

Using Mutual Information to Identify Regions of Analysis for Prosodic Analysis

Andrew Rosenberg

Department of Computer Science, Queens College (CUNY), USA

andrew@cs.qc.cuny.edu

Abstract

This paper presents a novel technique for empirically identifying regions of analysis for time/value information. The technique relies on analysis of mutual information between the contour, and some variable of interest. We present the use of this technique in the analysis of prosody in American English speech, where we identify valuable regions of analysis for the classification of phrase ending intonation. We also use the technique to investigate the most informative region of analysis for pitch accent detection.

Index Terms: Prosodic Analysis, Mutual Information, Region of Analysis

1. Introduction

Identification of a region of analysis is a necessary first step in the automatic analysis of prosody. It is common to extract acoustic information from a region of speech that is aligned with some lexical boundary, either a word or syllable. There are compelling arguments to generate predictions that are aligned with both the syllable and word level. When predicting prominence, words are *accented* to convey information such as contrast, focus, topicality or information status. The communicative implications of prominence influence the interpretation of the word or phrase. However, the acoustic excursions associated with accenting 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 regions of analysis.

A similar argument follows with regard to phrase ending intonation. The intonation preceding the end of a phrase influences the intended interpretation of the full phrase. The difference between declarative statements and questions like “You like John.” and “You like John?” is indicated by phrase ending intonation, specifically a high-rise in the question and falling intonation in the statement. While the communicative impact of phrase ending prosodic variation occurs at the phrase level, the acoustic realization of this variation occurs immediately preceding the phrase ending. There are few published results that analyze a full phrase to predict phrase ending intonation [1]. However, both the phrase final word [2] and phrase final syllable [14] have been explored as regions of analysis.

There is no reason that a region of acoustic analysis needs to be tied to the word or syllable level, even if the generated predictions will ultimately be aligned to these segmental regions. By separating the identification of an appropriate region of analysis from the segmentation decision, classification performance can be improved. In this paper, we describe an approach to identifying regions of analysis based on mutual information (MI). We investigate the MI between an acoustic contour and a variable

of interest at each time slice. We then threshold this value to isolate the region showing the greatest mutual information.

Throughout this paper we discuss the identification of a region of analysis in the context of prosodic analysis – classifying phrase ending intonation and detecting pitch accents. However, the approach we propose can be applied to other more task-based spoken language processing applications. For example, there is evidence that dialog participants can identify turn-taking cues prior to the cessation of speech [3]. Topic boundaries, and discourse boundaries are signaled, in part, by changes in pitch, duration and speaking rate across these boundaries [4]. This approach can be applied to the identification of appropriate regions preceding and following candidate boundaries to discriminating information for topic or discourse segmentation.

The remainder of this paper is structured as follows. The approach itself is described in Section 2. We apply this approach to two prosodic analysis tasks: classification of phrase ending intonation in Section 3.1 and detecting pitch accents at the word or syllable level 3.2. In Section 4 we describe related work. We conclude and describe future directions for this research in Section 5.

2. Normalized Mutual Information

We address the problem of identifying an optimal region of analysis using Mutual Information (MI) to measure the correlation between acoustic-prosodic contours and prosodic labels. Mutual information is an information theoretic measure that describes the mutual dependence of two random variables as measured in bits. This can be interpreted as how much information the two variables share.

We define a *contour*, $x = [x_1, \dots, x_T]$, as a vector of values extracted from a speech signal at evenly spaced intervals where T is number of intervals in the speech signal. At every time slice t , we calculate the mutual information between a set of contours $X = \{x^1, \dots, x^N\}$ and a prosodic variable c . Equation 2 describes the calculation of Mutual Information between the contours X and a prosody variable c at time t .

$$I(X; c, t) = \sum_{x^i \in X} \sum_{c \in C} p(x^i, c; \theta_t, \phi_t) \log \frac{p(x^i, c; \theta_t, \phi_t)}{p(x^i; \theta_t)p(c; \phi_t)}$$

We model the distribution of contour values x_t at each time t using a gaussian distribution with parameters θ_t . Since contours have different lengths, the distribution of c at each time slice t can vary. For example, prominent words are typically longer than non-prominent words. We model this as a multinomial distribution with parameters, ϕ_t . In the experiments reported in this paper, we will be using categorical prosodic labels to calculate mutual information. The inventory of labels, C , is defined by the ToBI standard [5]. However, the calculation of MI can be applied to continuous variables, by integrating rather

than summing over the C distribution. To apply this approach to continuous representations of prosody the calculation is modified to represent a continuous prosody variable, k where ψ_t are parameters of a distribution of c at time t .

$$I(X; k, t) = \sum_{x^i \in X} \int_k p(x^i, k; \theta_t, \psi_t) \log \frac{p(x^i, k; \theta_t, \psi_t)}{p(x^i; \theta_t)p(k; \psi_t)} dk$$

Mutual information, $I(X; c, t)$ has a minimal value of zero when two variables X and c are completely independent, and reaches a maximum of $\frac{\min(H(X_t), H(c))}{H(X_t) + H(c)}$ when one variable is completely redundant with knowledge of the other where $X_t = \{x^i\}$ for all $x^i \in X$. In order to compare mutual information measures across distributions with different values of $H(X_t)$ and $H(c)$, we normalize the mutual information:

$$D'(X, c, t) = \frac{I(X; c, t)}{\max(H(X_t), H(c))}$$

$D'(X, c, t)$ has a range between 0 and 1, where 0 indicates that the X and c are independent at time t and 1 if they are redundant at time t . $1 - D'(X, c, t)$ is a universal metric describing the independence of two distributions [6].

3. Experiments

In Sections 3.1 and 3.2, we describe experiments using mutual information to identify a region of analysis to classify phrase ending intonation or to detect prominence. Both sets of experiments use speech from one speaker, f2b, from the Boston University Radio News Corpus (BURNC) [7] as material. Recordings of this speaker total 78.3 minutes. Pitch and intensity contours are extracted using AuToBI [8] implementations of Praat's [9] To Pitch (ac)... and To Intensity... functions. Contours are made up of pitch values in log Hz, and intensity in d; both are extracted at 10ms intervals.

3.1. Analysis of Phrase Ending Intonation

In the ToBI system, phrasing is described by a hierarchy at two levels – the intermediate and intonational phrase. An intermediate phrase contains one or more words, and an intonational phrase contains one or more intermediate phrases. The phrase ending intonation at intermediate phrase boundaries is described by a *phrase accent* which can take either high (H-), downstepped high (!H-) or low tones (L-). For the purposes of these experiments we collapse the high and downstepped high tones into a single tone (H-). The local acoustic realizations of these two tones are very similar – their difference is based on the previous operating pitch range. Each intonational phrase boundary has a boundary tone which can also take a high (H%) or low (L%) tone. Since each intonational phrase boundary is also an intermediate phrase boundary, intonational phrase ending intonation is described by one of four phrase accent boundary tone pairs, L-L%, L-H%, H-L%, and H-H%. We treat these five tone pairs (!H-H% is not included in the ToBI standard) as the inventory for intonational phrase ending intonation because 1) boundary tones never occur in isolation from phrase accents and 2) determining where a phrase accent ends and a boundary tone starts is a difficult at best..

3.1.1. Identifying a region of analysis

We examine the $D'(X, c, t)$ for all time slices t for both pitch (log Hz) and intensity (dB) contours in order to identify the best region of analysis to discriminate between intonational phrase ending types. Since we are looking for a region aligned with the end of an intonational phrase, we align each contour at highest index to calculate the mutual information measure.

To identify a region of analysis for the classification of intonational phrase ending intonation, we compare the MI of pitch and intensity contours to the phrase accent-boundary tone pairs produced. We use every intonational phrase boundary produced by speaker f2b in this analysis. Plots of the mutual information for pitch and intensity can be found in Figure 1. To more

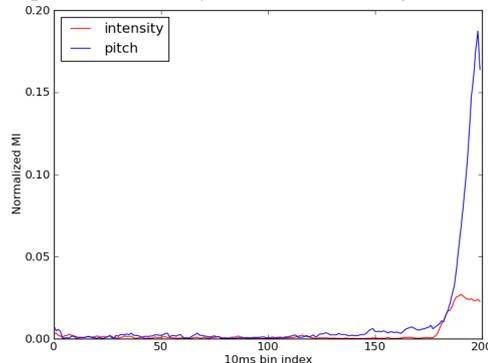


Figure 1: *Intonational Phrase Ending mutual information $D'(X, c, t)$ for each time slice t .*

clearly understand the derivation of this plot, Figure 2 includes the mean and standard deviation of pitch for each time slice preceding an intonational phrase boundary. This information is calculated over every intonational phrase produced by speaker f2b. The contours themselves, on average, follow typical ToBI

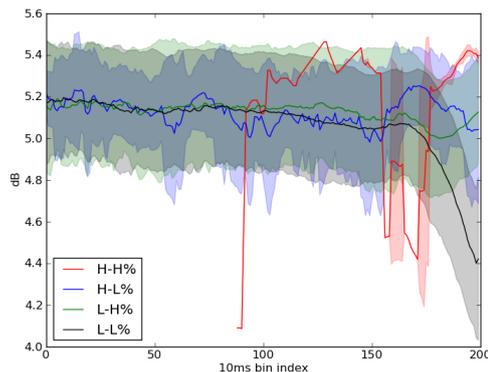


Figure 2: *Mean pitch contours for intonational phrases based on phrase ending tone, one standard deviation around the mean is shaded.*

behavior, with L-L% phrases showing a steep fall, L-H% a shallow rise, H-H% a high rise, and H-L% a high sustained plateau. The erratic behavior of the H-H% contour is due to the fact that the mean values are derived from only 4 contours produced by speaker f2b. We can observe that the the pitch contour of the phrases tend to diverge based on the phrase ending tone starting approximately 30 frames before the end, between slice 160 and 170. When we compare this to the mutual information plot, we see that this is the time index at which pitch contour and phrase ending intonation begin to show increased mutual information. When examining the mutual information with regards to intensity, we see significantly less redundancy with phrase ending tone. However, the value is greater than zero in approximately the same region as pitch – starting at bin 110.

In an idealized setting, the mutual information between a contour and a related variable would be zero in all places other than the appropriate region of analysis. Due to noise in the data, poor probability estimates and floating point precision noise, this is rarely the case. Therefore we establish a threshold, ϵ below which we define zero mutual information. We identify ϵ

such that only a single continuous region of analysis containing the highest value is defined. For some tasks this threshold identification process may be more appropriate than others. We identify the region of analysis for classification of phrase ending tone by thresholding the $D'(X, c, t)$ value at $\epsilon = 0.01$. In this data, this identifies a 190ms region from $t = 113$ to 132 as the region of analysis.

We repeat this analysis for intermediate phrase endings that are not intonational phrase endings. These points of modest disjuncture have phrase accents, but no boundary tones. The MI between pitch and intensity contours and phrase accent label can be found in Figure 3. For intermediate phrases, we identify

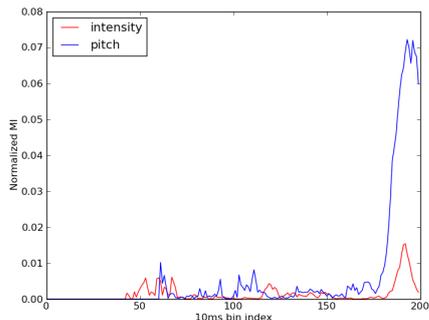


Figure 3: *Intermediate Phrase Ending mutual information $D'(X, c, t)$ for each time slice t .*

a 290ms region of analysis preceding phrase endings, using $\epsilon = 0.007$ to threshold $D'(X, c, t)$ values.

We now evaluate the value of this technique for the automatic classification of phrase ending intonation. We compare the empirically defined region of analysis to regions of analysis defined by the phrase-ending word, or syllable. We train classifiers on f2b material, and evaluate on the remaining 5 BURNC speakers. In this approach, we construct aggregate acoustic features extracted from within only the region of analysis. Based on these aggregate feature vectors, we use AuToBI to train a support vector machine (SVM) with linear kernel and sequential minimum optimization to classify the phrase ending tones. AuToBI [8] is an open-source toolkit for prosodic analysis. We use the weka API [10] through AuToBI to train and evaluate all models. The features that we extract are minimum, maximum, mean, and standard deviation of pitch (log Hz) and intensity (dB). We also extract these four features from the slopes of the contours. To allow the models trained on f2b material to be applied to material from other speakers, we perform z-score normalization on the pitch and intensity contours based on mean and standard deviation values calculated over all speech from a given speaker. The results of this experiment can be found in Table 1. The baseline is a majority class classification, H- for intermediate phrases, and L-L% for intonational phrases. On these

Region	Intermediate	Intonational
Baseline	68.92%	50.16%
Word	69.63%	66.30%
Syllable	63.33%	69.56%
MI	69.18%	69.95%

Table 1: *Classification accuracy of phrase ending intonation for intermediate and intonational phrases.*

two classification tasks, the performance of using word-based or syllable based regions of analysis is inconsistent. The syllable ROA generates better performance on the intonational phrase ending classification task by 6.23%. However, the word ROA yields phrase accent classification results that are 3.33% better than those obtained by extracting features from phrase-ending

syllables. By empirically defining the region of analysis, we are able to generate phrase ending classification results that are not significantly different from the best performance obtained by using either syllable or word ROAs; 0.45% lower than word based classification on phrase accent classification, and 0.39% higher than syllable based on intonational phrase endings. This shows that this technique generates more robust performance than either lexically-based region of analysis. The similarity in performance between the MI based ROA and the two lexical regions is not particularly surprising. In the BURNC, the mean f2b syllable length is 220ms, while the MI region for intonational phrase boundary classification is 190ms, a difference of 3 frames. The average word f2b word duration is 334ms compared to the MI region of 290ms, a difference of 4 frames.

Extracting features from suboptimal regions of analysis will lead to suboptimal classifier performance. The limitation of omitting discriminative regions from the analysis window is clear – this discriminative information will move the aggregate values further apart and lead to greater correlation with the label. On the other hand, when information from uninformative regions is included in an aggregate feature the feature is *less* discriminative. If the feature is represented by a normal distribution over contour values, inclusion of points from uninformative regions has the impact of either moving the class-conditional mean closer to the population mean or increasing the variance. In either case, the discriminative power of the feature is reduced. We can observe both of these effects in the experiment results. When classifying intermediate phrase endings using a syllable based region of analysis, the ROA is too short to capture all the available discriminative information. When classifying intonational phrase endings, the final word ROA includes too much non-discriminative information leading to reduced accuracy.

3.2. Analysis of Prominence in two-syllable words

In this section, we use the region of analysis identification technique to help describe why pitch accent detection at the word level typically generates higher accuracy than syllable-based detection. Pitch accents in English are approximately aligned with the lexically stressed syllable of the prominent word. In order to identify the appropriate region of analysis within the lexically stressed syllable we compared the mutual information between a binary accent variable, and pitch and intensity contours. If the conventional wisdom holds, we expect to find greater than zero $D'(X, c, t)$ values within the lexically stressed syllable, and very small values elsewhere. To test this, we examine only two-syllable words from the BURNC corpus spoken by f2b, and to further constrain this analysis we contrast those words with lexical stress on the first syllable to those with lexical stress on the second syllable. Figure 4 shows $D'(X, c, t)$ values for the first and second syllable in two-syllable words. In these plots, we isolate lexically stressed and unstressed syllables in each position. From these figure, we can clearly see that the discriminative information to accenting in two-syllable words is not confined to the lexically stressed syllable, particularly in the case where the first syllable bears lexical stress. While there is *more* discriminative information in the lexically stressed syllable, we find frames with discriminative information in *both* syllables. This observation helps explain previously reported results that indicate that pitch accent prediction at the word level yields higher accuracy than prediction at the syllable level [11]. The evidence is less convincing in the case of two-syllable words where the second syllable is stressed. The MI between both pitch and intensity and the presence of pitch accent is quite low in the first, unstressed syllable. We note,

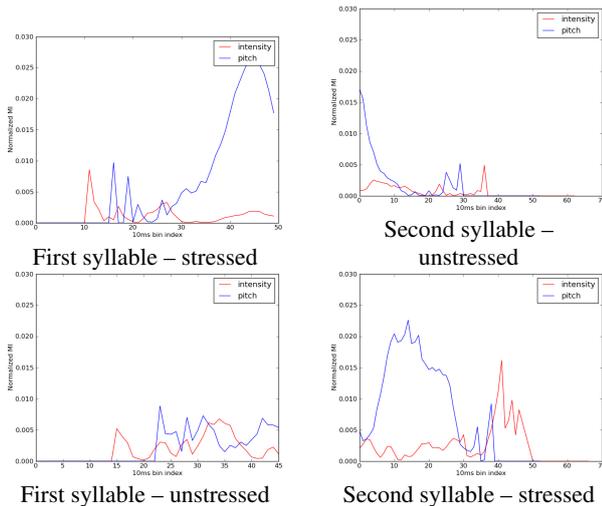


Figure 4: $D'(X, c, t)$ for pitch and intensity contours and a binary accenting variable in two syllable words where the second is lexically stressed.

however, that pitch accent detection is significantly aided by the inclusion of broader acoustic context [12, 11]. This region of analysis identification technique relies on only local observations without any notion of context.

4. Related work

It is common for prosodic analysis hypotheses to be generated at the word or syllable level. Typically, this decision also dictates the region of analysis for feature extraction. That is, when hypothesizing whether a syllable is accented, acoustic information is drawn from the region defined by the syllable boundaries. When detecting pitch accents, many researchers have explored extracting features from syllables (e.g. [13, 14]) and from words ([15, 2]). Levow also explored the use of context by constructing features using a region of analysis surrounding the syllable being classified as prominent or not [12]. The question of word versus syllable or vowel based prediction of pitch accents was directly explored by Rosenberg and Hirschberg [11]. Region of analysis identification also come into play in applications of what Shriberg and Stolcke [16] called “direct modeling of prosody”, where prosodic features are used in modeling a phenomena like turn-taking or topic segmentation without generating an intermediate representation of prosodic markers. Gravano and Hirschberg [17] investigated prosodic correlates to turn yielding over turn-final interpausal units. Levow extracted mean pitch from utterance final words and syllables when investigating turn-taking behavior in Mandarin [18]. In an effort to detect question-bearing turns in tutoring dialogs, Liscombe et al. [1] extracted prosodic features from both the entire turn and the final 200ms of the turn. The authors found that features drawn from the final 200ms of a turn were some of the most discriminative for the task. It is commonly held that questioning behavior is indicated through phrase ending intonation. This closely mirrors the result we observe in Section 3.1.1: the region 190ms preceding a phrase boundary is best suited for the classification of phrase-ending tones.

5. Conclusion and Future Work

Identification of an appropriate region of analysis is a necessary first step in the automatic analysis of prosody. In this paper, we describe a technique to use labeled training material to inform the feature extraction process. By extracting acoustic features only from regions where the mutual information between the

contour and the label is greater than a threshold, we are able to construct robust prosodic features.

This technique can clearly be applied to other prosody modeling tasks, including pitch accent classification, and phrase-boundary detection. It also has application in “direct-modeling” spoken language processing applications like turn-taking, question detection, and topic segmentation. It is clear that there are prosodic cues to this behavior, but the locus of the prosodic variation is not clearly defined. Throughout this paper, we have applied this technique to prosodic contours for spoken language processing. However, this could be applied to any time-value contour. Stock pricing is represented by such a contour, as is much biomedical information including blood-pressure and blood-glucose levels. Moreover, there is a growing interest in the prediction of seizures using EEG data. This technique could be applied to identify what region contains discriminative information. Regardless of the domain, moving towards a more empirical technique to identify regions of analysis will likely lead to improved and robust analyses.

6. References

- [1] J. Liscombe, J. Venditti, and J. Hirschberg, “Detecting question turns in spoken tutorial dialogues,” in *Interspeech*, 2006.
- [2] S. Ananthakrishnan and S. Narayanan, “Fine-grained pitch accent and boundary tone labeling with parametric f0 features,” in *ICASSP*, 2008.
- [3] C. T. Ishi, H. Ishiguro, and N. Hagita, “Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts,” in *ICSLP*, 2006.
- [4] A. Rosenberg and J. Hirschberg, “Varying input segmentation for story boundary detection in english, arabic and mandarin broadcast news,” in *Interspeech*, 2007.
- [5] K. Silverman, et al., “Tobi: A standard for labeling english prosody,” in *Proc. of the 1992 ICSLP*, vol. 2, 1992, pp. 12–16.
- [6] A. Kraskov and P. Grassberger, “Mic: Mutual information based hierarchical clustering,” in *Information Theory and Statistical Learning*, F. Emmert-Streib and M. Dehmer, Eds. Springer US, 2009, pp. 101–123.
- [7] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, “The boston university radio news corpus,” Boston University, Tech. Rep. ECS-95-001, March 1995.
- [8] A. Rosenberg, “Autobi – a tool for automatic tobi annotation,” in *Interspeech*, 2010.
- [9] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glott International*, vol. 5, no. 9-10, pp. 341–345, 2001.
- [10] I. Witten, et al., “Weka: Practical machine learning tools and techniques with java implementation,” in *ICONIP/ANZIIS/ANNES*, 1999, pp. 192–196.
- [11] A. Rosenberg and J. Hirschberg, “Detecting pitch accents at the word, syllable and vowel level,” in *HLT-NAACL*, 2009.
- [12] G.-A. Levow, “Context in multi-lingual tone and pitch accent recognition,” in *Interspeech*, 2005.
- [13] C. Wightman and M. Ostendorf, “Automatic labeling of prosodic patterns,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, 1994.
- [14] X. Sun, “Pitch accent predicting using ensemble machine learning,” in *ICSLP*, 2002.
- [15] A. Rosenberg and J. Hirschberg, “Detecting pitch accent using pitch-corrected energy-based predictors,” in *Interspeech*, 2007.
- [16] E. Shriberg and A. Stolcke, “Direct modeling of prosody: an overview of applications in automatic speech processing,” in *Speech Prosody*, 2004.
- [17] A. Gravano and J. Hirschberg, “Turn-yielding cues in task-oriented dialogue,” in *SigDial*, 2009.
- [18] G.-A. Levow, “Turn-taking in mandarin dialogue: Interactions of tone and intonation,” in *SigHAN*, 2005.