Discriminative Training and Unsupervised Adaptation for Labeling Prosodic Events with Limited Training Data

Raul Fernandez, Bhuvana Ramabhadran
IBM TJ Watson Research Lab, Yorktown Heights, NY, 10598
{fernra,bhuvana}@us.ibm.com

Abstract
Many applications of spoken-language systems can benefit from having access to annotations of prosodic events. Unfortunately, obtaining human annotations of these events, even sensible amounts to train a supervised system, can become a laborious and costly effort. In this paper we explore applying conditional random fields to automatically label major and minor break indices and pitch accents from a corpus of recorded and transcribed speech using a large set of fully automatically-extracted acoustic and linguistic features. We demonstrate the robustness of these features when used in a discriminative training framework as a function of reducing the amount of training data. We also explore adapting the baseline system in an unsupervised fashion to a target dataset for which no prosodic labels are available, and show how, when operating at point where only limited amounts of data are available, an unsupervised approach can offer up to an additional 3% improvement.

Index Terms: prosody labeling, conditional random fields

1. Introduction
Although a rich literature on the topic of automatic labeling of prosodic events has appeared during the last 15 years, there has been a skewed focus on supervised approaches that rely on pre-annotated labels. Since there is, however, a high cost associated with obtaining human annotation of these events, there is an interest in exploring approaches that dispense with large amounts of training data for supervised learning, that perform robustly when only limited amounts of training data are available, or that can be adapted using unlabeled data. In this paper, we investigate how a discriminative training framework applied to a large, rich feature set reflecting acoustic and linguistic information can perform robustly as a function of diminishing training set size, and how this loss in performance can be partially compensated with using unsupervised adaptation techniques. We apply this approach to automatically determine which words in a corpus bear pitch accent prominence, as well as the break indices between pairs of adjacent words.

2. Observation Features
The input to this procedure is a set of audio files and transcriptions, which are aligned with each other using forced alignment by iteratively re-estimating context-independent models from a flat start, and then further refining the alignments with context-dependent models and decision-tree state clustering. The pronunciation dictionary used for alignment also contains syllabification and stress-pattern information, which is used to post-process the alignment files to contain three tiers of segmentation at the phone-, syllable- and word levels, plus information about the syllable stress. The acoustic waveforms and their corresponding segmentation files are then used to extract different types of features, as explained in the next sections. Following the results of [1] on the effects of analysis scopes at the word, syllable and vowel levels on pitch-accent classification, the word is taken as the unit of analysis in this work.

2.1. F0 Features
Before computing F0 features, acoustic waveforms are first processed using the algorithm in [2] to automatically detect glottal closure instants (GCIs) and extract voicing information. The intra-glottal periods determined by the GCIs in voiced regions are smoothed out to obtain an F0 contour. Linear interpolation is used to fill in the unvoiced discontinuities of the contour, and spline interpolation to obtain a smoother version throughout voiced regions. This last smoothed F0 contour is then used as the input to an algorithm that finds a piecewise linear fit to the entire contour using a two-pass procedure to grow, and then prune, the set of inflection points determining the local linear pieces.

The log smoothed contour over every word is warped to a common length (50 points, in this case) and normalized by the mean and variance over the entire utterance. These 50 coefficients are retained as a descriptor (f1) of the contour shape. The following features are also computed: the mean (f2), range (f3), and difference between mean and minimum (f4) of the z scores of the normalized contour over the word; the number of slope changes of the linear fit over the word (f5), and the value of the last linear segment of the fit (f6); the overall linear trend of the F0 contour over the word (f7), and the second degree coefficient of a 2nd-order polynomial fit over the word to measure the convexity of the contour (f8). Finally, two measures of pitch-perturbation, or jitter, are computed from the GCIs over the last syllable of the word, namely the average of the relative intra-closure gaps (f9), and the average of the relative mean-normalized intra-closure gaps (f10).

2.2. Duration Features
Before extracting duration features, the alignment files are processed to extract statistics (sample mean and variance) of the log durations, both globally and class-dependent, where the classes are determined by (a) phone identity (for every phone in the alignment set), and (b) binary stress (stressed or not). The following 21 features are then extracted: the duration of each word (d1) and a speaking rate factor (d2), obtained by normalizing the speaking rate computed over the word by the speaking rate computed over the entire utterance (speaking rate, in this case, is
measured as number of phones per seconds); the average of the z-scores of log durations over all phones in the word, computed containing this corpus-based prior whenever it is statistically significant. This feature can be summarized as follows:

where \( n_w \) is the number of times a word \( w \) appears in the corpus, \( k_e \) the number of times that it is associated with a prosodic event \( e \), and \( B(k_e, n_w, 0.5) \) is a binomial distribution with parameter \( p = 0.5 \). This feature equals the fraction of “successes” whenever there is sufficient evidence in a corpus to establish how likely a word co-occurs with an event, as diagnosed by a binomial distribution. When not, the feature reflects uncertainty and is set to 0.5. The authors constructed this feature, which they called “Accent Ratio,” and showed a significant contribution when applied to the task of pitch-accent labeling. In this work we use this feature, but generalize it to other prosodic events as well, namely: a word \( w \) bears a pitch accent (E1); a word \( w \) follows a minor boundary (E2); a word \( w \) precedes a minor boundary (E3); a word \( w \) follows a major boundary (E4); a word \( w \) precedes a major boundary (E5). The motivation behind why it might make sense to look at events E2 through E4 is that, at least in the case of fluent speech, different words show different tendency to be coupled or uncoupled from its neighbors. (It is, in fact, a similar motivation behind flagging words as being in one of various classes of functional words, except that in this case the classes are not postulated a priori, and this tendency is derived from a corpus.) Given a training corpus, Eq. 1 can be tabulated for each of the prosodic events defined, and then later looked up for feature assignment. In this case, 5 “dictionaries” of prosodic event ratios were assembled, and used to define linguistic features \( l_{13} \) through \( l_{17} \).

2.5. Nominal Observations

The feature extraction procedure described yields a total of 258 feature types, all but one unidimensional features, plus a contour-shape 50-dimensional feature. As discussed in Section 3, we are working with an implementation of a learning model that takes as its inputs binary-valued indicators signaling when an output label co-occurs with a nominal-valued observation feature. Therefore, before we are in a position to define these indicator functions, we need to transform any real-valued observations extracted from the acoustics and texts into a nominal-valued observation; that is, we need to quantize and index the observation space. Any non-nominal and non-ordinal observation from the previous list was, therefore, transformed into a nominal-valued feature via K-means clustering, where the maximum number of clusters was set to 20 for the F0 contour type feature \( f_j \), and to 5 otherwise (in some cases, K-means found empty clusters, reducing the size of the codebook to 4 or 3 for that particular observation).

3. Conditional Random Fields

A Conditional Random Field (CRF) [6] is an undirected graph with nodes \( x \) and \( y \), corresponding to an observation sequence \( x \) and a latent sequence \( y \) to be inferred, which directly encodes the conditional distribution \( p(y|x) \) using a log-linear model. A linear-chain CRF, in particular, is a structure that assumes that, conditioned on \( x \), the dependencies of \( y \) form a Markov chain. In this paper, we make use of this assumption and restrict ourselves to first-order chains, in which case the conditional distribution can be written as:

\[
p(y|x) = \frac{\exp \left\{ \sum_{k=1}^{K} \lambda_k f_k(y_l, y_{l-1}, x_l) \right\}}{\sum_y \exp \left\{ \sum_{k=1}^{K} \lambda_k f_k(y_l, y_{l-1}, x_l) \right\}},
\]

2.3. Energy Features

The energy features are obtained from a sub-band analysis of 16kHz acoustic waveforms, where the bands are determined by the Bark scale (up to band 21, for this sampling rate). This approach is motivated by the results of [3], who found using committees of classifiers trained on different ranges of the Bark scale useful for pitch-accent classification. A bank of FIR, 512-order filters, maximally flat through the passband, is designed and used to decompose the speech into sub-band components, from each of which a short-time energy (STE) profile is extracted using a 20-msec. moving Hamming window every 10 msecs. For each of the 21 sub-bands, the following statistics are measured from the STE over each word: the minimum (\( e_1 \)), maximum (\( e_22 \)), mean (\( e_{43} \)), root-mean-square (\( e_{64} \)), and standard deviation (\( e_{85} \)). Additionally, the above local features are also computed in a normalized version, by dividing by the value estimated over a 5-word window centered on the analysis word (for each corresponding statistic and sub-band combination). This yields features \( e_{106} \) through \( e_{210} \) for a total of 210 energy features.

2.4. Linguistic Features

The linguistic features extracted include some broad class lexical information which has been previously explored, and which intuitively reflect the tendency of certain types of lexical items and collocations to be acceptable and/or dominate before particular types of boundaries (the latter, at least in the case of non-disfluent speech). These include binary flags indicating whether the word is in one of several lists containing function words \( l_1 \), auxiliary verbs \( l_2 \), adpositions \( l_3 \), conjunctions \( l_4 \) and WH-words \( l_5 \). Additionally, orthographic and positional features are extracted reflecting the type of punctuation that follows \( l_6 \), the number of words from the start \( l_7 \) and to the end \( l_8 \) of the utterance, whether the word starts in upper-case \( l_9 \), and whether the word ends a sentence \( l_{10} \). The degree of coupling between words is measured using bigram forward and reverse language models to evaluate, respectively, \( p(w_i|w_{i-1}) \) \( l_{11} \) and \( p(w_i|w_{i+1}) \) \( l_{12} \); features already explored and used for pitch accent detection in the work of [4]. Recently, [5] explored a memory-based feature that reflects prior knowledge about how likely a word is to be associated with a particular prosodic phenomenon, by collecting statistics from a labeled corpus and retaining this corpus-based prior whenever it is statistically significant. This feature can be summarized as follows:

\[
\text{Ratio}(w, w) = \begin{cases} \frac{n_{w}}{0.5} & \text{if } B(k_e, n_w; 0.5) \leq 0.05 \\ 0.5 & \text{otherwise} \end{cases}
\]
where each \( f(y_t, y_{t-1}, x_t) \) is a feature or indicator function\(^1\) associated with the token at time \( t \):
\[
f(y_t, y_{t-1}, x_t) = 1_{y_t=1}1_{y_{t-1}=1}1_{x_t=0}.
\]
We also assume in this formulation that both the latent and observation variables are discrete.

The strength of the model in Eq. 2 is that it directly encodes the posterior distribution over classes of interest given some observations (the quantity we need for classification), and avoids modeling the distribution over the variables \( x \). Avoiding modeling the joint distribution of input and output variables affords us the flexibility to define a rich set of overlapping features whose dependencies may be complex or unknown. The feature functions in Eq. 3 could instead be designed to capture domain knowledge about the problem without minding the interaction between them. For instance, a binary feature could be designed to be 1 whenever a word is pitch-accented and we observe a particular configuration of acoustic and linguistic observables in the input data. Given a training set of \( N \) sequences, the parameters \( \{ \lambda_k \}_{k=1}^{K} \) of the model can be carried out by optimizing the conditional log-likelihood of Eq. 2, plus, optionally, a regularizer on the parameter vector to encourage sparsity and prevent overfitting. This log-likelihood function is concave, and therefore has a single optimal point which can be found via, e.g., gradient ascent. Evaluating the gradients requires an inference step to compute the expectation of the \( f_k \) under the model, a step that can be efficiently computed with a forward-backward algorithm (see [6] for derivation and further details). Once the model has been trained, an input sequence can be classified as the state sequence \( x^* \) maximizing Eq. 2, a step that can be computed via dynamic programming.

3.1. Defining Input Features

As mentioned in the previous section, the power of CRFs arises from their ability to use rich feature representations. In this section we define the feature set we constructed from the acoustic and linguistic observables extracted from data, as described in section 2. The inputs to this stage are the nominal features obtained following K-means clustering and assignment of continuous-valued features to cluster indices (sect 2.5). The following were defined for a given nominal observation \( x \) associated with the word at time \( t \): (a) 1 when \( x_{t-1} \) falls in cluster \( i \); (b) 1 when \( x_t \) falls in cluster \( j \); (c) 1 when \( x_{t+1} \) falls in cluster \( k \); (d) 1 when \( x_{t-1} \) falls in cluster \( i \) and \( x_t \) falls in cluster \( j \); (e) 1 when \( x_t \) falls in clusters \( j \) and \( x_{t+1} \) falls in cluster \( k \); (f) 1 when \( x_{t-1} \) falls in cluster \( i \) and \( x_t \) falls in cluster \( j \) and \( x_{t+1} \) falls in cluster \( k \). These templates of unigram, bigram and trigram features were applied to the observations introduced earlier as follows: Templates 1 through 3 were applied to all 258 observations. The combination templates (4 through 6) were applied to all observations excluding the energy observations \( (e_1-e_{210}) \) and the contour-type observation \( (f_1) \), which, at higher cardinality than the rest, was likely to have some data sparsity issues. Each template, therefore, can generate up to \( K \times L \) features for each observation, where \( K \) is the number of classes and \( L \) the cardinality of the observation (or combinations thereof) considered by the template. The final set input to the CRF contained 495 indicator features. For the experiments reported next we made use of the CRF optimization toolkits [7] and [8], obtaining comparable results.

\footnote{The term “feature” is used differently from the previous section where it meant measurable observable; in this context it is binary-valued and encodes a configuration of output and input observables.}

4. Labeling of Prosodic Events with CRFs

In this section we explore the use of CRFs to automatically label prosodic events, specifically the presence or absence of a word-level pitch accent, and that of a major or minor boundary after a word. The baseline corpus we have worked with is a set of approximately 3700 utterances recorded by a professional female speaker, which was annotated using the full ToBI framework by an experienced ToBI annotator. No independent set of annotations, and therefore inter-rater agreement, is known for this corpus. From these annotations, a sparser labeling was produced indicating whether a word carried any kind of pitch accent, and whether its following boundary was one of three types: a minor boundary (corresponding to break indices 1 or 0), a major boundary (corresponding to break index 4), and any others (break indices 2 and 3 collapsed). 15% of the data was set aside for testing, and the remaining 85% was used for training. Since one of the goals of this work is to assess how robust the observations and system proposed are to limited training data, we additionally built training subsets of increasingly larger size from 1%, 5%, 10%, 25% and 66% of the full training set, ensuring that any subset was always contained within a subset of larger size: that is, we wish to explore the effect of adding more data to an experiment, not just of changing the size of the corpus. This “reference” speaker will be henceforth denoted by REF. Another goal is to explore how well such a baseline system generalizes to speakers for which no prosodic labels are available to train a supervised system, and how this could be improved with some unsupervised adaptation. To explore this we used data from 3 males and 3 female speakers from the Boston University Radio Speech Corpus (BUR, henceforth) for which similar prosodic annotations are available. No knowledge of the labels was used at any point of the training procedure; the labels were only used to score the final system predictions.

Fig. 1(a) shows the performance of the pitch-accent classifier on the REF and BUR corpora as a function of the amount of training data. For a given set size, the system is learned from the training subset, and evaluated on the REF test set and on all of the BUR set (with results averaged across all 6 speakers). F-Measure is used throughout to summarize performance; all systems were in fact well-balanced in terms of precision and recall, and the F-Measure does not differ by much from either of those two metrics, which have been omitted here due to space constraints. Fig. 1(a) also shows another set of curves to address the following: the memory features computed by Eq. 1 for the 5 prosodic events were tabulated once, based on the full REF training set, and stored for recall whenever needed. Inspection of systems trained on different set sizes reveals that these features are among the most discriminative, raising the objection that smaller sets are not in fact performing based on the limited training data, but are exploiting information outside the training set. Rather than retabulate these features at every step, we instead look at the extreme case of leaving them out altogether. The performance when no memory features are used is shown by the curves \( REF_n \) and \( BUR_n \), which almost match that of the original feature set. This is an interesting result because it shows how the classifier is still able to find an alternate set of features that compensate for the most discriminative ones with barely a loss of performance, and points to redundancies in our feature set which make our results very robust. Since a simi-
lar trend is observed with boundary classification, we resume exploration with the full feature set for the rest of the experiments.

The adaptation experiments are conducted as follows. For a given size of the training subset, the baseline REF system is used to decode the BUR corpus and obtain the probability of the most likely sequence of labels. All utterances whose probability exceeds a threshold are filtered and used in a further retraining step according to one of two strategies. In the first of these, which we call the “joint” strategy, the TRF training subset is augmented with the filtered data (and corresponding decoded labels) to retrain a system. In the second “unlabeled” strategy, only the filtered data and decoded labels are used to retrain the system. Fig. 1(b) shows the performance of these two strategies (labeled $\text{BUR}_{J}$ and $\text{BUR}_{UL}$, respectively) after filtering above a threshold of $p = 0.5$ compared with the baseline, unadapted REF system. Figs. 1(c) through 1(f) show an analogue set of experiments to those just described for the case of minor and major boundary classification tasks respectively, where adaptation has been done with a filtering threshold of $p = 0.4$.

5. Discussion

Fig. 1(a) shows that the pitch-accent labeling system is robust to changes in the training size, both when applied to the REF speaker and, though with a drop in performance, data outside a training set. At 83.5% this counts among the highest figures reported for pitch-accent classification on the BUR corpus, especially in light of the fact the system is speaker-independent and uses no labeled training data from that corpus. On the REF dataset, it is gratifying to see that the baseline performance of 90.5% degrades by less than 5% when only a small fraction of the training data is used, a result that suggests more optimal strategies for collecting annotations from human labelers to minimize time and costs. The simple adaptation strategies we have used seem most useful when operating with small amounts of training data, when we partially recover from the drop in performance. With larger training subsets, however, this adaptation can actually hurt. More sophisticated schemes for adapting are the focus of current and future work, though we feel these preliminary results are the focus of current and future work, though we feel these preliminary results

6. References


