Combining Acoustic, Lexical, and Syntactic Evidence for Automatic Unsupervised Prosody Labeling

Sankaranarayanan Ananthakrishnan, Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory
Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089
ananthak@usc.edu, shri@sipi.usc.edu

Abstract

Automatic labeling of prosodic events in speech has potentially significant implications for spoken language processing applications, and has received much attention over the years, especially after the introduction of annotation standards such as ToBI. Current labeling techniques are based on supervised learning, relying on the availability of a corpus that is annotated with the prosodic labels of interest in order to train the system. However, creating such resources is an expensive and time-consuming task. In this paper, we examine an unsupervised labeling algorithm for accent (prominence) and prosodic phrase boundary detection at the linguistic syllable level, and evaluate their performance on an standard, manually annotated corpus. We obtain labeling accuracies of 77.8% and 88.5% for the accent and boundary labeling tasks, respectively. These figures compare well against previously reported performance levels for supervised labelers.

Index Terms: prosody recognition, accent, stress, prominence, prosodic boundary, spoken language processing.

1. Introduction

Prosody is an all-inclusive term that is commonly used to refer to a large class of supra-segmental phenomena that accompany spoken language. These higher-level cues are influenced by a variety of factors that range from the word sequence and the syntactic structure of the utterance, to the utterance category and the speaker's emotional state. They play a significant role in spoken language understanding; as humans, we use them for determining the dialogue context, the speaker's emotional state and intent, and at a lower level, for word and structure disambiguation.

Prosodic phenomena manifest themselves in speech in different ways, including changes in relative intensity (energy) to impart emphasis, or stress, to specific words or syllables; variations of the fundamental frequency (F0) range and contour; and subtle timing variations, such as syllable lengthening, insertion of pauses, etc. These cues are likely to be very useful for spoken language understanding systems; however, it is often difficult to precisely understand and exploit their relationship to the segmental, lexical, syntactic and semantic structure of the utterance from just their acoustic correlates. As a result, the past couple of decades have seen the evolution of categorical annotation schemes for prosodic events, the best-known among which is the Tones and Break Indices (ToBI) standard [1]. The most important prosodic phenomena captured within this framework include pitch accents and prosodic phrase boundaries. Within the ToBI framework, pitch accents roughly correspond with syllable prominence, or "stress", whereas phrase boundaries serve to separate linguistically related chunks of speech and are loosely related to punctuation and syntactic phrase boundaries.

Corpora annotated with ToBI-like labels can be very useful for learning the relationship between prosodic events and the lexical, syntactic and semantic structure of the utterance. However, creating such corpora manually is expensive and time-consuming. Hence, they are severely restricted in size and scope; to date, only one standard publicly available corpus exists that has been labeled with ToBI events - the Boston University Radio News Corpus, which contains about 3 hours of broadcast news-style speech.

Automatic labeling of prosodic events is, therefore, an attractive alternative that has received attention over the past decade. Initial attempts using acoustic evidence alone have been reported in [2]; subsequent efforts harnessing acoustic, lexical and syntactic cues have been reported in [3], [4] and [5]. However, the major drawback with these methods is that they are all based on supervised learning, and require a hand-labeled training corpus. This has the effect of tying the labeling system to the specific corpus with which it was trained. An unsupervised word prominence labeling algorithm is discussed in [6], but this method assigns prominence scores on a continuous scale using only acoustic features.

In this paper, we develop an automatic, unsupervised prosody labeling algorithm that annotates accent and boundary events in speech without the need for a hand-labeled training set. The system is therefore not tied to any single data set or domain. Our algorithm is grounded in modifications of simple clustering techniques to make use of acoustic, lexical and syntactic cues for prosody labeling. The remainder of this paper is organized as follows. Section 2 describes the data set and features we use for labeling and evaluation. Section 3 gives an overview of our unsupervised labeling technique. In Section 4, we present our experimental results, and in Section 5, we present a brief discussion of the results and suggest future directions to explore.

2. Data corpus and features

The Boston University Radio News Corpus (BU-RNC) is a broadcast news-style read speech corpus containing speech from 7 speakers (3 females, 4 males), totaling about 3 hours of acoustic data. A significant fraction of this data has been manually anno-

tated with ToBI labels. In addition to the prosody labels, it also contains the orthography corresponding to each spoken utterance, ASR-generated phone-level alignments, and part-of-speech (POS) annotation for each word in the orthography. Since our intention is to explore unsupervised labeling techniques, our only motivation for choosing to work with this corpus is to allow an easy, concrete evaluation of our algorithm.

We collapse all categories of ToBI pitch accents into a single accent label, and similarly all categories of prosodic phrase boundaries into a single boundary label, reducing the annotation task to two independent binary classification problems. All prosody labels are assigned at the linguistic syllable level. The following subsections describe the acoustic, lexical and syntactic features we use for the unsupervised labeling task.

2.1. Acoustic features

The prosodic features we extract from the acoustic data and auxiliary sources essentially capture information about the intensity, intonation and timing effects in the speech. Since we assign prosodic labels to individual syllables, our acoustic feature vectors are aligned at the linguistic syllable level. Syllabification of the transcripts corresponding to the speech data is carried out using a deterministic algorithm based on the phonological rules of English [7]. We extract the following acoustic features from the data corpus:

- *Intensity* - within-syllable energy range, difference between minimum and average within-syllable energy
- *F0-related* - within-syllable F0 range, difference between minimum and average within-syllable F0
- *Timing* - syllable nucleus duration, pause duration (for boundary labeling only)

Previous work in this area suggests that these features are likely to be useful for labeling prosodic events. We extract features independently for every syllable and do not consider dependencies across syllable boundaries.

2.2. Lexical and syntactic features

The lexical representation of an utterance as well as its syntactic structure play an important role in determining its prosody. A previous study [3] shows that certain syllable tokens, such as *k_aa_n*, which occur mostly in content words, are much more likely to be associated with accents than tokens such as *dh_ax*, which occur mostly in function words. Similarly, nouns are much more likely to coincide with phrase boundaries than adjectives. However, training labels are absent in the unsupervised context, and this relationship cannot be learned directly from the data. We present an incremental technique that allows us to use these cues to refine the clusters generated using the acoustic features alone.

Our lexical features include the syllable tokens that make up the sequence of words, as well as canonical stress patterns from a standard pronunciation dictionary. We use part-of-speech (POS) tags as shallow syntactic features.

3. Unsupervised labeling algorithm

Our unsupervised labeling system is based on widely-used clustering algorithms for metric data. In this paper, we consider both model-free techniques such as k-means and its probabilistic variant, fuzzy k-means, as well as model-based clustering algorithms such as Gaussian mixtures. The basic approach is as follows. We use the clustering algorithm to partition the acoustic feature space into two clusters. One cluster represents the positive labels (accent or boundary) and the other, the negative labels (no accent, or no boundary). This step provides an initial, rough separation of the samples into the two desired categories.

Assuming that we are able to uncover the mapping between cluster and prosody labels, the separation induced by the clustering algorithms is likely to produce output labels of average accuracy. We would like to improve on the results produced by acoustic feature clustering using lexical and syntactic features. In order to do this, we identify the samples that are the most *reliable* representatives of their respective clusters. Using these reliable samples, for which the true classification accuracy is likely to be quite high, we estimate conditional probability distributions over the lexical and syntactic features given the cluster labels. We use this information to reassign cluster labels to less reliable samples. This step is carried out iteratively until no further samples have to be reassigned cluster labels. In essence, given a sample $x_i = (a_i, s_i, l_i, pos_i)$, where the factors represent the acoustic feature vector, syllable token, canonical stress label, and POS tag of the containing word, respectively, we implement a MAP classifier with the cluster labels as target classes.

$$c_i^* = \arg\max_{c \in (c_0, c_1)} p(c|a_i, s_i, l_i, pos_i) \quad (1)$$

$$= \arg\max_{c \in (c_0, c_1)} p(a_i, s_i, l_i, pos_i|c)\, p(c) \quad (2)$$

Given the large joint vocabulary, it is not possible to estimate this joint conditional distribution from the data corpus. Hence, we invoke a naïve-Bayesian approximation, which simplifies the above equation as follows.

$$c_i^* = \arg\max_{c \in (c_0, c_1)} p(a_i|c)\, p(s_i, l_i|c)\, p(pos_i|c)\, p(c) \quad (3)$$

$$= \arg\max_{c \in (c_0, c_1)} p(c|a_i)\, p(s_i, l_i|c)\, p(pos_i|c) \quad (4)$$

The cluster posteriors $p(c|a_i)$ are obtained from the partitioning algorithm; the conditional distributions $p(s_i, l_i|c)$ and $p(pos_i|c)$ are estimated from the reliably clustered samples. The performance of this algorithm is dependent on the quality of the initial partitioning algorithm, and on the choice of a good reliability metric. These, and the iterative method used to reclassify less reliable samples, are discussed in the following subsections.

3.1. Clustering algorithms

We employ widely-used metric data clustering algorithms in order to obtain an initial partitioning of the acoustic feature space.

- `kmeans`: This model-free algorithm produces a "hard" partition of the feature space. Samples belong to their assigned cluster with unit probability and to other clusters with zero probability.
- `fuzzkm`: Fuzzy k-means produces a soft partition of the feature space. Samples may belong to more than one cluster, the "degree of belongingness" being expressed by a *membership function*. This function is usually so chosen that its sum over all clusters is unity, permitting membership values to be treated as cluster posterior probabilities.

- `gmm`: This is a model-based clustering technique that fits a Gaussian-mixture probability distribution over the data. The number of mixtures in the distribution represents the number of clusters. The EM algorithm is used to estimate parameters for this GMM (priors, means and covariances).

In our experiments, we partition the acoustic feature space into two clusters, which represent the positive and negative categories for the accent and boundary labeling tasks.

3.2. Evaluating reliability

We identify the most reliable representatives of each cluster by evaluating a reliability metric for each sample x_i from its acoustic correlates a_i. We expect that the acoustic confusion associated with reliable samples is small, and by extension, that their cluster assignments yield high accuracy when the cluster labels are mapped to the correct prosody labels. We experiment with the following reliability measures.

- `km-euclid`: This reliability metric compares the Euclidean distances between the vector a_i and the two cluster means, and returns the absolute difference between them (Eq. 5). The most reliable samples are those that lie close to one cluster mean, but are distant from the mean of the competing cluster.

$$R_E(x_i) = |d_E(a_i, m_0) - d_E(a_i, m_1)| \tag{5}$$

- `km-mahal`: This is a slight variation of `km-euclid` in that we estimate covariance matrices for each cluster, and use the Mahalanobis distance instead of the Euclidean distance to evaluate sample reliability (Eq. 6).

$$R_M(x_i) = |d_M(a_i, m_0, \Sigma_0) - d_M(a_i, m_1, \Sigma_1)| \tag{6}$$

- `fkm-post`: When properly initialized, the cluster membership function returned by fuzzy k-means for each sample x_i can be treated as the posterior probability $p(c_i|a_i)$ of the cluster given the sample. A straightforward choice for the reliability metric in this case is just the maximum membership value of the sample across the two clusters (Eq. 7).

$$R_P(x_i) = \max_{c \in (c_0, c_1)} p(c|a_i) \tag{7}$$

- `gmm-mahal`: This is a variation of `km-mahal` applied to the `gmm` clustering technique. We use the Gaussian covariance matrices to evaluate the reliability metric.

We evaluate the chosen reliability metric for each data point x_i, and re-rank the samples in descending order of reliability for the next step - bootstrapping conditional probability models for the lexical and/or syntactic features given the cluster labels.

3.3. Bootstrapping lexical and syntactic probability models

The lexical and syntactic conditional probability distributions are estimated from the most reliably clustered samples. We expect that the labeling accuracy for these samples is high enough that the noise in estimating these distributions is low. There is a trade-off in choosing T, the fraction of samples with which we estimate these distributions. If T is small, the corresponding set of samples will be very reliable, but we will only have a small number of samples from which to estimate the probability models. This leads to sparsity issues and increases the estimation error. On the other hand, choosing a larger value of T leads to a larger "training set", but the samples closer to the tail of this set will not be very reliable, leading to noisy estimates. In our experiments, we vary this parameter and observe its effects on labeling performance.

Since we are dealing with potential sparsity issues, smoothing the probability estimates becomes very important, even though we are only dealing with "unigrams". For these initial experiments, we use Lidstone ("add-λ") smoothing. For instance, the term $p(NNS|c_0)$ is estimated as follows.

$$p(NNS|c_0) = \frac{\lambda + \mathcal{C}(NNS, c_0)}{\lambda \cdot V_{POS} + \mathcal{C}(c_0)} \tag{8}$$

where $\mathcal{C}(\cdot)$ represents the count of the argument, and V_{POS} is the number of unique POS tags. The smoothing parameter λ is set to a suitable small value (say $\lambda = 0.01$). These estimates are used to compute cluster posteriors for each sample by substituting into the product term in Eq. 4 and normalizing across the two clusters. Note that we do not use the acoustic cluster posteriors for `kmeans`, which produces a unimodal posterior. The samples are then re-ranked with the cluster posteriors as the reliability metric, and this process is carried out iteratively to the point where the cluster posteriors converge (i.e., cluster reassignments dip below a threshold, say 0.5% of all samples).

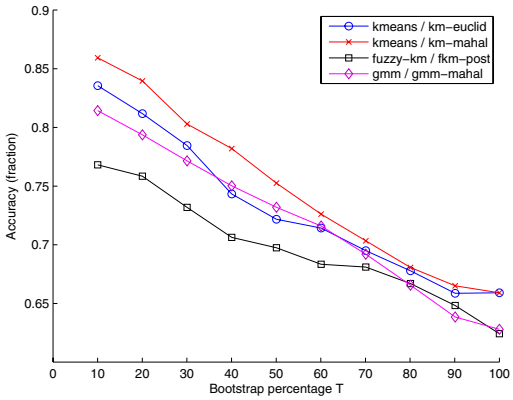
The final step is to map cluster labels to the correct prosody labels. We know from previous work on supervised labeling that positive labels are much less likely than negative labels. For instance, the fraction of syllables in the BU-RNC associated with accent and boundary events is 34% and 17%, respectively. We use this heuristic to map the cluster with fewer samples to the positive label and the competing cluster to the negative label.

4. Experimental results

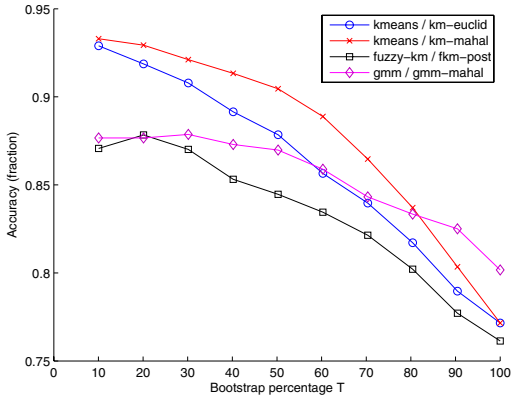
We compare the performance of our algorithm against two methods described in detail in [3]: a heuristic baseline that combines label chance levels with an n-gram model of label sequences, and a supervised labeling technique that uses a combination of neural networks and lexical-syntactic-prosodic language models to uncover prosody labels.

We first cluster the acoustic features using one of the techniques mentioned in Sec. 3, evaluate the reliability metric for each clustered sample, and re-order the samples in decreasing order of reliability. Figures 1(a) and 1(b) show the variation in accent / boundary clustering accuracy as a function of T, the fraction of samples chosen from the top of this list. It is clear that the accuracy is very high for small values of T, but drops off as T increases. The accuracy drop-off is much slower for boundary labeling than it is for accent labeling.

We examined the effects of different values for T (5%, 10%, 15% for accent labeling; 20%, 30%, 40% for boundary labeling) on labeling accuracy, and determined that the algorithm is relatively robust for smaller values of T and only begins to degrade when T is much larger. Tables 1 and 2 present the accent and boundary labeling performance, respectively, for different clustering techniques / reliability measures averaged across the above choices for T. We report labeling accuracy, precision (P), recall (R), and F-score (F) for each method. For the boundary labeling task, we evaluate both overall (all syllables) and word-final (WF) syllable labeling performance; overall accuracy is always higher than WF accuracy, since we map all non-word-final syllables to the negative label.



(a) Accent clustering: accuracy vs. T



(b) Boundary clustering: accuracy vs. T

Figure 1: Variation of accent and boundary cluster-label accuracy as a function of "training" fraction T.

Table 1: Unsupervised accent labeling performance

Method	Acc.	P	R	F
Baseline	67.9%	0.55	0.40	0.46
Supervised	86.4%	0.78	0.85	0.81
kmeans / km-euclid	77.7%	0.62	0.87	0.73
kmeans / km-mahal	77.8%	0.63	0.87	0.73
fuzzkm / fkm-post	77.5%	0.62	0.87	0.73
gmm / gmm-mahal	77.8%	0.63	0.87	0.73

Table 2: Unsupervised boundary labeling performance

Method	All	WF	P	R	F
Baseline	82.8%	72.8%	0.81	0.04	0.08
Supervised	91.6%	87.3%	0.72	0.85	0.78
kmeans / km-euclid	87.0%	78.6%	0.58	0.74	0.65
kmeans / km-mahal	88.5%	81.1%	0.64	0.69	0.66
fuzzkm / fkm-post	86.3%	77.6%	0.57	0.69	0.62
gmm / gmm-mahal	88.1%	80.5%	0.68	0.51	0.58

5. Discussion and future directions

In this paper, we presented an unsupervised prosody labeling algorithm for detecting accent and boundary events in speech. With very few assumptions, our algorithm achieved accent and boundary labeling accuracies of 77.8% and 88.5%, respectively. These figures significantly exceed the baseline performance, and compare well to a supervised labeling system. The performance of this system can be further improved by a clever choice of the reliability metric - the measures used in this paper are obvious, but we would like to explore other possibilities in the future. A limitation of this algorithm is that it operates independently on each sample and cannot incorporate context-based constraints on the label sequence, which have been shown to be important. We would like to modify the algorithm in order to incorporate such constraints.

In the longer term, we plan to use this unsupervised labeling system to annotate different types of data, such as spontaneous speech. This will enable us to explore the usefulness of such labels for spoken language applications such as speech recognition and speech-to-speech translation.

6. References

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard scheme for labeling prosody," in *Proceedings of the International Conference on Spoken Language Processing*, 1992, pp. 867–869.

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

[3] S. Ananthakrishnan and S. Narayanan, "Automatic prosody labeling using acoustic, lexical and syntactic evidence," *submitted to the IEEE Transactions on Speech and Audio Processing*, 2006.

[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 *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2004, pp. 509–512.

[5] S. Ananthakrishnan and S. Narayanan, "An automatic prosody labeling system using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2005, pp. 269–272.

[6] D. Wang and S. Narayanan, "An unsupervised quantitative measure for word prominence in spontaneous speech," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2005, pp. 377–380.

[7] D. Kahn, "Syllable-based generalizations in English phonology," Ph.D. dissertation, University of Massachusetts, 1976.