision allows us to closely examine the indicators of accent that are present in the speech signal in isolation from linguistic effects that may indicate that a word or syllable may be accented. The choice of domain for automatic pitch accent prediction it also related to how that prediction is to be used and impacts how it can be evaluated in comparison with other research efforts. While some downstream spoken language processing tasks benefit by knowing *which* syllable in a word is accented, such as clarification of communication misunderstandings, such as "I said **un**lock the door – not lock it!", most applications care only about which *word* is intonationally prominent. For the identification of contrast, given/new status, or focus, only word-level information is required. While the performance of nucleus- or syllable-based predictions can be translated to word predictions, such a translation is rarely performed, making it difficult to compare performance and thus determine which approach is best.

In this paper, we describe experiments in pitch accent detection comparing the use of vowel nuclei, syllables and words as units of analysis. In Section 2, we discuss related work. We describe the materials in Section 3, the experiments themselves in Section 4 and conclude in Section 5.

2 Related Work

Acoustic-based approaches to pitch accent detection have explored prediction at the word, syllable, and

Proceedings of NAACL HLT 2009: Short Papers, pages 81–84,
Boulder, Colorado, June 2009.

vowel level, but have rarely compared prediction accuracies across these different domains. An exception is the work of Ross and Ostendorf (1996), who detect accent on the Boston University Radio News Corpus (BURNC) at both the syllable and word level. Using CART predictions as input to an HMM, they detect pitch accents on syllables spoken by a single speaker from BURNC with 87.7% accuracy, corresponding to 82.5% word-based accuracy, using both lexical and acoustic features. In comparing the discriminative usefulness of syllables vs. syllable nuclei for accent detection, Tamburini (2003) finds syllable nuclei (vowel) duration to be as useful to full syllables. Rosenberg and Hirschberg (2007) used an energy-based ensemble technique to detect pitch accents with 84.1% accuracy on the read portion of the Boston Directions Corpus, without using lexical information. Sridhar *et al.* (2008) obtain 86.0% word-based accuracy using maximum entropy models from acoustic and syntactic information on the BURNC. Syllable-based detection by Ananthakrishnan and Narayanan (2008) combines acoustic, lexical and syntactic FSM models to achieve a detection rate of 86.75%. Similar suprasegmental features have also been explored in work at SRI/ICSI which employs a hidden event model to model intonational information for a variety of tasks including punctuation and disfluency detection (Baron et al., 2002). However, while progress has been made in accent detection performance in the past 15 years, with both word and syllable accuracy at about 86%, these accuracies have been achieved with different methods and some have included lexico-syntactic as well as acoustic features. It is still not clear which domain of acoustic analysis provides the most accurate cues for accent prediction. To address this issue, our work compares accent detection at the syllable nucleus, full syllable, and word levels, using a common modeling technique and a common corpus, to focus on the question of which domain of acoustic analysis is most useful for pitch accent prediction.

3 Boston University Radio News Corpus

Our experiments use 157.9 minutes (29,578 words) from six speakers in the BURNC (Ostendorf et al., 1995) recordings of professionally read radio news. This corpus has been prosodically annotated with full ToBI labeling (Silverman et al., 1992), including the presence and type of accents; these are annotated at the syllable level and 54.7% (16,178) of words are accented. Time-aligned phone boundaries generated by forced alignment are used to identify vowel regions for analysis. There are 48,359 vowels in the corpus and 34.8 of these are accented. To generate time-aligned syllable boundaries, we align the forced-aligned phones with a syllabified lexicon included with the corpus.

The use of BURNC for comparative accent prediction in our three domains is not straightforward, due to anomalies in the corpus. First, the lexicon and forced-alignment output in BURNC use distinct phonetic inventories; to align these, we have employed a *minimum edit distance* procedure where aligning any two vowels incurs zero cost. This guarantees that, at a minimum the vowels will be aligned correctly. Also, the number of syllables per word in the lexicon does not always match the number of vowels in the forced alignment. This leads to 114 syllables containing two forced-aligned vowels, and 8 containing none. Instead of performing *post hoc* correction of the syllabification results, we include all of the automatically identified syllables in the data set. This syllabification approach generates 48,253 syllables, 16,781 (34.8%) bearing accent.

4 Pitch Accent Detection Experiments

We train logistic regression models to detect the presence of pitch accent using acoustic features drawn from each word, syllable and vowel, using Weka (Witten et al., 1999). The features we use included pitch (f0), energy and duration, which have been shown to correlate with pitch accent in English. To model these, we calculate pitch and energy contours for each token using Praat (Boersma, 2001). Duration information is derived using the vowel, syllable or word segmentation described in Section 3. The feature vectors we construct include features derived from both raw and speaker z-score normalized[1] pitch and energy contours. The feature vector used in all three analysis scenarios is comprised of minimum, maximum, mean, standard de-

[1]Z-score normalization:$x_{norm} = \frac{x-\mu}{\sigma}$, where x is a value to normalize, μ and σ are mean and standard deviation. These are estimated from all pitch or intensity values for a speaker.

viation and the z-score of the maximum of these raw and normalized acoustic contours. The duration of the region in seconds is also included.

The results of ten-fold cross validation classification experiments are shown in Table 1. Note that, when running ten-fold cross validation on syllables and vowels, we divide the folds by words, so that each syllable within a word is a member of the same fold. To allow for direct comparison of the three approaches, we generate word-based results from vowel- and syllable-based experiments. If any syllable or vowel in a word is hypothesized as accented, the containing word is predicted to be accented. Vowel/syllable accuracies should be higher

Region	Accuracy (%)	F-Measure
Vowel	68.5 ± 0.319	0.651 ± 0.00329
Syllable	75.6 ± 0.125	0.756 ± 0.00188
Word	82.9 ± 0.168	0.845 ± 0.00162

Table 1: *Word-level accuracy and F-Measure*

than word-based accuracies since the baseline is significantly higher. However, we find that the F-measure for detecting accent is consistently higher for word-based results. A prediction of **accented** on any component syllable is sufficient to generate a correct word prediction.

Our results suggest, first of all, that there is discriminative information beyond the syllable nucleus. Syllable-based classification is significantly better than vowel-based classification, whether we compare accuracy or F-measure. It is possible that the narrow region of analysis offered by syllable and vowel-based analysis makes the aggregated features more susceptible to the effects of noise. Moreover, errors in the forced-alignment phone boundaries and syllabification may negatively impact the performance of vowel- and syllable-based approaches. Until automatic phone alignment improves, word-based prediction appears to be more reliable. An automatic, acoustic syllable-nucleus detection approach may be able generate more discriminative regions of analysis for pitch accent detection than the forced-alignment and lexicon alignment technique used here. This remains an area for future study.

However, if we accept that the feature representations accurately model the acoustic information contained in the regions of analysis and that the BURNC annotation is accurate, the most likely explanation for the superiority of word-based prediction over syllable- or vowel-based strategiesis is that the acoustic excursions correlated with accent occur outside a word's lexically stressed syllable. In particular, complex pitch accents in English are generally realized on multiple syllables. To examine this possibility, we looked at the distribution of misses from the three classification scenarios. The distribution of pitch accent types of missed detections using evaluation of the three scenarios is shown in Table 2. In the ToBI framework, the complex pitch accents include L+H*, L*+H, H+!H* and their downstepped variants. As we suspected, larger units of analysis lead to improved performance on complex tones; χ^2 analysis of the difference between the error distributions yields a χ^2 of 42.108, p< 0.0001.

Since accenting is the perception of a word as more prominent than surrounding words, features that incorporate local contextual acoustic information should improve detection accuracy at all levels. To represent surrounding acoustic context in feature vectors, we calculate the z-score of the maximum and mean pitch and energy over six regions. Three of these are "short range" regions: one previous region, one following region, and both the previous and following region. The other three are "long range" regions. For words, these regions are defined as two previous words, two following words, and both two previous and two following words. To give syllable- and vowel-based classification scenarios access to a comparable amount of acoustic context, the "long range" regions covered ranges of three syllables or vowels. There are approximately 1.63 syllables/vowels per word in the BURNC corpus; thus, on balance, a window of two words is equivalent to one of three syllables. Duration is also normalized relative to the duration of regions within the contextual regions. Accuracy and f-measure results from ten-fold cross validation experiments are shown in Table 3. We find dramatic

Analysis Region	Accuracy (%)	F-Measure
Vowel	77.4 ± 0.264	0.774 ± 0.00370
Syllable	81.9 ± 0.197	0.829 ± 0.00195
Word	84.2 ± 0.247	0.858 ± 0.00276

Table 3: *Word-level accuracy and F-Measure with Contextual Features*

increases in the performance of vowel- and syllable-

Region	H*	L*	Complex	Total Misses
Vowel	.6825 (3732)	.0686 (375)	.2489 (1361)	1.0 (5468)
Syllable	.7033 (2422)	.0851 (293)	.2117 (729)	1.0 (3444)
Word	.7422 (2002)	.0610 (165)	.1986 (537)	1.0 (2704)

Table 2: *Distribution of missed detections organized by H*, L* and complex pitch accents.*

based performance when we include contextual features. Vowel-based classification shows nearly 10% absolute increase accuracy when translated to the word level. The improvements in word-based classification, however, are less dramatic. It may be that word-based analysis already incorporates much the contextual information that is helpful for detecting pitch accents. The feature representations in each of these three experiments include a comparable amount of acoustic context. This suggests that the superiority of word-based detection is not simply due to the access to more contextual information, but rather that there is discriminative information outside the accent-bearing syllable.

5 Conclusion and Future Work

In this paper, we describe experiments comparing the detection of pitch accents on three acoustic domains – words, syllables and vowels – using acoustic features alone. To permit direct comparison between accent prediction in these three domains of analysis, we generate word-, syllable-, and vowel-based results directly, and then transfer syllable- and nucleus-based predictions to word predictions.

Our experiments show that word-based accent detection significantly outperforms syllable- and vowel-based approaches. Extracting features that incorporate acoustic information from surrounding context improves performance in all three domains. We find that there is, in fact, acoustic information discriminative to pitch accent that is found within accented words, outside the accent-bearing syllable. We achieve 84.2% word-based accuracy — significantly below the 86.0% reported by Sridhar *et al.* (2008) using syntactic and acoustic components. However, our experiments use only acoustic features, since we are concerned with comparing domains of acoustic analysis within the larger task of accent identification. Our 84.2% accuracy *is* significantly higher than the 80.09% accuracy obtained by the 10ms frame-based acoustic modeling described in (Sridhar et al., 2008). Our aggregations of pitch and energy contours over a region of analysis appear to be more helpful than short frame modeling.

In future work, we will explore a number of techniques to transfer word based predictions to syllables. This will allow us to compare word-based detection to published syllable-based results. Preliminary results suggest that word-based detection is superior regardless of the domain of evaluation.

References

S. Ananthakrishnan and S. Narayanan. 2008. Automatic prosodic event detection using acoustic, lexical and syntactic evidence. *IEEE Transactions on Audio, Speech & Language Processing*, 16(1):216–228.

D. Baron, E. Shriberg, and A. Stolcke. 2002. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In *ICSLP*.

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

M. Ostendorf, P. Price, and S. Shattuck-Hufnagel. 1995. The boston university radio news corpus. Technical Report ECS-95-001, Boston University, March.

A. Rosenberg and J. Hirschberg. 2007. Detecting pitch accent using pitch-corrected energy-based predictors. In *Interspeech*.

K. Ross and M. Ostendorf. 1996. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech & Language*, 10(3):155–185.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. Tobi: A standard for labeling english prosody. In *Proc. of the 1992 International Conference on Spoken Language Processing*, volume 2, pages 12–16.

V. R. Sridhar, S. Bangalore, and S. Narayanan. 2008. Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. *IEEE Transactions on Audio, Speech & Language Processing*, 16(4):797–811.

F. Tamburini. 2003. Prosodic prominence detection in speech. In *ISSPA2003*, pages 385–388.

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