Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence

Sankaranarayanan Ananthakrishnan, *Student Member, IEEE*, and Shrikanth S. Narayanan, *Senior Member, IEEE*

***Abstract*—With the advent of prosody annotation standards such as tones and break indices (ToBI), speech technologists and linguists alike have been interested in automatically detecting prosodic events in speech. This is because the prosodic tier provides an additional layer of information over the short-term segment-level features and lexical representation of an utterance. As the prosody of an utterance is closely tied to its syntactic and semantic content in addition to its lexical content, knowledge of the prosodic events within and across utterances can assist spoken language applications such as automatic speech recognition and translation. On the other hand, corpora annotated with prosodic events are useful for building natural-sounding speech synthesizers. In this paper, we build an automatic detector and classifier for prosodic events in American English, based on their acoustic, lexical, and syntactic correlates. Following previous work in this area, we focus on accent (prominence, or "stress") and prosodic phrase boundary detection at the syllable level. Our experiments achieved a performance rate of 86.75% agreement on the accent detection task, and 91.61% agreement on the phrase boundary detection task on the Boston University Radio News Corpus.**

***Index Terms*—Accent, prominence, prosodic phrase boundary, prosody recognition, prosody–syntax interface, spoken language processing, stress.**

I. Introduction

SPOKEN utterances are characterized not only by segment-level (spectral) correlates of each sound unit, but also by a variety of suprasegmental effects that operate at a level higher than the local phonetic context [1]. The most prominent among these are as follows:

- modulation of intensity to impart emphasis to certain syllables or words;
- modulation of intonation patterns which reflect the class of the utterance (question, affirmation, etc.) as well as the speaker's intent and emotional state;
- timing, which refers to subtle variations in the rate and length of syllables, coupled with pauses that serve to separate linguistic "phrases" within the utterances.

These suprasegmental effects occur at the syllable, word, and utterance level. Together, they encode rhythm, intonation, and lexical stress, which constitute the prosody of spoken utterances. As human listeners make heavy use of the above cues in the understanding process, they evidently carry a lot of information that is likely to be useful for spoken language understanding and generation systems [2], [3].

A. Motivation for Prosodic Event Annotation

As mentioned previously, prosody can be very useful because it encodes aspects of higher level information not completely revealed by segmental acoustics. Below, we list sample scenarios where prosody can play an important role in augmenting the abilities of spoken language systems.

1) *Speech Act Detection*: Intonation patterns at the end of an utterance can provide an indication of specific speech acts or utterance categories (question, statement, exclamation, etc.).
2) *Word Disambiguation*: Knowledge of syllable stress or accent patterns can help in word or word-category disambiguation; a common example for this is the word *content*, which functions as a noun when stress is imparted to the first syllable (*con*-tent), and as an adjective when stress is imparted to the second syllable (con-*tent*).
3) *Speech Recognition*: The correlation between accent/prominence patterns and words can be exploited to build joint lexical-prosodic models which can improve speech recognition performance in terms of reducing word-error rate (WER).
4) *Natural Speech Synthesis*: One of the challenges in natural-sounding speech synthesis systems is to generate human-like prosody to accompany the segmental acoustic properties. This includes local effects (such as syllable accent), suitably timed boundaries, which reflect the syntactic structure of the utterance, as well as modulation of pitch at a global level to produce appropriate intonation patterns.

However, most current systems either completely disregard such information, or use it in limited, unprincipled ways for the simple reason that there is no established way to employ them. The main issues with using prosodic cues for spoken language applications are 1) the asynchronous nature of acoustic–prosodic features and consequently 2) the difficulty in modeling the relationship between the acoustic–prosodic features, segmental acoustics, lexical items, and syntactic structure of the utterance. Having a symbolic representation of prosodic events in terms of discrete labels greatly simplifies the task of learning these relationships; however, such discretization, if not performed carefully, may result in loss of information from the prosodic tier.

The tones and break indices (ToBI) annotation standard [4], [5] was developed in the early 1990s in an attempt to solve this problem, and to address the broader issue of representing prosodic events in spoken language in an unambiguous fashion.

Manuscript received April 12, 2006; revised August 14, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Helen Meng.

The authors are with the Signal and Image Processing Institute (SIPI), University of Southern California, Los Angeles, CA 90089 USA (e-mail: ananthak@usc.edu; shri@sipi.usc.edu).

Digital Object Identifier 10.1109/TASL.2007.907570

As such, ToBI is not a perfect scheme, and has been accused over the years of harboring several deficiencies [6], but it is the closest there is to a standard annotation system, and has been accepted as such by speech technologists and linguists working in this area.

B. ToBI Annotation Scheme

The ToBI standard uses four interrelated "tiers" of annotation in order to capture prosodic events in spoken utterances.

1) The orthographic tier contains a plain-text transcription of the spoken utterance.
2) The tone tier marks the presence of *pitch accents* and *prosodic phrase boundaries*, which are defined as follows. A pitch accent can be broadly thought of as a prominence or stress mark. Two basic types of accents, high (H) and low (L) are defined, based on the value of the fundamental frequency (F0) with respect to its vicinity; more fine-grained accent marks, such as low-high $(\mathrm{L} + \mathrm{H}^*)$ and high-low $(\mathrm{H} + \mathrm{L}^*)$ are based on the shape of the F0 contour in the immediate vicinity of the accent. Prosodic phrase boundaries serve to group together semantic units in the utterance. These are divided in two coarse categories, weak *intermediate phrase boundaries* and *full intonational phrase boundaries*, each of which can be high (H) or low (L).
3) The break-index tier marks the perceived degree of separation between lexical items (words) in the utterance. Break indices range in value from 0 through 4, with 0 indicating no separation, or *cliticization*, and 4 indicating a full pause, such as at a sentence boundary. This tier is strongly correlated with phrase boundary markings on the tone tier—boundary locations usually score 3 or above on the break index tier.
4) The miscellaneous tier is used to annotate any other information relevant to the utterance that is not covered by the other tiers. This may include annotation of nonspeech events such as disfluencies, etc.

A ToBI labeling guide along with several sample utterances annotated with ToBI labels is available in [5].

Although ToBI is by far the most well-known and widely used prosody annotation standard, it is not the only one in existence. The International Transcription System for Intonation [7] (INTSINT) is a standard that is intended to function as an International Phonetic Alphabet (IPA) for describing the intonation contour of an utterance. Eight discrete symbols (T: top, M: mid, B: bottom, H: higher, L: lower, S: same, U: upstep, and D: downstepped), are used to parameterize the intonation contour. Of these, the first three (T, M, and B) are *absolute*, i.e. defined with respect to the speaker's pitch range, while the other five (H, L, S, U, and D) are relative to the preceding target. The primary utility of INTSINT is to provide a parameterization of the overall intonation structure of the utterance, whereas ToBI is geared towards annotation of events that are linguistic in nature. As a result, INTSINT is more or less language independent, whereas a different version of ToBI has to be provided for each language (English, German, Japanese, Korean, and Greek are some of the languages for which complete ToBI system descriptions exist [8]).

Other prosody annotation systems include Intonational Variation in English [9] (IViE), which is derived from ToBI and is geared towards analysis and comparison of the intonational variation among different dialects/varieties of English, and TILT [10], which provides a numerical (continuous) parameterization of the intonation contour (as opposed to symbolic parameterization in INTSINT).

Although some annotation or parameterization systems may be better suited to specific tasks than ToBI (for instance, INTSINT, and TILT are more suitable for parameterizing the intonation contour), ToBI is more general purpose and is well suited for capturing the connection between intonation and prosodic structure. Another factor that has motivated most previous work in automatic prosodic annotation to use ToBI or its subsets is the wide availability of the Boston University Radio News Corpus, described in Section II. This corpus has been hand-annotated with ToBI labels, and is now a standard data set for training and evaluating automatic prosody annotation techniques.

In this paper, we focus on detecting a simpler subset of elements in the ToBI tone tier—specifically, we are interested in determining the presence or absence of pitch accents and phrase boundaries, regardless of their fine categories. As the title of the paper suggests, these events can be detected not only from their acoustic correlates (energy, syllable duration, F0 range, and contour, etc.), but also from the lexical and syntactic elements contained in an enriched textual representation of the utterance. Such a representation might include the orthography, part-of-speech (POS) and even a syntactic parse of the orthography of the utterance. The relationship of prosody to the acoustic, lexical, and syntactic structure of spoken utterances is discussed in further detail in following sections.

C. Previous Work on ToBI-Like Prosodic Event Detection

Initial attempts at automatic detection of prosodic events are presented in the work by Wightman *et al.* [11] and Ross and Ostendorf [12]. In [11], binary prominence and boundary labels were assigned to syllables based on posterior probabilities computed from acoustic evidence (such as F0, energy, and duration features) using a decision tree, combined with a probabilistic (bigram) model of accent and boundary patterns. Their method achieved an accuracy of 84% for prominence detection and 71% accuracy for boundary detection at the syllable level on the Boston University corpus. Thus, for prominence detection, they obtain performance levels that approach levels of agreement between human labelers (quoted as 86%–94%) for this task. However, their boundary detection performance is lower than agreement levels between human annotators (95%–98% for intonational phrase boundaries). In addition to prominence and boundary detection, they also conduct experiments on break index labeling, achieving an accuracy rate of 60% for exact index match and 90% for a match within $+/-1$ of the true index.

In [12], the authors present an automatic pitch accent and boundary tone labeling system which predicts pitch accent labels and boundary tone types using a multilevel hierarchical model based on a decision tree framework. In addition to detecting presence versus absence of pitch accents, they also attempt to perform fine-grained labeling of accent and

boundary types. The fine-pitch accent categories include high, down-stepped, and low; fine boundary categories include L-L%, H-L%, and L-H%. With a single speaker training and test set, they obtained 87.7% accuracy for binary presence versus absence of pitch accent at the syllable level. Pitch accents detected using the syllable level decision tree were then classified into fine categories using a pitch accent type classifier. They obtain an accuracy of 72.4% for the three-class pitch accent categorization task, measured over the subset of syllables that were correctly marked by the pitch accent detector as being accented. However, since very few syllables carry the "low" pitch accent, this three-way classifier was only marginally better than a chance-level accent type assignment that assigned a "high" pitch accent to all syllables (chance level accuracy was 71.8%). Their boundary tone classifier operates at the intonational phrase level. Intonational phrases are identified as those segments marked with a break index value of 4 or above on the ToBI break index tier. For boundary locations identified in this deterministic fashion, the three-way boundary tone classifier produced boundary labels that were 66.9% accurate, as opposed to the chance level of 61.1%, where all boundary tones were labeled as L-L%. These accuracy figures are quoted at the intonational phrase level rather than at the word or syllable level.

Syrdal *et al.* [13] attempt to predict binary pitch accents and intonational boundary tones labels directly from lexical cues (text, punctuation, part-of-speech, etc.) using a text-to-speech (TTS) engine to obtain a "default" starting point for manual labelers. They determined that manual labeling of ToBI labels with this starting point was significantly faster than starting from "scratch" i.e. with no prior knowledge of pitch accent and boundary tone placement.

More recent efforts are reported in Chen *et al.* [14] and Ananthakrishnan and Narayanan [15]. The former used a Gaussian mixture model (GMM)-based acoustic–prosodic model and an artificial neural network (ANN)-based syntactic–prosodic model built from POS tags in a maximum-likelihood framework to achieve binary pitch accent detection accuracy of 84.21% and intonational boundary detection accuracy of 93.07% at the word level. The latter experimented with an ASR-like structure for prosodic event detection, using a coupled-HMM structure to model the dynamic prosodic features and an n-gram-based syntactic–prosodic model to obtain 75% agreement on the prominence detection task and 88% agreement on the boundary detection task (combining both intermediate and intonational phrase boundaries) at the syllable level. This system also has binary pitch accent and boundary event targets.

D. Our Current Approach

In this paper, we attempt to combine different sources of information to improve accent and boundary detection performance. While our focus in this paper is automatic annotation of corpora with prosodic event tags, we develop our model structure in a way that makes it easy to integrate with existing ASR architectures. We assume that, in addition to the speech data, we also have available the corresponding orthographic transcription annotated with POS tags. We collapse all categories of ToBI-style accent and boundary labels to single "accent" and "boundary" categories, respectively. Thus, we have two binary classification problems that we treat independently—presence versus absence of pitch accents, and presence versus absence of boundaries. We associate prosodic events with specific syllables, because the latter are traditionally regarded as the smallest linguistic units at which these phenomena manifest themselves.

Using statistics, we analyze the effect of these prosodic events on their acoustic correlates, such as F0, short-time energy, and timing cues; we also study their relationship to the syntactic part-of-speech, and to the lexical entities with which they correspond. Armed with this knowledge, we build classifiers that assign prosodic events to syllables from the unlabeled test data set using acoustic evidence extracted from the speech data. The classifiers also generate posterior probability scores for each class given the acoustic evidence. The relationship of prosody to syntactic POS and individual lexical items is exploited by building factored n-gram language models that capture such dependencies. Finally, for the pitch accent detection problem, we also incorporate prior knowledge from existing lexica that provide canonical pronunciation information (including stress marks) for a large body of words. Our work differs from previous efforts in the following respects.

- We use acoustic, lexical, and syntactic features as opposed to [11], who use only acoustic evidence, and [12] and [13], who use only lexical and syntactic features. Our lexical and syntactic feature set is much simpler than that used in [14]. In particular, we do not use syntactic phrase boundaries obtained from parsing the text for the boundary detection task.
- We detect pitch accent at the linguistic syllable level, similar to [11] and [12], but different from [14], who do so at the word level. This is because prosodic events such as pitch accents are associated with specific syllables, rendering this approach more suitable for tasks such as word disambiguation, where two words may have the same phonetic pronunciation, but different syllable accent patterns (see example in Section I-A)
- Previous work on boundary detection emphasizes intonational phrase boundaries only. Our boundary detection task is different, because we consider intermediate as well as intonational phrase boundaries as part of our "boundary" category. This is a much more difficult task than just intonational phrase boundary detection described in previous work.
- We use a maximum *a posteriori* (MAP) framework for prosodic event detection as opposed to the maximum-likelihood (ML) framework used in [14]. Moreover, we use an n-gram structure for our prosodic language model, which makes for easier processing and decoding using the Viterbi algorithm, as well as integration with existing automatic speech recognition (ASR) systems. Another novelty of our work is the use of factored backoff to estimate smooth probabilities for the prosodic language model (see Section IV-B)
- The work described in [14] makes use of a prosodic lexicon that encodes all possible combinations of pitch accent and phrase boundaries for a given word. While this improves performance by restricting the search space, building such

a lexicon is a time-consuming task that does not scale to other corpora. Our solution is to incorporate canonical stress information from a public domain electronic pronunciation dictionary within the statistical classification framework. We show that this corpus-independent approach leads to significant gains in pitch accent detection accuracy over using the lexical tokens alone.

The remainder of this paper is organized as follows. Section II discusses the data corpus used, and the acoustic, syntactic, and lexical features extracted from the data for training and testing. Section III presents analyses of the acoustic, syntactic, and lexical correlates of accent and boundary events. Section IV explains the basic architecture of our prosodic event-detection system, and the assumptions that underly the structure. Section V details the experiments we conducted and the prosody recognition results we obtained. Finally, Section VI contains a brief discussion of some of the open problems in this area, the limitations of our current approach, and how it may be improved and applied to spoken language systems.

II. Data Corpus and Features

The Boston University Radio News Corpus (BU-RNC) [16] is a database of broadcast news style read speech that contains ToBI-style prosodic annotations for part of the data. The availability of these annotations have made it the corpus of choice for most experiments on prosodic event detection and labeling, including all those cited in Section I-C. The database contains speech from three female ($f1a$, $f2b$, and $f3a$) and four male speakers ($m1b$, $m2b$, $m3b$, and $m4b$). Data labeled with ToBI-style labels is available for six speakers, namely $f1a$, $f2b$, $f3a$, $m1b$, $m2b$, and $m3b$, which amounts to about 3 h of speech. In addition to the raw speech and prosodic annotation, the BU-RNC also contains the following:

- orthographic (text) transcription corresponding to each utterance;
- word- and phone-level time-alignments from automatic forced-alignment of the transcription;
- POS tags corresponding to each token in the orthographic transcription.

In order to obtain time-alignments at the linguistic syllable level, we syllabify the orthographic transcriptions using a deterministic algorithm based on the rules of English phonology [17], and since the resultant syllables are simply vowel-centric collections of the underlying phone sequences, we are able to generate syllable-level time alignments from phone-level alignments, which are available in the corpus.

For our experiments, we pooled all utterances that were ToBI-transcribed and created five cross-validation training and test sets. We then pruned the test sets so that no story repetitions by the same speaker coexisted in the training and test partitions of a given cross-validation set. This resulted in a training set size of 37 047 syllables and a test set size of 7343 syllables, averaged across the five cross-validation sets. The average syllable vocabulary (number of unique syllables) of the training sets was 2850, while that of the test sets was 1623. The average number of out-of-vocabulary syllables in the test sets was 250 (15.4% relative to the test vocabulary). Of the syllables in the training sets, an average of 12 705 (34.3%) carried pitch accents, while 6307 (17.0%) were associated with boundary events (counting both intermediate and intonational phrase boundaries). Of the syllables in the test sets, an average of 2560 (34.9%) carried pitch accents, and 1304 (17.7%) were associated with boundary events. Thus, the training and test sets exhibit similar chance levels for pitch accent and boundary events.

With the enriched transcriptions available in this corpus, we are then able to extract a variety of acoustic, lexical, and syntactic features as described below.

A. Acoustic Features

Prosody has a marked effect on suprasegmental features such as F0, energy, and timing in the vicinity of the event. Accent and boundary events are marked by exaggerated movements of the F0 contour. Accented syllables show an increase in the local energy profile. Preboundary syllable lengthening is a subtle timing variation found in the vicinity of boundary events [18]. Our acoustic features are derived from these cues, and are listed below.

- Features derived from F0 include within-syllable F0 range ($f0\_range$), difference between maximum and average within-syllable F0 ($f0\_maxavg\_diff$), difference between minimum and average within-syllable F0 ($f0\_avgmin\_diff$), and difference between within-syllable average and utterance average F0 ($f0\_avgutt\_diff$).
- Features derived from timing cues include normalized vowel nucleus duration for each syllable (n_dur) and pause duration after the word-final syllable (p_dur, for boundary detection only)
- Features derived from energy include within-syllable energy range (e_range), difference between maximum and average within-syllable energy (e_maxavg_diff), and difference between minimum and average energy within the syllable (e_avgmin_diff).

The use of differences rather than absolute values for F0- and energy-related features serves to normalize the data against variation between speakers, especially between males and females, but preserves the variations produced by prosody. We normalized the syllable nucleus (vowel) duration on a per vowel-type basis, such that for each vowel-type, the normalized duration feature is zero mean and unit variance. This serves to eliminate absolute duration differences due to vowel-instrinsic properties (for example, the high-front vowel iy is usually much longer than the neutral $schwa$), while preserving differences due to pitch accent or boundary events. The pause duration feature used for boundary detection was not normalized. In addition to the above features, which are extracted directly from the speech data and the F0 track, we also include the number of phonemes in a syllable as an additional dimension to the acoustic feature vector. The complete set of acoustic features used for pitch accent and boundary detection is shown in Table I. Thus, our acoustic features are encoded as nine-dimensional vectors, one for each syllable. We do not consider acoustic dependencies across syllables.

B. Lexical and Syntactic Features

As we will demonstrate in Sections III and V, prosodic events in an utterance can be accurately predicted from the lexical and

TABLE I
ACOUSTIC FEATURES RANKED BY IMPORTANCE

Accent features	Boundary features
n_dur	*p_dur*
f0_avgmin_diff	*n_dur*
e_avgmin_diff	*f0_maxavg_diff*
e_range	*f0_range*
f0_range	*e_range*
f0_maxavg_diff	*f0_avgmin_diff*
n_phones	*e_maxavg_diff*
f0_avgutt_diff	*e_avgmin_diff*
e_maxavg_diff	*f0_avgutt_diff*

syntactic content of the underlying orthography [19]. For example, content words such as nouns, adjectives, and verbs are much more likely to contain prominent syllables than function words, such as articles and determiners. Phrase boundaries, too, are more likely to follow content words than function words. Similarly, certain syllables occur much more frequently in content words than in function words and are more likely to be accented than syllables that appear mostly in function words.

We use individual syllable tokens as lexical features and POS tags as syntactic features. For the accent detection problem, we also include the canonical stress pattern for the word; this is obtained from a standard pronunciation dictionary that includes stress marks. These features are used to build probabilistic models of prosodic event sequences. These "prosodic language models" have a structure similar to the word-level n-grams used in speech recognition, and are used to constrain and refine hypotheses generated by classifiers that operate only on the acoustic evidence.

III. STATISTICAL ANALYSES OF ACOUSTIC, LEXICAL, AND SYNTACTIC FEATURES

Before implementing classification algorithms using the features described in Section II, we analyze these features in order to determine which ones are important for classification, and to verify if indeed they are capable of discriminating between the prosodic categories that we wish to separate. For the acoustic features, the former is accomplished using a feature selection algorithm, and the latter, using statistical hypothesis tests. In the case of lexical and syntactic features, we collect frequency counts to establish what types of lexical items or POS tags correspond well with accent and boundary events. These tests are conducted on the entire labeled corpus. Details of these analyses are presented next.

A. Analysis of Acoustic Features

We conduct a feature selection experiment using the *information gain* criterion [20] in order to rank the acoustic features by importance. This was implemented using the WEKA machine learning toolkit [21]. Table I lists acoustic features for pitch accent and boundary detection in decreasing order of importance based on this criterion. According to this ranking criterion, syllable nucleus duration is the most important determinant of pitch accent. Pause duration and nucleus duration are key indicators of boundary events. F0 and energy range also play an important role in discriminating between presence versus nonpresence of accent and boundary events.

Given the nature of our acoustic features, a simple two-way hypothesis test can also be conducted in order to determine whether the acoustic features are likely to be useful for classification. We test each feature independently in order to determine whether the mean value of the feature differs significantly across the positive and negative samples for each classification problem. Here, "positive" samples refer to syllables that carry a pitch accent or boundary; conversely, "negative" samples correspond to those syllables that do not carry accent or boundary events. We define the null and alternate hypotheses as follows.

$H0$: the mean value of feature f_i does not differ between positive and negative samples.

$H1$: the mean value of feature f_i differs between positive and negative samples

Analysis of variance (ANOVA) [22] is a commonly used statistical test to determine if two population means are different. However, standard ANOVA assumes that the variable being tested is normally distributed within each category label. In our case, most of the features are nonnegative; hence, this assumption is invalid. We therefore use a nonparametric form of ANOVA known as the Kruskal–Wallis test, which only makes the assumption that the samples are drawn independently. The significance level (p-value) reported by this test is lower than 0.001 for all but two acoustic features for pitch accent detection ($f0\_avgutt\_diff$ and e_maxavg_diff), for which $p \leq 0.15$, corresponding to their low ranking by the information gain criterion. However, since they do carry some discrimination information, we include them in the acoustic feature set for pitch accent detection. The significance value reported by this test is below 0.001 for all features used in the boundary detection task. This indicates that the null hypothesis can be rejected with high confidence for most features. We conclude from this test that the acoustic features are likely to contain information that will discriminate between accent/boundary events and nonevents.

B. Analysis of Lexical and Syntactic Features

We use individual syllable tokens as lexical features and POS tags (at the word level) as syntactic features in order to predict prosodic events from nonacoustic evidence. We gather unigram frequency counts from these features in order to establish their relationship to accent and boundary events in the speech data. Fig. 1(a) and (b) shows the distribution of five randomly chosen syllable tokens between positive and negative samples of accent and boundary events, respectively. Each token appears more than 80 times in the training corpus. Fig. 1(c) and (d) shows a similar distribution for five randomly chosen POS tags in the corpus. Each of the tags considered appears several hundred, if not a few thousand times, in the corpus. In this case, since POS tags are associated with whole words rather than individual syllables, an accent label associated with a POS tag implies that one of the syllables that constitute the word is accented. Boundary events are associated only with word-final syllables; hence, a boundary label associated with a POS tag simply means that the

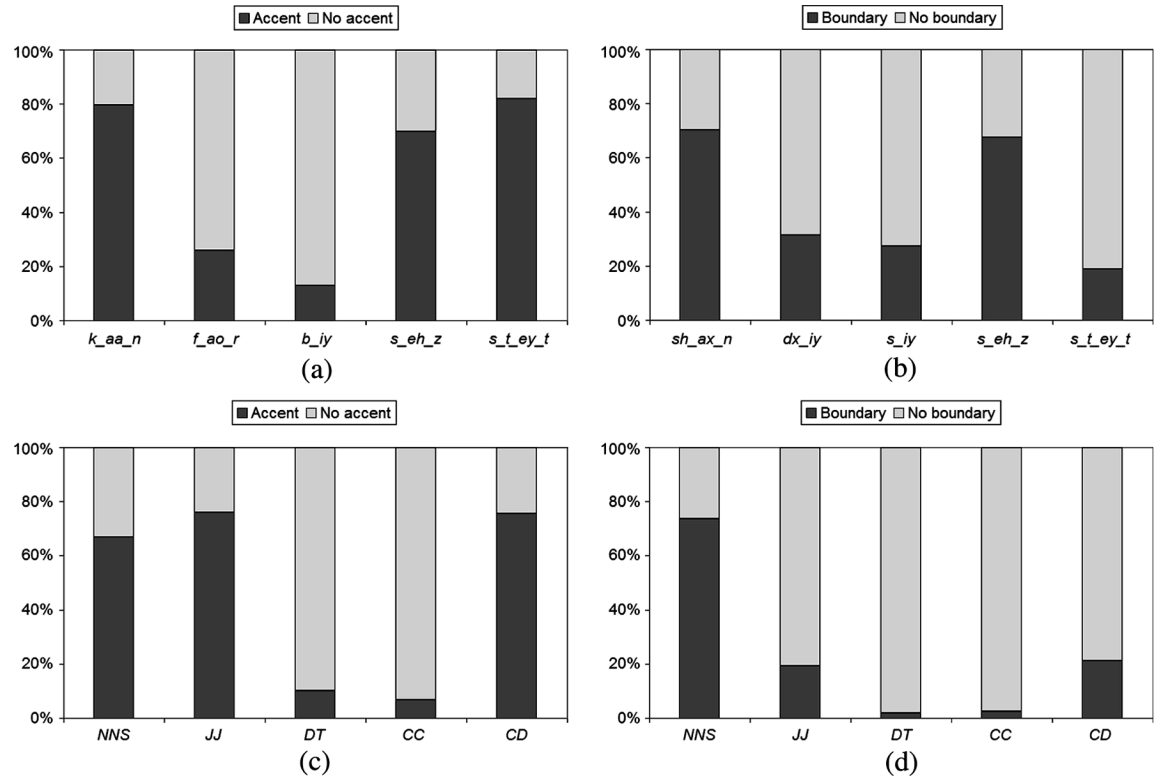

Fig. 1. Unigram frequency distributions of selected syllable tokens and part-of-speech (POS) tags between positive and negative classes for pitch accent and boundary detection tasks. The figures show a clear preference of syllable tokens for specific categories. POS tags corresponding to content words (NNS, JJ, etc.) are much more likely to be associated with accented words than those that correspond to function words (DT, CC, etc.). (a) Syllable-accent distribution. (b) Syllable-boundary distribution. (c) POS-accent distribution. (d) POS-boundary distribution.

final syllable of the corresponding word is at a boundary location.

These figures show a clear preference of certain syllable tokens and POS tags for specific prosodic events. For instance, in test data that statistically resemble the corpus used to compute the above statistics, there is approximately an 80% chance that the syllable token k_aa_n will be accented; on the other hand, the there is only a 13% chance that the token b_iy will be accented. Similarly, nouns (indicated by the tag NNS) have a 73% chance of being associated with boundary events, whereas adjectives (JJ) have only a 20% chance of being located at a prosodic phrase boundary. From this analysis of unigram frequency counts, we conclude that lexical and syntactic cues are likely to play an important role in recognition of accent and boundary events in speech.

IV. Architecture of the Prosodic Event Detector

Our prosodic event detector has a MAP structure and is modeled on the lines of a standard ASR system. We seek the sequence of prosodic events that maximizes the posterior probability of the event sequence given the acoustic, lexical, and syntactic evidence. In the following subsections, we develop the system architecture for each feature type separately, and then discuss feasible ways to merge them for performance improvement.

A. Prosodic Event Detection Using Acoustic Evidence

We wish to find the sequence of prosodic events $\mathbf{P}^* = \{p_1^*, p_2^*, \ldots, p_n^*\}$ such that

$$\mathbf{P}^* = \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{P}|\mathbf{A}) \tag{1}$$

$$= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{A}|\mathbf{P})p(\mathbf{P}) \tag{2}$$

where $\mathbf{A} = \{a_1, a_2, \ldots, a_n\}$ is the sequence of acoustic feature vectors, one for each syllable. Since our acoustic–prosodic classifiers return posterior probabilities $p(p_i|a_i)$, we can classify each syllable independently, in which case we approximate (1) as

$$\mathbf{P}^* = \arg\max_{\mathbf{P}\in\mathcal{P}} \prod_{i=1}^{n} p(p_i|a_i). \tag{3}$$

We can incorporate context information by using the form of (2), where it is possible to model $p(\mathbf{P})$ as an n-gram of prosodic labels (we call this a *de-lexicalized* prosodic language model). For a trigram language model

$$\begin{aligned}\mathbf{P}^* &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(a_1|p_1)p(a_2|p_2)p(p_1)p(p_2|p_1) \\ &\quad \cdot \prod_{i=3}^{n} p(a_i|p_i)p(p_i|p_{i-1}, p_{i-2}) \\ &= \arg\max_{\mathbf{P}\in\mathcal{P}} \alpha(p_1|a_1)\alpha(p_2|a_2)p(p_1)p(p_2|p_1) \\ &\quad \cdot \prod_{i=3}^{n} \alpha(p_i|a_i)p(p_i|p_{i-1}, p_{i-2})\end{aligned} \tag{4}$$

where

$$\alpha(p_i|a_i) = \frac{p(a_i|p_i)}{p(a_i)} = \frac{p(p_i|a_i)}{p(p_i)}$$

Eqation (4) is the architecture employed by Wightman *et al.* [11]. They use a bigram prosodic language model with a decision tree providing the label posterior probabilities. In this paper, we compare linear discriminant (LD), GMM, and neural network (NN) classifiers [23], [24] trained on acoustic features. Since the de-lexicalized prosodic LM has a binary vocabulary, it can be estimated very robustly even from small amounts of data. Thus, it is possible to model the prosody sequence using more context than it is to model word or syllable sequences; we use a 4-gram context for the prosodic LM. A small variation of this method is used for boundary detection; we specify that boundaries can only coincide with word-final syllables. Therefore, the terms $\alpha(p_i|a_i)$ are computed only for these syllables. For the word-initial and word-medial syllables, they are set to unimodal values so that the "no-boundary" event is always chosen.

B. Prosodic Event Detection Using Lexical Evidence

The most likely sequence of prosodic events $\mathbf{P}^*$ given only the sequence of syllables $\mathbf{S}$ can be found as follows:

$$\begin{aligned}\mathbf{P}^* &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{P}|\mathbf{S}) \\ &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{S},\mathbf{P}) \end{aligned} \tag{5}$$

where the joint distribution $p(\mathbf{S},\mathbf{P})$ can be modeled in an n-gram fashion; for example, a trigram approximation gives

$$p(\mathbf{S},\mathbf{P}) = p(s_1,p_1)p(s_2,p_2|s_1,p_1) \cdot \prod_{i=3}^{n} p(s_i,p_i|s_{i-1},p_{i-1},s_{i-2},p_{i-2}).$$

However, as detailed in Section II, the vocabulary of syllables is quite large in relation to the training corpus, and it is difficult to robustly estimate this distribution even with the n-gram approximation. Moreover, the test data exhibits a significant out-of-vocabulary (OOV) rate for the syllables (15.4% relative to the test vocabulary). We therefore employ a factored backoff scheme [25], where the probability of the current syllable-event pair is conditioned on previous syllable-event pairs, but backs off to lower order distributions by dropping syllable tokens if reliable estimates cannot be obtained for the full conditional distribution. Since the syllable token sequence is known, the distribution $p(s_i,p_i|s_{i-1},p_{i-1},s_{i-2},p_{i-2})$ may be replaced, without loss of generality, by the expression $p(p_i|s_i,s_{i-1},p_{i-1},s_{i-2},p_{i-2})$. In this scheme, if an unseen factor, such as an OOV syllable, occurs as a conditioning variable in the term $p(p_i|s_i,s_{i-1},p_{i-1},s_{i-2},p_{i-2})$, the backed off estimate that does not contain this variable is substituted for the complete expression. Fig. 2 shows the backoff graph we used for building the lexical–prosodic language model. The graph shows that we keep dropping lexical factors up to the point where we back off to the de-lexicalized prosodic LM described in Section IV-A. We use a fixed backoff path in this case. In practice, we use more (4-gram) history for the prosodic event factors and less (trigram) history for the syllable tokens.

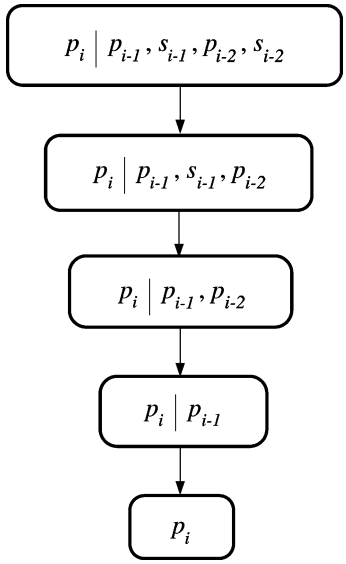

Fig. 2. Backoff graph for estimating lexical-prosodic LM. At each step, we drop a conditioning variable. Lexical tokens are dropped first.

C. Integrating Information From a Pronunciation Lexicon

The CMU dictionary [26] is a widely available pronunciation lexicon of over 125 000 words that is commonly used in large-vocabulary ASR tasks. In addition to a phonetic transcription of each word, it also encodes the canonical stress pattern for each word. We can look up each word in the test set from the lexicon and use the canonical stress pattern as another stream of evidence for pitch accent detection. We have, then, in addition to the syllable sequence $\mathbf{S}$, the sequence of canonical stress labels $\mathbf{L}$, whose elements are binary features. The problem then reduces to finding

$$\begin{aligned}\mathbf{P}^* &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{P}|\mathbf{S},\mathbf{L}) \\ &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{S},\mathbf{L},\mathbf{P}) \end{aligned} \tag{6}$$

where the joint distribution can be approximated by its n-gram factors in a manner similar to that described in Section IV-B. The sparsity problem can again be alleviated by the use of factored backoff, where in this case there are three factors per syllable instead of two.

D. Prosodic Event Detection Using Syntactic Evidence

We use syntactic evidence in the same way as we used lexical evidence to determine the most likely sequence of prosodic events

$$\begin{aligned}\mathbf{P}^* &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{P}|\mathbf{POS}) \\ &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{POS},\mathbf{P}) \end{aligned} \tag{7}$$

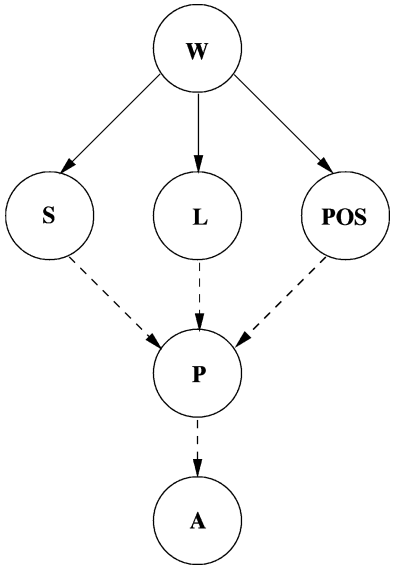

Fig. 3. Directed graph illustrating dependencies among variables. $\mathbf{W}$ is the sequence of words. $\mathbf{S}$, $\mathbf{L}$, and $\mathbf{POS}$ are the corresponding sequence of syllable tokens, canonical stress labels, and part-of-speech tags, respectively. $\mathbf{P}$ is the sequence of prosodic events and $\mathbf{A}$ is the sequence of acoustic–prosodic features. We treat the prosody labels as hidden variables influenced by (observed) lexical and syntactic features of the underlying orthography. The hidden prosodic event sequence generates acoustic observations.

where, as above, the joint distribution can be expressed as a product of its n-gram factors

$$p(\mathbf{POS},\mathbf{P}) = p(pos_1,p_1)p(pos_2,p_2|pos_1,p_1) \cdot \prod_{i=3}^{m} p(pos_i,p_i|pos_{i-1},p_{i-1},pos_{i-2},p_{i-2}).$$

This syntactic–prosodic distribution is much easier to estimate than the lexical-prosodic distribution, because the vocabulary of POS tags is quite small (ca. 30–35 tags in all), and hence it is easy to obtain robust estimates even from limited amounts of training data.

One difference between the syntactic– and lexical–prosodic models is that the former is built at the word level. This is not an issue for determining boundaries, because the boundary is constrained to coincide with the final syllable of the word; hence, there can only be one boundary event per word. No such restriction is placed on pitch accents, as they can be associated with any syllable within the word. Thus, for the pitch accent detection task, the syntactic–prosodic model indicates whether some syllable within the word is accented, but does not provide information as to which one is. Even so, this model helps eliminate false-positive decisions by constraining all syllables within nonaccented words to the "no-accent" tag. Note that we do not use information obtained from a syntactic parse of the orthography.

E. Combining Acoustic, Lexical, and Syntactic Evidence

We would like to combine each of the above streams of evidence $\{\mathbf{A},\mathbf{S},\mathbf{POS},\mathbf{L}\}$ in a principled fashion in order to maximize performance of the prosody recognizer. Fig. 3 illustrates the dependencies between these variables in the form of a directed graph. The solid arrows indicate deterministic relationships, while the dotted ones represent probabilistic dependencies that we must model. We take the view that the word sequence $\mathbf{W}$, or equivalently, its features $\{\mathbf{S},\mathbf{POS},\mathbf{L}\}$ are responsible for generating the prosody of the spoken utterance, which in turn modulates the acoustic parameters such as F0, energy, and duration, producing the acoustic feature sequence $\mathbf{A}$. In doing so, we ignore higher-level factors such as the utterance class (question, statement, etc.) and the speaker's emotional state that also play a role in determining the sequence of prosodic events. The observed variables in this graph are $\mathbf{W}$, the corresponding lexico-syntactic feature sequence $\{\mathbf{S},\mathbf{POS},\mathbf{L}\}$, and the acoustic feature sequence $\mathbf{A}$. The sequence of prosodic events $\mathbf{P}$ is to be inferred. Hence

$$\begin{aligned}\mathbf{P}^* &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{P}|\mathbf{A},\mathbf{S},\mathbf{L},\mathbf{POS}) \\ &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{A},\mathbf{S},\mathbf{L},\mathbf{POS}|\mathbf{P})p(\mathbf{P}) \\ &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{A}|\mathbf{P})p(\mathbf{S},\mathbf{L},\mathbf{POS}|\mathbf{P})p(\mathbf{P}) \qquad (8)\end{aligned}$$

where (8) follows from our assumption that the acoustic observations are conditionally independent of the lexical and syntactic features given the prosody labels. However, the distribution $p(\mathbf{S},\mathbf{L},\mathbf{POS}|\mathbf{P})$ cannot be robustly estimated because the joint vocabulary (ca. $2850\times 2\times 35$) is very large as compared to the available training data. We therefore use a naïve-Bayesian approximation such that the factors are easily and robustly estimated

$$p(\mathbf{S},\mathbf{L},\mathbf{POS}|\mathbf{P}) \approx p(\mathbf{S},\mathbf{L}|\mathbf{P})p(\mathbf{POS}|\mathbf{P}). \qquad (9)$$

Note that the feature sequence $\mathbf{L}$ is not available for the boundary detection task. Making this approximation, and substituting (9) into (8) gives

$$\begin{aligned}\mathbf{P}^* &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{A}|\mathbf{P})p(\mathbf{S},\mathbf{L}|\mathbf{P})p(\mathbf{POS}|\mathbf{P})p(\mathbf{P}) \\ &= \arg\max_{\mathbf{P}\in\mathcal{P}} p(\mathbf{A}|\mathbf{P})p(\mathbf{S},\mathbf{L},\mathbf{P})p(\mathbf{POS}|\mathbf{P}) \\ &= \arg\max_{\mathbf{P}\in\mathcal{P}} \frac{p(\mathbf{A}|\mathbf{P})}{p(\mathbf{P})}p(\mathbf{S},\mathbf{L},\mathbf{P})p(\mathbf{POS},\mathbf{P}) \qquad (10) \\ &= \arg\max_{\mathbf{P}\in\mathcal{P}} \frac{p(\mathbf{P}|\mathbf{A})}{p^2(\mathbf{P})}p(\mathbf{S},\mathbf{L},\mathbf{P})p(\mathbf{POS},\mathbf{P}) \\ &= \arg\max_{\mathbf{P}\in\mathcal{P}} \beta(\mathbf{P}|\mathbf{A})p(\mathbf{S},\mathbf{L},\mathbf{P})p(\mathbf{POS},\mathbf{P})\end{aligned}$$

where

$$\beta(\mathbf{P}|\mathbf{A}) = \frac{p(\mathbf{P}|\mathbf{A})}{p^2(\mathbf{P})}.$$

The combined recognition model hence reduces to a product of the individual acoustic, lexical, and syntactic models, respectively.

V. Experimental Results

We conduct a number of prosodic event detection experiments using acoustic, lexical, and syntactic cues, as discussed in Section IV. In this section, we describe our experimental setup and recognition results for the individual and combined models.

All performance figures in this section are obtained using five-fold cross validation with training and test splits as described in Section II. For each classification experiment, we list the accent and boundary detection accuracy as well as the corresponding false positive (FP) percentages. For the boundary detection task, we list overall detection accuracy as a fraction of all syllables, as well as word-final detection accuracy as a fraction of just the word-final (WF) syllables. The latter is a more useful metric, since word-initial and -medial syllables are always forced to the "no-boundary" category by our classifiers. We also report confidence intervals in terms of significance values (p-values) wherever we make comparisons between the performance of different classifiers and feature sets. We use the Wilcoxon signed rank test to compute significance values, because it is nonparametric, works with small sample sizes, and makes no assumptions regarding the distribution of the values (in this case, accuracy rates) being compared.

TABLE II
ACOUSTIC PROSODY RECOGNIZER: PERFORMANCE

	Accent	Accent FP
LD	71.15%	7.24%
GMM	72.18%	9.75%
NN	74.10%	8.64%
NN + de-lex LM	80.07%	10.14%

	Boundary		Boundary FP	
	All	**WF**	**All**	**WF**
LD	89.30%	83.47%	1.02%	1.68%
GMM	89.41%	83.65%	2.20%	3.63%
NN	89.99%	84.61%	2.30%	3.80%
NN + de-lex LM	89.59%	83.95%	5.09%	8.41%

A. Baseline

We set up a simple baseline based on the chance level of pitch accent and boundary events computed from the training data. Approximately 34% of training syllables carry an accent, while only about 17% of syllables coincide with boundaries. We form a lattice where each test syllable can take on positive or negative labels with the corresponding *a priori* chance level computed from the training corpus and rescore this lattice with the de-lexicalized prosodic LM to obtain a baseline system. The baseline pitch accent and boundary detection accuracies were 67.94% and 82.82% (overall), respectively. Note that our baseline boundary detection accuracy (based on the chance level for boundaries) is higher than the IPB detection accuracy of 71% reported in [11] for the radio news task. However, unlike [11], we provide figures for intermediate and intonational boundaries together.

B. Acoustic Prosodic Event Detector

We employed three different classifiers (LD, GMM, and NN) to obtain the prosody labels from the acoustic evidence. The GMM and NN classifiers also provide posterior probabilities for the prosodic events given the evidence. We first tested these classifiers in "independent-syllable" mode (3), and chose the best performing one for combination with the de-lexicalized prosodic LM.

1) LD Classifier: The LD classifier was used to obtain a simple baseline for classification based on acoustic evidence. The weights are trained using standard batch least-squares (the "pseudoinverse" method). This classifier achieved an independent syllable classification accuracy of 71.15% for pitch accent detection and 89.30% (overall) for the boundary detection task.

2) GMM Classifier: We trained GMM-based classifiers for pitch accent and boundary events using the EM algorithm. The number of component mixtures was chosen using the Bayesian information criterion (BIC). Although it not optimal in the sense of minimizing classification error, the BIC score provides a convenient way to select the number of mixtures based on a minimum-description length criterion. Specifically, the BIC score is simply the log-likelihood of the training data given the GMM parameters penalized by a function of the number of parameters and training samples. Based on this metric, the best choice for the number of mixtures was 18. This classifier achieved an independent syllable classification accuracy of 72.18% for the pitch accent detection task and 89.41% (overall) for the boundary detection task.

3) NN Classifier: The small difference in performance between the GMM and LD classifiers despite the large difference in the number of model parameters suggests that the acoustic features are not modeled well by GMMs. This led us to use a neural network for classifying prosodic features. We built a two-layer feedforward neural network with nine input units, 25 hidden units, and two output units, one for each class. We used linear activation for the input units, sigmoidal activation for the hidden units, and softmax activation for the output units. The neural network was trained using standard back-propagation. This classifier achieved an independent syllable classification accuracy of 74.10% for pitch accent detection task and 89.99% for the boundary detection task.

4) Acoustic Classifier + De-Lexicalized LM: We combine posterior label probabilities from the best performing acoustic classifier, the neural network, with label sequence constraints imposed by a 4-gram de-lexicalized prosodic LM. This is achieved by constructing a sausage lattice with prosodic variants of each syllable forming the lattice arcs. Each arc is weighted by the posterior probability assigned by the acoustic classifier (neural network). This lattice is then rescored with the n-gram de-lexicalized prosodic LM. This resulted in an absolute accuracy improvement of 5.97% for pitch accent detection (significant at $p \leq 0.05$). However, accuracy actually decreased by 0.4% ($p \leq 0.05$) for the boundary detection task. This is probably due to the fact that boundary events are quite far apart, and their context cannot be captured by narrow n-gram models. In the BU corpus, boundary events occur on average once every six to seven syllables; constructing such long range n-gram LMs is not feasible even for a binary vocabulary. We found this to be the case empirically as well; a 5-gram LM performed worse than the 4-gram with which we report the above results. Table II summarizes classification accuracy results using acoustic evidence.

TABLE III
LEXICAL/SYNTACTIC PROSODY RECOGNIZER: PERFORMANCE

	Accent	Accent FP
Tokens only	82.92%	8.09%
Incl. lexicon	85.17%	8.65%
Syntax only	70.70%	2.13%

	Boundary		Boundary FP	
	All	**WF**	**All**	**WF**
Tokens only	85.73%	77.59%	4.08%	6.75%
Syntax only	87.99%	81.31%	2.28%	3.77%

TABLE IV
COMBINED PROSODY RECOGNIZER: PERFORMANCE

	Accent	Accent FP
Baseline	67.94%	11.33%
Acoustic + Lexical (with pron.)	86.37%	7.64%
Acoustic + Syntactic	76.04%	7.25%
Acous. + Lex. + Syn. (no pron.)	86.06%	6.58%
Acous. + Lex. + Syn. (with pron.)	86.75%	8.08%
Word-level baseline	72.73%	6.66%
Combined system (word-level)	84.59%	9.33%

	Boundary		Boundary FP	
	All	**WF**	**All**	**WF**
Baseline	82.82%	72.78%	0.17%	0.28%
Acoustic + Lexical	90.41%	85.31%	4.63%	7.65%
Acoustic + Syntactic	91.61%	87.29%	4.61%	7.61%
Acous. + Lex. + Syn.	91.38%	86.91%	5.51%	9.11%

C. Lexical Prosodic Event Detector

In this setup, we attempt to uncover prosodic events using only lexical evidence, i.e. the syllable tokens and, for the accent detection task, the canonical stress sequence obtained from a pronunciation lexicon. The lexical-prosodic LMs were implemented using a factored backoff scheme according to Fig. 2 in order to alleviate problems due to data sparsity. We built these models using the $fngram$ tools that are part of the well-known SRILM toolkit [27]. The test transcriptions were used to construct unweighted lattices for each utterance; these lattices have a sausage structure and encode all possible combinations of syllable tokens and prosodic events for the corresponding utterances. They were then scored with the language model, and the best paths through the lattices were obtained using Viterbi search. This yielded the most likely sequence of prosodic events.

This experiment was conducted both with and without the canonical stress patterns from the pronunciation lexicon (for pitch accent detection) in order to study the effects of such *a priori* knowledge on system performance. The results are summarized in Table III. We observe that classification accuracy from syllable tokens alone exceeds the performance of a purely acoustic evidence based classifier by a significant margin (82.92% versus 80.07%, $p \leq 0.05$). However, prediction of boundary events using lexical evidence alone was 4.26% less accurate ($p \leq 0.05$) than predicting them using acoustic evidence. We also note, for the accent classification task, that the use of a pronunciation lexicon leads to an absolute classification accuracy improvement of 2.25% ($p \leq 0.05$) over a classifier that uses only the syllable tokens.

D. Syntactic Prosodic Event Detector

The structure of the syntactic prosodic event detector is similar to that of the lexical prosody recognizer, except for two differences. The first is that we use a standard backoff trigram to model the joint distribution of POS tags and prosodic events, which are treated as compound tokens. As mentioned earlier, the POS vocabulary is quite small, and no sparsity issues are likely to arise even with a relatively small training set. The second difference is that this recognizer detects prosodic events at the word level rather than at the syllable level. This is not an issue for the boundary detection task, as we force all non-word-final syllables to the negative label. However, the syntactic-prosodic LM does not influence classification of individual syllables that comprise the accented variant of a word. Syllables within an accented word are assigned pitch accents according to the chance level observed in training data. Table III summarizes accent and boundary detection accuracy for this recognizer. As expected, we observe that this method results in a performance gain over the lexical classifier for the boundary classification task (87.99% versus 85.73%, $p \leq 0.05$), but produces significantly worse results on the pitch accent detection task (70.70% versus 85.17%, $p \leq 0.05$), only slightly better than the baseline. This is expected, because for multisyllabic words that are identified as being accented, the syntactic model does not predict which syllable carries the pitch accent.

E. Combined Acoustic, Lexical, and Syntactic Prosodic Event Detector

Having tested prosodic event detection performance with each feature stream separately, we now combine them in accordance with (10). The issue of combining the syntactic-prosodic LM and lexical-prosodic LM arises again, because the former is built at the word level and the latter, at the syllable level. We address this problem by representing the syntactic lattice as a finite-state acceptor (FSA) and the word-to-syllable mapping as a finite-state transducer (FST). Scores from the acoustic model are embedded in the FST. The syntactic FSA is scored with the syntactic–prosodic LM and then composed with the mapping FST. This produces an syllable-token level FSA that incorporates syntactic and acoustic scores, which is finally rescored with the lexical-prosodic LM to obtain the best sequence of labels. We implemented the composition and other FSM operations with the AT&T FSM toolkit [28].

In addition to combining all feature streams, we also tested classifiers that used only acoustic and lexical features, and another that combined only acoustic and syntactic features. These experiments were conducted in order to examine the effects of the assumption underlying (9). Table IV summarizes the performance of the combined feature classifiers. We note that the combining all feature streams produces the most accurate pitch accent classification results, whereas boundary classification accuracy is highest for the classifier that combines only acoustic and syntactic evidence. Addition of the lexical feature stream actually decreases performance by 0.23%; however, this result was

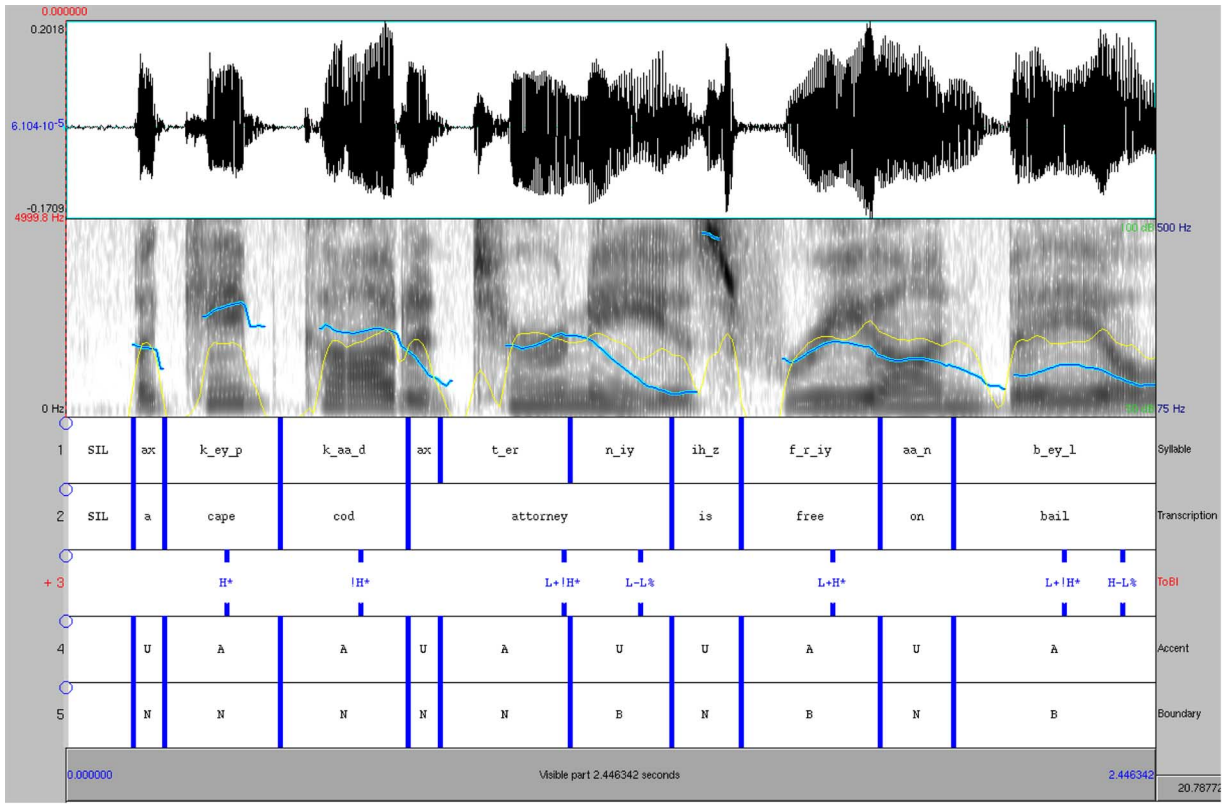

Fig. 4. Sample prosodic event detector output for utterance $f1as01p1$. The first 2.4 s of the utterance is shown. Tier 1 shows the speech signal; tier 2 shows the spectrogram with superimposed F0 and intensity tracks; tier 3 shows syllable-level transcription with time-alignments; tier 4 shows time-aligned word-level transcriptions; tier 5 shows ToBI pitch accents and boundaries as annotated in the corpus; tier 6 shows accent events assigned to syllables (U: unaccented, A: accented); tier 7 shows boundary events aligned with syllables (N: no boundary, B: boundary).

not significant at $p \leq 0.05$. This lack of performance improvement probably arises from the fact that lexical features (syllable tokens) are poorer indicators of boundary events than POS tags and therefore do not provide any additional information over the syntactic features. We note that for pitch accent detection, the combined system that uses canonical stress patterns from the pronunciation dictionary performs better than the combined system that does not use these stress patterns (86.75% versus 86.06%, $p \leq 0.05$). Finally, we also derive word-level pitch accent detection performance from syllable level annotations—a word carries a pitch accent if any syllable within that word is identified as carrying an accent. The baseline performance for this task was 72.73%, and on combining acoustic, lexical and syntactic features, we obtained a significant performance improvement up to 84.59%.

Fig. 4 shows a sample output of the combined prosodic event detector for the first 2.4 s of the utterance $f1bs01p1$ in a Praat TextGrid display. The figure contains seven tiers. Tier 1 shows the speech signal; tier 2 shows the spectrogram with superimposed F0 and intensity tracks; tier 3 shows syllable-level transcription with time-alignments; tier 4 shows time-aligned word-level transcriptions; tier 5 shows ToBI-style pitch accents and boundaries as annotated in the corpus; tier 6 shows accent events assigned to syllables (U = unaccented, A = accented); tier 7 shows boundary events aligned with syllables (N = no boundary, B = boundary). In this example, all pitch accent events were correctly identified and assigned to their corresponding syllables, but there is one error in the boundary tier, where the syllable f_r_iy has been assigned a boundary event, where, in fact, there is none (this may be attributed to the statistical nature of the boundary detector, similar to word errors in ASR—our system is, after all, a "speech recognizer" for prosodic events).

VI. Discussion and Future Work

In this paper, we developed a pitch accent detection system that obtained an accuracy of 86.75%, and a prosodic phrase boundary detector that obtained an accuracy of 91.61% at the linguistic syllable level on a human-annotated test set derived from the BU-RNC. Both systems approach the agreement level between human labelers for these tasks. We incorporated acoustic, lexical and shallow syntactic features within a MAP framework, which, combined with the n-gram prosodic language models, makes it easy to integrate with existing ASR systems. We determined from our experiments that lexical syllable tokens are useful for pitch accent detection, but not so effective for boundary detection. On the other hand, syntactic POS tags play an important role in boundary detection, but are not as useful for predicting pitch accent events. We also determined that canonical stress labels from a pronunciation dictionary are useful for pitch accent detection.

Our pitch accent detector performs better than that described in [11] (86.75% versus 84% at the syllable level). Although [14]

also report 84.21% accuracy on this task, the systems are not directly comparable, as we assign pitch accents to syllables, while their system operates at the word level. The work by Ross and Ostendorf [12] reports pitch accent detection accuracy of 87.7% at the syllable level; however, this is on a very small test set using data from only one speaker, whereas we report results on six speakers (three male, three female) using speaker-independent prosody models.

Our boundary detection task is more difficult than that described in previous work, because they focus only on intonational phrase boundary detection, whereas we consider both intermediate and intonational phrase boundaries as valid boundary events. Our boundary detection performance significantly exceeds that reported in [11] (91.61% versus 71% at the syllable level), but lags that reported in [14] (87.29% versus 93.07% at the word level). However, the figures quoted in [11] and [14] are for intonational phrase boundary detection only. Also, unlike [14], we do not use phrase opening/closing information from a syntactic parse of the text for the boundary detection task. Our task cannot be compared with the boundary tone classification problem described in [12], because they perform classification of boundary tones that have been deterministically identified from the ToBI break index tier, and not boundary tone detection itself.

As discussed in Section I-A, automatic recognition of pitch accent and boundary events in speech can be very useful for tasks such as word disambiguation, where a group of words may be phonetically similar but differ in placement of pitch accent, or in their location with respect to a boundary. At a higher level, knowledge of these prosodic events can be useful for spoken language understanding systems. For instance, in building a speech-to-speech translation system, we would like the suprasegmental structure in the target language to be equivalent to that of the utterance in the source language. Mapping prosodic events to a finite set of categories is a good starting point for this task.

There are several open problems that still need to be addressed. First, we work with binarized versions of the ToBI label set, disregarding the fine categories i.e. types of pitch accents and boundaries. These fine categories are annotated on the basis of the intonation pattern in the vicinity of the syllable associated with the prosodic event. However, to distinguish between these types using automatically extracted features is a difficult problem because 1) we rely on syllable time alignments generated from automatic forced alignment of the speech, which is not very accurate, and 2) intonation patterns used by human annotators to make these fine distinctions often occur in an asynchronous fashion, and do not always lie within the time-window indicated by forced alignment. As a result, extracting reliable features for distinguishing fine categories becomes very difficult. Indeed, most previous work, including that cited in this paper, focuses on binary pitch accent and boundary tone detection. An exception is [12], who report results on fine categorization of pitch accents and boundary tones. However, as their results show, fine categorization does not yield significant improvement over chance level category assignment (72.4% versus 71.8%) for three-way pitch accent categorization. For boundary tone locations deterministically identified from the ToBI break index tier, three-way classification of tone category was somewhat better (66.9% versus 61.1% for chance level assignment).

Second, for our current approach to be useful, we require a training corpus that is annotated with pitch accent and boundary labels. The BU-RNC is a broadcast news style corpus, and models trained with this data may not generalize well to spontaneous speech, which is usually the input for most spoken language understanding systems. We would therefore like to experiment with semisupervised and unsupervised techniques to perform the labeling task where such annotations are not available in the training set. Previous work on unsupervised prosodic event detection has focused exclusively on acoustic evidence [29]. In [30], we describe an unsupervised algorithm for accent and boundary event detection using acoustic, lexical, and part-of-speech evidence. The algorithm described in that paper uses information from an unsupervised clustering process to bootstrap lexical and syntactic probability models for improved performance.

Finally, in our current approach, we assume that the orthography and syntactic features (POS tags) corresponding to the spoken utterances are available. In many cases, however, we have only the speech utterance and wish to detect prosodic events directly from the acoustic signal, either for improving speech recognition performance or to extract other paralinguistic information such as speech acts, emotion, etc. One possible approach is discussed in Hasegawa-Johnson *et al.* [31], who use a lexical–syntactic–prosodic LM in order to simultaneously obtain word hypotheses as well as accent and boundary labels. More generally, incorporating prosodic cues in ASR to improve word recognition performance is a difficult problem, and we would like to see if operating at a lower level of granularity (such as accent and boundary events) will improve performance. In recent experiments [32], we obtained modest but statistically significant word-error rate improvement by reranking ASR n-best lists with prosody models similar to the ones described in this paper.

REFERENCES

[1] J. Terken, T. Hermes, M. Ostendorf, and N. Campbell, *Prosody: Theory and Experiment*. Norwell, MA: Kluwer, 2000, ch. 4,9,10.

[2] R. Kompe, *Prosody in Speech Understanding Systems*. New York: Springer-Verlag, 1997.

[3] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Proc. ISCA Workshop Prosody in Speech Recognition and Understanding*, 2001, pp. 13–16.

[4] 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 *Proc. Int. Conf. Spoken Lang. Process.*, 1992, pp. 867–869.

[5] M. Beckman and G. Elam, "Guidelines for ToBI Labeling." [Online]. Available: http://www.ling.ohio-state.edu/research/phonetics/E_ToBI

[6] C. Wightman, "ToBI or not ToBI?," in *Proc. Speech Prosody Conf.*, 2002, pp. 25–30.

[7] D. Hirst and A. Di Cristo, *Intonation Systems: A Survey of Twenty Languages*, D. Hirst and A. Di Cristo, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[8] "ToBI: The Ohio State University Department of Linguistics," The Ohio State Univ., Columbus, 1999 [Online]. Available: http://www.ling.ohio-state.edu/~tobi

[9] E. Grabe, F. Nolan, and K. Farrar, "IViE—A comparative transcription system for intonational variation in English," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, pp. 1259–1262.

[10] P. Taylor, "The TILT intonation model," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, vol. 4, pp. 1383–1386.
[11] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 469–481, Oct. 1994.
[12] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Comput. Speech Lang.*, vol. 10, pp. 155–185, 1996.
[13] A. Syrdal, J. Hirschberg, J. McGory, and M. Beckman, "Automatic ToBI prediction and alignment to speed manual labeling of prosody," *Speech Commun.*, vol. 33, pp. 135–151, 2001.
[14] 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. Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 1, pp. 509–512.
[15] S. Ananthakrishnan and S. Narayanan, "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2005, vol. 1, pp. 269–272.
[16] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," 1995.
[17] D. Kahn, "Syllable-based generalizations in English phonology," Ph.D. dissertation, Univ. Massachusetts, Boston, 1976.
[18] C. Wightman, "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Amer.*, pp. 1707–1717, 1992.
[19] S. Arnfield, "Prosody and syntax in corpus-based analysis of spoken English," Ph.D. dissertation, School Comput. Studies, Univ. Leeds , Leeds, U.K., 1994.
[20] J. R. Quinlan, "Induction of decision trees," in *Machine Learning*. Norwell, MA: Kluwer, 1986, vol. 1, pp. 81–106.
[21] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Mateo, CA: Morgan Kaufmann, 2005.
[22] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences*, 4th ed. Englewood Cliffs, NJ: Prentice-Hall, 1995.
[23] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2000.
[24] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1996.
[25] J. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proc. HLT/NAACL*, 2003, pp. 4–6.
[26] "The CMU Pronunciation Dictionary." [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[27] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. Int. Conf. Spoken Lang. Process.*, 2002, pp. 901–904.
[28] M. Mohri and M. Riley, "Weighted finite-state transducers in speech recognition (tutorial)," in *Proc. Int. Conf. Spoken Lang. Process.*, 2002.
[29] G.-A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent," in *Proc. HLT/NAACL*, Jun. 2006, pp. 224–231.
[30] S. Ananthakrishnan and S. Narayanan, "Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling," in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, pp. 297–300.
[31] M. Hasegawa-Johnson, J. Cole, C. Shih, K. Chen, A. Cohen, S. Chavarria, H. Kim, T. Yoon, S. Borys, and J.-Y. Choi, "Speech recognition models of the interdependence among syntax, prosody and segmental acoustics," in *Proc. HLT/NAACL*, 2004, pp. 56–63.
[32] S. Ananthakrishnan and S. Narayanan, "Improved speech recognition using acoustic and lexical correlates of pitch accent in a N-best rescoring framework," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Honolulu, HI, Apr. 2007, pp. IV-873–IV-876.

Sankaranarayanan Ananthakrishnan (S'07) received the B.Eng. degree in electronics and telecommunication engineering from the University of Bombay, Mumbai, India, in 2002 and the M.S. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2004. He is currently pursuing the Ph.D. degree in the Signal and Image Processing Institute (SIPI), USC.

His research interests include integration of higher level prosodic information for spoken language applications (speech recognition, parsing of spoken utterances, and spoken document matching and retrieval). Another aspect of his research involves experimental, data-driven verification of linguistic theories of prosody. In 2005, he interned with AT&T Labs-Research, where he was involved in building a speech recognizer for the Arabic language.

He is the recipient of a Best Student Paper Award (ICASSP '05).

Shrikanth S. Narayanan (S'88–M'95–SM'02) received the Ph.D. degree from the University of California, Los Angeles, in 1995.

He is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), Los Angeles, where he holds appointments as Professor in electrical engineering and jointly in computer science, linguistics, and psychology. Prior to joining USC, he was with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal Member, of its Technical Staff from 1995 to 2000. At USC, he is a member of the Signal and Image Processing Institute and a Research Area Director of the Integrated Media Systems Center, an NSF Engineering Research Center. He has published over 235 papers and has 14 granted/pending U.S. patents.

Dr. Narayanan is a recipient of an NSF CAREER Award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost Fellowship from the USC Center for Interdisciplinary Research, a Mellon Award for Excellence in Mentoring, and a recipient of a 2005 Best Paper Award from the IEEE Signal Processing Society. Papers by his students have won best student paper awards at ICSLP'02, ICASSP'05, and MMSP'06. He is an Editor for the *Computer Speech and Language Journal* (2007-present) and an Associate Editor for the IEEE *Signal Processing Magazine*. He was also an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2000–2004). He serves on the Speech Processing and Multimedia Signal Processing technical committees of the IEEE Signal Processing Society and the Speech Communication Committee of the Acoustical Society of America. He is a Fellow of the Acoustical Society of America and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu.