Published in final edited form as:
IEEE Trans Audio Speech Lang Processing. 2008 ; 16(4): 797–811. doi:10.1109/TASL.2008.917071.

Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework

Vivek Kumar Rangarajan Sridhar [Student Member][1], **Srinivas Bangalore**[2], and **Shrikanth S. Narayanan [Senior Member]**[1]

[1] The Viterbi School of Engineering, Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089-2564 USA (e-mail: vrangara@usc.edu; shri@sipi.usc.edu)

[2] AT&T Labs-Research, Florham Park, NJ 07932 USA (e-mail: srini@research.att.com)

Abstract

In this paper, we describe a maximum entropy-based automatic prosody labeling framework that exploits both language and speech information. We apply the proposed framework to both prominence and phrase structure detection within the Tones and Break Indices (ToBI) annotation scheme. Our framework utilizes novel syntactic features in the form of supertags and a quantized acoustic–prosodic feature representation that is similar to linear parameterizations of the prosodic contour. The proposed model is trained discriminatively and is robust in the selection of appropriate features for the task of prosody detection. The proposed maximum entropy acoustic–syntactic model achieves pitch accent and boundary tone detection accuracies of 86.0% and 93.1% on the Boston University Radio News corpus, and, 79.8% and 90.3% on the Boston Directions corpus. The phrase structure detection through prosodic break index labeling provides accuracies of 84% and 87% on the two corpora, respectively. The reported results are significantly better than previously reported results and demonstrate the strength of maximum entropy model in jointly modeling simple lexical, syntactic, and acoustic features for automatic prosody labeling.

Index Terms

Acoustic; prosodic representation; maximum entropy model; phrasing; prominence; spoken language processing; supertags; suprasegmental information; ToBI annotation

I. Introduction

PROSODY is generally used to describe aspects of a spoken utterance's pronunciation which are not adequately explained by segmental acoustic correlates of sound units (phones). The prosodic information associated with a unit of speech, say, syllable, word, phrase, or clause, influences all the segments of the unit in an utterance. In this sense, they are also referred to as suprasegmentals [1] that transcend the properties of local phonetic context.

Prosody encoded in the form of intonation, rhythm, and lexical stress patterns of spoken language conveys linguistic and paralinguistic information such as emphasis, intent, attitude, and emotion of a speaker. On the other hand, prosody is also used by speakers to provide cues to the listener and aid in the appropriate interpretation of their speech. This facilitates a method to convey the intent of the speaker through meaningful chunking or phrasing of the sentence, and is typically achieved by breaking long sentences into smaller prosodic phrases. Two key prosodic attributes described above include **prominence** and **phrasing** [2].

Prosody in spoken language correlates with acoustic and syntactic features. Acoustic correlates of duration, intensity and pitch, such as syllable nuclei duration, short time energy, and fundamental frequency (f0) are some of the acoustic features that are used to express prosodic prominence or stress in English. Lexical and syntactic features such as parts-of-speech, syllable nuclei identity, syllable stress of neighboring words have also been shown to exhibit a high degree of correlation with prominence. Humans realize phrasing acoustically by pausing after a major prosodic phrase, accentuating the final syllable in a phrase, and/or by lengthening the final syllable nuclei before a phrase boundary. Prosodic phrase breaks typically coincide with syntactic boundaries [3]. However, prosodic phrase structure is not isomorphic to the syntactic structure [4], [5].

Incorporating prosodic information can be beneficial in speech applications such as text-to-speech synthesis, automatic speech recognition, and natural language understanding, dialog act detection and even speech-to-speech translation. Accounting for the correct prosodic structure is essential in text-to-speech synthesis to produce natural sounding speech with appropriate pauses, intonation, and duration. Speech understanding applications also benefit from being able to interpret the recognized utterance through the placement of correct prosodic phrasing and prominence. Speech-to-speech translation systems can also greatly benefit from the marking of prosodic phrase boundaries, e.g., providing this information could directly help in building better phrase-based statistical machine translation systems. The integration of prosody in these applications is preempted by two main requirements:

1. a suitable and appropriate representation of prosody (e.g., categorical or continuous);
2. algorithms to automatically detect and seamlessly integrate the detected prosodic structure in speech applications.

Prosody is highly dependent on the individual speaker style, gender, dialect, and phonological factors. Nonuniform acoustic realizations of prosody are characterized by distinct intonation patterns and prosodic constituents. These distinct intonation patterns are typically represented using either symbolic or parametric prosodic labeling schemes such as Tones and Break Indices (ToBI) [6], TILT intonational model [7], Fujisaki model [8], Intonational Variation in English (TViE) [9], and International Transcription System for Intonation (INTSINT) [10]. These prosodic labeling approaches provide a common framework for characterizing prosody and hence facilitate development of algorithms and computational modeling frameworks for automatic detection and subsequent integration of prosody within various speech applications. While detailed categorical representations are suitable for text-to-speech synthesis, speech, and natural language understanding tasks, simpler prosodic representations in terms of raw or speaker normalized acoustic correlates of prosody have also been shown to be beneficial in many speech applications such as disfluency detection [11], sentence boundary detection [12], parsing [13], and dialog act detection [14]. As long as the acoustic correlates are reliably extracted under identical conditions during training and testing, an intermediate symbolic or parametric representation of prosody can be avoided, even though they may provide additional discriminative information if available. In this paper, we use the ToBI labeling scheme for categorical representation of prosody.

Prior efforts in automatic prosody labeling have utilized a variety of machine learning techniques, such as decision trees [2], [15], rule-based systems [16], bagging and boosting on decision trees [17], hidden Markov models (HMMs) [18], coupled HMMs [19], neural networks [20], and conditional random fields [21]. These algorithms typically exploit lexical, syntactic, and acoustic features in a supervised learning scenario to predict prosodic constituents characterized through one of the aforementioned prosodic representations.

The interplay between acoustic, syntactic, and lexical features in characterizing prosodic events has been successfully exploited in text-to-speech synthesis [22], [23], dialog act modeling

[24], [25], speech recognition [20], and speech understanding [2]. The procedure in which the lexical, syntactic, and acoustic features are integrated plays a vital role in the overall robustness of automatic prosody detection. While generative models using HMMs typically perform a front-end acoustic–prosodic recognition and integrate syntactic information through back-off language models [19], [20], stand-alone classifiers use a concatenated feature vector combining the three sources of information [21], [26]. We believe that a discriminatively trained model that jointly exploits lexical, syntactic, and acoustic information would be the best suited for the task of prosody labeling. We present a brief synopsis of the contribution of this paper in the following section.

A. Contributions of This Work

We present a discriminative classification framework using maximum entropy modeling for automatic prosody detection. The proposed classification framework is applied to both prominence and phrase structure prediction, two important prosodic attributes that convey vital suprasegmental information beyond the orthographic transcription. The prominence and phrase structure prediction is carried out within the ToBI framework designed for categorical prosody representation. We perform automatic pitch accent and boundary tone detection, and break index prediction, that characterize prominence and phrase structure, respectively, with the ToBI annotation scheme.

The primary motivation for the proposed work is to exploit lexical, syntactic, and acoustic–prosodic features in a discriminative modeling framework for prosody modeling that can be easily integrated in a variety of speech applications. The following are some of the salient aspects of our work.

1. Syntactic Features:
    - We propose the use of novel syntactic features for prosody labeling in the form of supertags which represent dependency analysis of an utterance and its predicate-argument structure, akin to a shallow syntactic parse. We demonstrate that inclusion of supertag features can further exploit the prosody-syntax relationship compared to that offered by using parts-of-speech tags alone.
2. Acoustic Features:
    - We propose a novel representation scheme for the modeling of acoustic–prosodic features such as energy and pitch. We use *n*-gram features derived from the quantized continuous acoustic–prosodic sequence that is integrated in the maximum entropy classification scheme. Such an *n*-gram feature representation of the prosodic contour is similar to representing the acoustic–prosodic features with a piecewise linear fit as done in parametric approaches to modeling intonation.
3. Modeling:
    - We present a maximum entropy framework for prosody detection that jointly exploits lexical, syntactic, and prosodic features. Maximum entropy modeling has been shown to be favorable for a variety of natural language processing tasks such as part-of-speech tagging, statistical machine translation, sentence chunking, etc. In this paper, we demonstrate the suitability of such a framework for automatic prosody detection. The proposed framework achieves state-of-the-art results in pitch accent, boundary tone, and break index detection on the Boston University (BU)

Radio News Corpus [30] and Boston Directions Corpus (BDC) [31], two publicly available read speech corpora with prosodic annotation.

- Our framework for modeling prosodic attributes using lexical, syntactic, and acoustic information is at the word level, as opposed to syllable level. Thus, the proposed automatic prosody labeler can be readily integrated in speech recognition, text-to-speech synthesis, speech translation, and dialog modeling applications.

The rest of the paper is organized as follows. In Section II, we describe some of the standard prosodic labeling schemes for representation of prosody, particularly, the ToBI annotation scheme that we use in our experiments. We discuss related work in automatic prosody labeling in Section III followed by a description of the proposed maximum entropy algorithm for prosody labeling in Section IV. Section V describes the lexical, syntactic, and acoustic–prosodic features used in our framework and Section VI-A describes the data used. We present results of pitch accent and boundary tone detection, and break index detection in Sections VII and VIII, respectively. We provide discussion of our results in Section IX and conclude in Section X along with directions for future work.

II. Prosodic Labeling Standards

Automatic detection of prosodic prominence and phrasing requires appropriate representation schemes that can characterize prosody in a standardized manner and hence facilitate design of algorithms that can exploit lexical, syntactic, and acoustic features in detecting the derived prosodic representation. Existing prosody annotation schemes range from those that seek comprehensive representations for capturing the various multiple facets of prosody to those that focus on exclusive categorization of certain prosodic events.

Prosodic labeling systems can be categorized into two main types: linguistic systems, such as ToBI [6], which encode events of linguistic nature through discrete categorical labels and parametric systems, such as TILT [7] and INTSINT [10] that aim only at providing a configurational description of the macroscopic pitch contour without any specific linguistic interpretation. While TILT and INTSINT are based on numerical and symbolic parameterizations of the pitch contour and hence are more or less language independent, ToBI requires expert human knowledge for the characterization of prosodic events in each language (e.g., Spanish ToBI [28] and Japanese ToBI [29]). In contrast, the gross categorical descriptions within the ToBI framework offer a level of uncertainty in the human annotation to be incorporated into the labeling scheme and hence provide some generalization, considering that prosodic structure is highly speaker dependent. They also provide more general-purpose description of prosodic events encompassing acoustic correlates of pitch, duration, and energy compared to TILT and INTSINT that exclusively model the pitch contour. Furthermore, the availability of large prosodically labeled corpora with manual ToBI annotations, such as the Boston University (BU) Radio News Corpus [30] and Boston Directions Corpus (BDC) [31], offer a convenient and standardized avenue to design and evaluate automatic ToBI-based prosody labeling algorithms.

Several linguistic theories have been proposed to represent the grouping of prosodic constituents [6], [32], [33]. In the simplest representation, prosodic phrasing constituents can be grouped into *word*, *minor phrase*, *major phrase*, and *utterance* [1]. The ToBI break index representation [6] uses indices between 0 and 4 to denote the perceived disjuncture between each pair of words, while the perceptual labeling system described in [32] represents a superset of prosodic constituents by using labels between 0 and 6. In general, these representations are mediated by rhythmic and segmental analysis in the orthographic tier and associate each word with an appropriate index.

In this paper, we evaluate our automatic prosody algorithm on the Boston University Radio News Corpus and Boston Directions Corpus, both of which are hand annotated with ToBI labels. We perform both prominence and phrase structure detection that are characterized within the ToBI framework through the following parallel tiers: 1) a tone tier, and 2) a break-index tier. We provide a brief description of the ToBI annotation scheme and the associated characterization of prosodic prominence and phrasing by the parallel tiers in the following section.

A. ToBI Annotation Scheme

The ToBI [6] framework consists of four parallel tiers that reflect the multiple components of prosody. Each tier consists of discrete categorical symbols that represent prosodic events belonging to that particular tier.[1] A concise summary of the four parallel tiers is presented below. The reader is referred to [6] for a more comprehensive description of the annotation scheme.

- Orthographic Tier: The orthographic tier contains the transcription of the orthographic words of the spoken utterance.
- Tone tier: Two types of tones are marked in the tonal tier: pitch events associated with intonational boundaries, *phrasal tones or boundary tones*, and pitch events associated with accented syllables, *pitch accents*. The basic tone levels are high (H) and low (L), and are defined based on the relative value of the fundamental frequency in the local pitch range. There are a total of five pitch accents that lend prominence to the associated word: {H*, L*, L*+H, L+H*, H+ !H*}. The phrasal tones are divided in two coarse categories, weak *intermediate phrase boundaries* {L−, H−}, and *full intonational phrase boundaries* {L – L%, L – H%, H – H%, H – L%} that group together semantic units in the utterance.
- Break index tier: The break-index tier marks the perceived degree of separation between lexical items (words) in the utterance and is an indicator of prosodic phrase structure. 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 tone markings on the tone tier.
- Miscellaneous tier: This may include annotation of non-speech events such as disfluencies, laughter, etc.

The detailed representation of prosodic events in the ToBI framework, however, suffers from the drawback that all the prosodic events are not equally likely, and hence a prosodically labeled corpus would consist of only a few instances of one event while comprising a majority of another. This in turn creates serious data sparsity problems for automatic prosody detection and identification algorithms. This problem has been circumvented to some extent by decomposing the ToBI labels into intermediate or coarse categories such as presence or absence of pitch accents, phrasal tones, etc., and performing automatic prosody detection on the decomposed inventory of labels. Such a grouping also reduces the effects of labeling inconsistency. A detailed illustration of the label decompositions is presented in Table I. In this paper, we use the coarse representation (presence versus absence) of pitch accents, boundary tones, and break indices to alleviate the data sparsity and compare our results with previous work.

[1]On a variety of speaking styles, Pitrelli *et al.* [38] have reported inter-annotator agreements of 83%–88%, 94%–95%, and 92.5%, respectively, for pitch accent, boundary tone, and break index detection within the ToBI annotation scheme.

III. Related work

In this section, we survey previous work in prominence and phrase break prediction with an emphasis on ToBI-based pitch accent, boundary tones, and break index prediction. We present a brief overview of speech applications that have used such prosodic representations along with algorithms and their corresponding performance on the various prosody detection and identification tasks.

A. Pitch Accent and Boundary Tone Labeling

Automatic prominence labeling through pitch accents and boundary tones, has been an active research topic for over a decade. Wightman and Ostendorf [2] developed a decision-tree algorithm for labeling prosodic patterns. The algorithm detected phrasal prominence and boundary tones at the syllable level. Bulyko and Ostendorf [22] used a prosody prediction module to synthesize natural speech with appropriate pitch accents. Verbmobil [39] incorporated prosodic prominence into a translation framework for improved linguistic analysis and speech understanding.

Pitch accent and boundary tone labeling has been reported in many past studies [15], [19], [20]. Hirschberg [15] used a decision-tree based system that achieved 82.4% speaker-dependent accent labeling accuracy at the word level on the BU corpus using lexical features. Wang and Hirschberg [37] used a CART-based labeling algorithm to achieve intonational phrase boundary classification accuracy of 90.0%. Ross and Ostendorf [34] also used an approach similar to [2] to predict prosody for a text-to-speech (TTS) system from lexical features. Pitch accent accuracy at the word level was reported to be 82.5% and syllable-level accent accuracy was 87.7%. Hasegawa-Johnson *et al.* [20] proposed a neural network based syntactic–prosodic model and a Gaussian mixture model-based acoustic–prosodic model to predict accent and boundary tones on the BU corpus that achieved 84.2% accuracy in accent prediction and 93.0% accuracy in intonational boundary prediction. With syntactic information alone, they achieved 82.7% and 90.1% for accent and boundary prediction, respectively. Ananthakrishnan and Narayanan [19] modeled the acoustic–prosodic information using a coupled hidden Markov model that modeled the asynchrony between the acoustic streams. The pitch accent and boundary tone detection accuracy at the syllable level were 75% and 88%, respectively. Yoon [40] has recently proposed memory-based learning approach and has reported accuracies of 87.78% and 92.23% for pitch accent and boundary tone labeling. The experiments were conducted on a subset of the BU corpus with 10 548 words and consisted of data from same speakers in the training and test set.

More recently, pitch accent labeling has been performed on spontaneous speech in the Switchboard corpus. Gregory and Atlun [21] modeled lexical, syntactic, and phonological features using conditional random fields and achieved pitch accent detection accuracy of 76.4% on a subset of words in the Switchboard corpus. Ensemble machine learning techniques such as bagging and random forests on decision trees were used in the 2005 JHU Workshop [36] to achieve pitch accent detection accuracy of 80.4%. The corpus used was a prosodic database consisting of spontaneous speech from the Switchboard corpus [41]. Nenkova *et al.* [35] have reported a pitch accent detection accuracy of 76.6% on a subset of the Switchboard corpus using a decision tree classifier.

Our proposed maximum entropy discriminative model outperforms previous work on prosody labeling on the BU and BDC corpora. On the BU corpus, with syntactic information alone we achieve pitch accent and boundary tone accuracy of 85.2% and 91.5% on the same training and test sets used in [20] and [27]. These results are statistically significant by a difference of proportions test.[2] Further, the coupled model with both acoustic and syntactic information results in accuracies of 86.0% and 93.1%, respectively. The pitch accent improvement is

statistically significant compared to results reported in [27] by a difference of proportions test. On the BDC corpus, we achieve pitch accent and boundary tone accuracies of 79.8% and 90.3%. The proposed work uses speech and language information that can be reliably and easily extracted from the speech signal and orthographic transcription. It does not rely on any hand-coded features [35] or prosody labeled lexicons [20]. The results of previous work on pitch accent and boundary tone detection on the BU corpus are summarized in Table II.

B. Prosodic Phrase Break Labeling

Automatic intonational phrase break prediction has been addressed mainly through rule-based systems developed by incorporation of rich linguistic rules, or, data-driven statistical methods that use labeled corpora to induce automatic labeling information [2], [26], [42], [43]. Typically, syntactic information like part-of-speech (POS) tags, syntactic structure (parse features), as well as acoustic correlates like duration of pre-boundary syllables, boundary tones, pauses and f0 contour have been used as features in automatic detection and identification of intonational phrase breaks. Algorithms based on machine learning techniques such as decision trees [2], [26], [44], HMM [42], or combination of these [43] have been successfully used for predicting phrase breaks from text and speech.

Automatic detection of phrase breaks has been addressed mainly from the intent of incorporating the information in text-to-speech systems [26], [42], to generate appropriate pauses and lengthening at phrase boundaries. Phrase breaks have also been modeled from the interest of their utility in resolving syntactic ambiguity [13], [44], [45]. Intonational phrase break prediction is also important in speech understanding [2], where the recognized utterance needs to be interpreted correctly.

One of the first efforts in automatic prosodic phrasing was presented by Ostendorf and Wightman [2]. Using the seven-level break index proposed in [32], they achieved an accuracy of 67% for exact identification and 89% correct identification within ±1. They used a simple decision tree classifier for this task. Wang and Hirschberg [37] have reported an overall accuracy of 81.7% in detection of phrase breaks through a CART-based scheme. Ostendorf and Veilleux [45] achieved 70% accuracy for break correct prediction, while, Taylor and Black [42], using their HMM-based phrase break prediction based on POS tags have demonstrated 79.27% accuracy in correctly detecting break indices. Sun and Applebaum [43] have reported F-scores of 77% and 93% on break and nonbreak prediction. Recently, ensemble machine learning techniques such as bagging and random forests that combined decision tree classifiers were used at the 2005 JHU workshop [36] to perform automatic break index labeling. The classifiers were trained on spontaneous speech [41] and resulted in break index detection accuracy of 83.2%. Kahn *et al.* [13] have also used prosodic break index labeling to improve parsing. Yoon [40] has reported break index accuracy of 88.06% in a three-way classification between break indices using only lexical and syntactic features.

We achieve a break index accuracy of 83.95% and 87.18% on the BU and BDC corpora using lexical and syntactic information alone. Our combined maximum entropy acoustic–prosodic model achieves a break index detection accuracy of 84.01% and 87.58%, respectively, on the two corpora. The results from previous work are summarized in Table III.

IV. Maximum Entropy Discriminative Model for Prosody Labeling

Discriminatively trained classification techniques have emerged as one of the dominant approaches for resolving ambiguity in many speech and language processing tasks. Models trained using discriminative approaches have been demonstrated to outperform generative

[2]Results at a level ≤ 0.001 were considered significant.

models as they directly optimize the conditional distribution without modeling the distribution of all the underlying variables. The maximum entropy approach can model the uncertainty in labels in typical NLP tasks and hence is desirable for prosody detection due to the inherent ambiguity in the representation of prosodic events through categorical labels. A preliminary formulation of the work in this section was presented by the authors in [46] and [47].

We model the prosody prediction problem as a classification task as follows: given a sequence of words w_i in an utterance $W = \{w_1,\ldots,w_n\}$, the corresponding syntactic information sequence $S = \{s_1,\ldots,s_n\}$ (e.g., parts-of-speech, syntactic parse, etc.), a set of acoustic–prosodic features $A = \{\mathbf{a}_1,\ldots, \mathbf{a}_n\}$, where $\mathbf{a}_i=(a_i^1,\ldots,a_i^{t_{w_i}})$ is the acoustic–prosodic feature vector corresponding to word w_i with a frame length of t_{w_i} and a prosodic label vocabulary $\mathcal{L} = \{l_1,\ldots,l_v\}$, the best prosodic label sequence $L^* = \{l_1,l_2,\ldots,l_n\}$ is obtained as follows:

$$L^* = \arg\max_L P(L|W,S,A). \tag{1}$$

We approximate the string level global classification problem, using conditional independence assumptions, to a product of local classification problems as shown in (3). The classifier is then used to assign to each word a prosodic label conditioned on a vector of local contextual features comprising the lexical, syntactic, and acoustic information:

$$L^* = \arg\max_L P(L|W,S,A) \tag{2}$$

$$\approx \arg\max_L \prod_{i=1}^{n} p\left(l_i|w_{i-k}^{i+k}, s_{i-k}^{i+k}, \mathbf{a}_{i-k}^{i+k}\right) \tag{3}$$

$$= \arg\max_L \prod_{i=1}^{n} p(l_i|\mathbf{\Phi}(W,S,A,i)) \tag{4}$$

Where $\mathbf{\Phi}(W,S,A,i)=(w_{i-k}^{i+k}, s_{i-k}^{i+k}, \mathbf{a}_{i-k}^{i+k})$ is a set of features extracted within a bounded local context k. $\mathbf{\Phi}$ (W, S, A, i) is shortened to $\mathbf{\Phi}$ in the rest of the section.

To estimate the conditional distribution $P(l_i|\mathbf{\Phi})$, we use the general technique of choosing the maximum entropy (maxent) distribution that estimates the average of each feature over the training data [48]. This can be written in terms of the Gibbs distribution parameterized with weights λ_l where l ranges over the label set and V is the size of the prosodic label set. Hence

$$P(l_i|\mathbf{\Phi}) = \frac{e^{\lambda_{l_i}\cdot\mathbf{\Phi}}}{\sum_{l=1}^{V} e^{\lambda_l\cdot\mathbf{\Phi}}}. \tag{5}$$

To find the global maximum of the concave function in (5), we use Sequential L1-Regularized Maxent algorithm (SL1-Max) [49]. Compared to iterative scaling (IS) and gradient descent procedures, this algorithm results in faster convergence and provides L1-regularization as well as efficient heuristics to estimate the regularization meta-parameters. We use the machine learning toolkit LLAMA [50] to estimate the conditional distribution using maxent. LLAMA

encodes multiclass maxent as binary maxent to increase the training speed and to scale the method to large data sets. We use here V one-versus-other binary classifiers. Each output label l is projected onto a bit string, with components $b_j(l)$. The probability of each component is estimated independently:

$$\begin{aligned} P(b_j(l)|\mathbf{\Phi}) &= 1 - P(\bar{b}_j(l)|\mathbf{\Phi}) \\ &= \frac{e^{\lambda_j \cdot \mathbf{\Phi}}}{e^{\lambda_j \cdot \mathbf{\Phi}} + e^{\bar{\lambda}_j \cdot \mathbf{\Phi}}} \\ &= \frac{1}{1 + e^{-(\lambda_j - \bar{\lambda}_j) \cdot \mathbf{\Phi}}} \end{aligned} \tag{6}$$

where λ_j is the parameter vector for $\bar{b}_j(y)$. Assuming the bit vector components to be independent, we have

$$P(l_i|\mathbf{\Phi}) = \prod_{j=1}^{V} P(b_j(l_i)|\mathbf{\Phi}). \tag{7}$$

Therefore, we can decouple the likelihoods and train the classifiers independently. In this paper, we use the simplest and most commonly studied code, consisting of V one-versus-others binary components. The independence assumption states that the output labels or classes are independent.

V. Lexical, Syntactic, and Acoustic Features

In this section, we describe the lexical, syntactic, and acoustic features that we use in our maximum entropy discriminative modeling framework. We use only features that are derived from the local context of the text being tagged, referred to as static features hereon (see Table IV). One would have to perform a Viterbi search if the preceding prediction context were to be added. Using static features is especially suitable for performing prosody labeling in lockstep with recognition or dialog act detection, as the prediction can be performed incrementally instead of waiting for the entire utterance or dialog to be decoded.

A. Lexical and Syntactic Features

The lexical features used in our modeling framework are simply the words in a given utterance. The BU and BDC corpora that we use in our experiments are automatically labeled (and hand-corrected) with POS tags. The POS inventory is the same as the Penn treebank which includes 47 POS tags: 22 open class categories, 14 closed class categories, and 11 punctuation labels. We also automatically tagged the utterances using the AT&T POS tagger. The POS tags were mapped into function and content word categories[3] and were added as a discrete feature.

In addition to the POS tags, we also annotate the utterance with Supertags [51]. Supertags encapsulate predicate-argument information in a local structure. They are the elementary trees of Tree-Adjoining Grammars (TAGs) [52]. Similar to part-of-speech tags, Supertags are associated with each word of an utterance, but provide much richer information than part-of-speech tags, as illustrated in the example in Table V. Supertags can be composed with each other using substitution and adjunction operations [52] to derive the predicate-argument structure of an utterance.

[3]Function and content word features were obtained through a look-up table based on POS.

There are two methods for creating a set of Supertags. One approach is through the creation of a wide coverage English grammar in the lexicalized tree adjoining grammar formalism, called XTAG [53]. An alternate method for creating supertags is to employ rules that decompose the annotated parse of a sentence in Penn Treebank into its elementary trees [54], [55]. This second method for extracting supertags results in a larger set of supertags. For the experiments presented in this paper, we employ a set of 4726 supertags extracted from the Penn Treebank.

There are many more supertags per word than part-of-speech tags, since supertags encode richer syntactic information than part-of-speech tags. The task of identifying the correct supertag for each word of an utterance is termed as supertagging [51]. Different models for supertagging that employ local lexical and syntactic information have been proposed [56]. For the purpose of this paper, we use a maximum entropy supertagging model that achieves a supertagging accuracy of 87% [57].[4]

While there have been previous attempts to employ syntactic information for prosody labeling [44], [58], which mainly exploited the local constituent information provided in a parse structure, supertags provide a different representation of syntactic information. First, supertags localize the predicate and its arguments within the same local representation (e.g., *give* is a di-transitive verb) and this localization extends across syntactic transformations (relativization, passivization, wh-extraction), i.e., there is a different supertag for each of these transformations for each of the argument positions. Second, supertags also factor out recursion from the predicate-argument domain. Thus, modification relations are specified through separate supertags as shown in Table V. For this paper, we use the supertags as labels, even though there is a potential to exploit the internal representation of supertags as well as the dependency structure between supertags as demonstrated in [59]. Table V shows the supertags generated for a sample utterance in the BU corpus.

B. Acoustic–Prosodic Features

The BU corpus contains the corresponding acoustic–prosodic feature file corresponding to each utterance. The f0 and root mean square (rms) energy (e) of the utterance along with features for distinction between voiced/unvoiced segments, cross-correlation values at estimated f0 values, and ratio of first two cross correlation values are computed over 10-ms frame intervals. The pitch values for unvoiced regions are smoothed using linear interpolation. In our experiments, we use these values rather than computing them explicitly which is straightforward with most audio processing toolkits. Both the energy and the f0 levels were range normalized (znorm) with speaker specific means and variances. Delta and acceleration coefficients were also computed for each frame. The final feature vector has six dimensions comprising f0, Δf0, Δ^2f0, e, Δe, and Δ^2e per frame.

We model the frame level continuous acoustic–prosodic observation sequence as a discretized sequence through quantization (see Fig. 1). We perform this on the normalized pitch and energy extracted from the time segment corresponding to each word. The quantized acoustic stream is then used as a feature vector. For this case, (3) becomes

$$L^* \approx \arg\max_L \prod_i^n p(l_i|\mathbf{\Phi}) = \arg\max_L \prod_i^n p(l_i|\mathbf{a}_i) \tag{8}$$

[4]The model is trained to disambiguate among the supertags of a word by using the lexical and part-of-speech features of the word and of six words in the left and right context of that word. The model is trained on one million words of supertag annotated text.

Where $\mathbf{a}_i=(a_i^1,\ldots,a_i^{t_{w_i}})$, the acoustic–prosodic feature vector corresponding to word w_i with a frame length of t_{w_i}.

The quantization, while being lossy, reduces the vocabulary of the acoustic–prosodic features, and hence offers better estimates of the conditional probabilities. The quantized acoustic–prosodic cues are then modeled using the maximum entropy model described in Section IV. The *n*-gram representation of quantized continuous features is similar to representing the acoustic–prosodic features with a piecewise linear fit as done in the TILT international model [7]. Essentially, we leave the choice of appropriate representations of the pitch and energy features to the maximum entropy discriminative classifier, which integrates feature selection during classification.

The proposed scheme of quantized *n*-gram prosodic features as input to the maxent classifier is different from previous work [60]. Shriberg *et al.* [60] have proposed *n*-grams of Syllable-based Nonuniform Extraction Region Features (SNERF-grams) for speaker recognition. In their approach, they extract a large set of prosodic features such as maximum pitch, mean pitch, minimum pitch, durations of syllable onset, coda, nucleus, etc., and quantize these features by binning them. The resulting syllable-level features, for a particular bin resolution, are then modeled as either unigram (using current syllable only), bigram (current and previous syllable or pause), or trigram (current and previous two syllables or pauses). They use support vector machines (SVMs) for subsequent classification. Our framework, on the other hand, models the macroscopic prosodic contour in its entirety by using *n*-gram feature representation of the quantized prosodic feature sequence. This representation coupled with the strength of the maxent model to handle large feature sets and in avoiding overtraining through regularization makes our scheme attractive for capturing characteristic pitch movements associated with prosodic events.

VI. Experimental Evaluation

A. Data

All the experiments reported in this paper are performed on the Boston University (BU) Radio News Corpus [30] and the Boston Directions Corpus (BDC) [31], two publicly available speech corpora with manual ToBI annotations intended for experiments in automatic prosody labeling. The BU corpus consists of broadcast news stories including original radio broadcasts and laboratory simulations recorded from seven FM radio announcers. The corpus is annotated with orthographic transcription, automatically generated and hand-corrected part-of-speech tags and automatic phone alignments. A subset of the corpus is also hand annotated with ToBI labels. In particular, the experiments in this paper are carried out on four speakers similar to [27], two males and two females referred to hereafter as **m1b**, **m2b**, **f1a**, and **f2b**. The BDC corpus is made of elicited monologues produced by subjects who were instructed to perform a series of direction-giving tasks. Both spontaneous and read versions of the speech are available for four speakers **h1**, **h2**, **h3**, and **h4** with hand-annotated ToBI labels and automatic phone alignments, similar to the BU corpus. Table VI shows some of the statistics of the speakers in the BU and BDC corpora.

In all our prosody labeling experiments, we adopt a leave-one-out speaker validation similar to the method in [20] for the four speakers with data from one speaker for testing and those from the other three for training. For the BU corpus, speaker **f2b** was always used in the training set since it contains the most data. In addition to performing experiments on all the utterances in BU corpus, we also perform identical experiments on the train and test sets reported in [27] which is referred to as Hasegawa-Johnson *et al.* set.

VII. Pitch Accent and Boundary Tone Labeling

In this section, we present pitch accent and boundary tone labeling results obtained through the proposed maximum entropy prosody labeling scheme. We first present some baseline results, followed by the description of results obtained from our classification framework.

A. Baseline Experiments

We present three baseline experiments. One is simply based on chance where the majority class label is predicted. The second is a baseline only for pitch accents derived from the lexical stress obtained through look-up from a pronunciation lexicon labeled with stress. Finally, the third baseline is obtained through prosody detection in current off-the-shelf speech synthesis systems. The baseline using speech synthesis systems is comparable to our proposed model that uses lexical and syntactic information alone. For experiments using acoustics, our baseline is simply chance.

1) Acoustic Baseline (Chance)—The simplest baseline we use is chance, which refers to the majority class label assignment for all tokens. The majority class label for pitch accents is presence of a pitch accent (**accent**) and that for boundary tone is absence (**none**).

2) Prosody Labels Derived From Lexical Stress—Pitch accents are usually carried by the stressed syllable in a particular word. Lexicons with phonetic transcription and lexical stress are available in many languages. Hence, one can use these lexical stress markers within the syllables and evaluate the correlation with pitch accents. Even when the lexicon has a closed vocabulary, letter-to-sound rules can be derived from it for unseen words. For each word carrying a pitch accent, we find the particular syllable where the pitch accent occurs from the manual annotation. For the same syllable, we assign a pitch accent based on the presence or absence of a lexical stress marker in the phonetic transcription. The CMU pronunciation lexicon was used for predicting lexical stress through simple lookup. Lexical stress for out-of-vocabulary words was predicted through a CART based letter-to-sound rule derived from the pronunciation lexicon. The results are presented in Table VII.

3) Prosody Labels Predicted Using TTS Systems—We perform prosody prediction using two off-the-shelf speech synthesis systems, namely, AT&T NV speech synthesizer and Festival. The AT&T NV speech synthesizer [61] is a half phone speech synthesizer. The toolkit accepts an input text utterance and predicts appropriate ToBI pitch accent and boundary tones for each of the selected units (in this case, a pair of phones) from the database. The toolkit uses a rule-based procedure to predict the ToBI labels from lexical information [15]. We reverse mapped the selected half phone units to words, thus obtaining the ToBI labels for each word in the input utterance. The pitch accent labels predicted by the toolkit are $L_{\text{accent}} \varepsilon\{$**H***, **L***, **none**$\}$ and the boundary tones are $L_{\text{btone}} \varepsilon\{$**L** – **L**%, **H** – **H**%, **L** – **H**%, none$\}$.

Festival [62] is an open-source unit selection speech synthesizer. The toolkit includes a CART-based prediction system that can predict ToBI pitch accents and boundary tones for the input text utterance. The pitch accent labels predicted by the toolkit are $L_{\text{accente}} \varepsilon$ {**H***, **L**+**H***, !**H***, **none**}, and the boundary tones are $L_{\text{btone}} \varepsilon\{$**L** – **L**%, **H** – **H**%, **L** – **H**%, **none**$\}$. The prosody labeling results obtained through both the speech synthesis engines are presented in Table VII.

B. Maximum Entropy Pitch Accent and Boundary Tone Classifier

In this section, we present results of our maximum entropy pitch accent and boundary tone classification. We first present a maximum entropy syntactic–prosodic model that uses only lexical and syntactic information for prosody detection, followed by a maximum entropy

acoustic–prosodic model that uses an *n*-gram feature representation of the quantized acoustic–prosodic observation sequence.

1) Maximum Entropy Syntactic-Prosodic Model—The maximum entropy syntactic–prosodic model uses only lexical and syntactic information for prosody labeling. Our prosodic label inventory consists of $L_{accent}\ \varepsilon\{$**accent**, **none**$\}$ for pitch accents and $L_{btone}\ \varepsilon\{$**btone**, **none**$\}$ for boundary tones. Such a framework is beneficial for text-to-speech synthesis that relies on lexical and syntactic features derived predominantly from the input text to synthesize natural sounding speech with appropriate prosody. The results are presented in Table VIII. In Table VIII, correct POS tags refer to hand-corrected POS tags present in the BU corpus release and POS tags refers to parts-of-speech tags predicted automatically.

Prosodic prominence and phrasing can also be viewed as joint events occurring simultaneously. Previous work by [2] suggests that a joint labeling approach may be more beneficial in prosody labeling. In this scenario, we treat each word to have one of the four labels $l_i\ \varepsilon\mathcal{L}$ = {**accent** – **btone**, **accent** – **none**, **none** – **btone**, **none** – **none**}. We trained the classifier on the joint labels and then computed the error rates for individual classes. The joint modeling approach provides a marginal improvement in the boundary tone prediction but is slightly worse for pitch accent prediction.

2) Maximum Entropy Acoustic–Prosodic Model—We quantize the continuous acoustic–prosodic values by binning and extract *n*-gram features from the resulting sequence. The quantized acoustic–prosodic *n*-gram features are then modeled with a maxent acoustic–prosodic model similar to the one described in Section 5. Finally, we append the syntactic and acoustic features to model the combined stream with the maxent acoustic-syntactic model, where the objective criterion for maximization is (1). The two streams of information were weighted in the combined maximum entropy model by performing optimization on the training set (weights of 0.8 and 0.2 were used on the syntactic and acoustic vectors, respectively). The pitch accent and boundary tone prediction accuracies for quantization performed by considering only the first decimal place is reported in Table IX. As expected, we found the classification accuracy to drop with increasing number of bins used in the quantization due to the small amount of training data. In order to compare the proposed maxent acoustic–prosodic model with conventional approaches such as HMMs, we also trained continuous observation density HMMs to represent pitch accents and boundary tones. This is presented in detail in the following section.

C. HMM Acoustic–Prosodic Model

In this section, we compare the proposed maxent acoustic–prosodic model with a traditional HMM approach. HMMs have been demonstrated to capture the time-varying pitch patterns associated with pitch accents and boundary tones effectively [18], [19]. We trained separate context-independent HMMs with three state left-to-right topology with uniform segmentation. The segmentations need to be uniform due to lack of an acoustic–prosodic model trained on the features pertinent to our task to obtain forced segmentation. The acoustic observations of the HMM were unquantized acoustic–prosodic features described in Section V-B. The label sequence was decoded using the Viterbi algorithm.

The final label sequence using the maximum entropy syntactic–prosodic model and the HMM based acoustic–prosodic model was obtained by combining the syntactic and acoustic probabilities. Essentially, the prosody labeling task reduces to the following:

$$\begin{aligned} L^* &= \arg\max_L P(L|A, W) \\ &= \arg\max_L P(L|W) \cdot P(A|L, W) \\ &\approx \arg\max_L P(L|\Phi(W)) \cdot P(A|L)^\gamma \end{aligned} \tag{9}$$

where $\Phi(W)$ is the syntactic feature encoding of the word sequence W. The first term in (9) corresponds to the probability obtained through our maximum entropy syntactic model. The second term in (9) computed by an HMM corresponds to the probability of the acoustic data stream which is assumed to be dependent only on the prosodic label sequence. γ is a weighting factor to adjust the weight of the two models.

The syntactic–prosodic maxent model outputs a posterior probability for each class per word. We formed a lattice out of this structure and composed it with the lattice generated by the HMM acoustic–prosodic model. The best path was chosen from the composed lattice through a Viterbi search. The procedure is illustrated in Fig. 2. The acoustic–prosodic probability $P(A|L, W)$ was raised by a power of γ to adjust the weighting between the acoustic and syntactic model. The value of γ was chosen as 0.008 and 0.015 for pitch accent and boundary tone, respectively, by tuning on the training set. The results of the HMM acoustic–prosodic model and the coupled model are shown in Table IX. The weighted maximum entropy syntactic–prosodic model and HMM acoustic–prosodic model performs the best in pitch accent and boundary tone classification. We conjecture that the generalization provided by the acoustic HMM model is complementary to that provided by the maximum entropy model, resulting in slightly better accuracy when combined together as compared to that of a combined maxent-based acoustic and syntactic model.

VIII. Prosodic Break Index Labeling

We presented pitch accent and boundary tone labeling results using our proposed maximum entropy classifier in the previous section. In the following section, we address phrase structure detection by performing automatic break index labeling within the ToBI framework. Prosodic phrase break prediction has been especially useful in text-to-speech [42] and sentence disambiguation [44], [45] applications, both of which rely on prediction based on lexical and syntactic features. We follow the same format as the prominence labeling experiments, presenting baseline experiments followed by our maximum entropy syntactic and acoustic classification schemes. All the experiments are performed on the entire BU and BDC corpora.

A. Baseline Experiments

We present baseline experiments, both chance and break index labeling results using an off-the-shelf speech synthesizer. The AT&T Natural Voices speech synthesizer does not have a prediction module for prosodic break prediction, and hence we present results from using the Festival [62] speech synthesizer alone. Festival speech synthesizer produces simple binary break presence or absence distinction, as well as more detailed ToBI-like break index prediction.

1) Break Index Prediction in Festival—Festival can predict break index at the word level based on the algorithm presented in [42]. The toolkit can predict both, ToBI-like break values ($L_{\mathbf{tobi\_break}} \varepsilon\{\mathbf{0,1,2,3,4}\}$) and simple presence versus absence ($L_{\text{binary\_break}} \varepsilon\{\mathbf{B, NB}\}$). Only lexical and syntactic information is used in this prediction without any acoustics. Baseline classification results are presented in Table X.

B. Maximum Entropy Model for Break Index Prediction

1) Syntactic–Prosodic Model—The maximum entropy syntactic–prosodic model uses only lexical and syntactic information for prosodic break index labeling. Our prosodic label inventory consists of $L_{\mathbf{tobi\_break}} \varepsilon\{\mathbf{0,1,2,3,4}\}$ for ToBI based break indices and $L_{\text{binary\_break}} \varepsilon$ {**B, NB**} for binary break versus no-break distinction. The {**B, NB**} categorization was obtained by grouping break indices **0,1,2** into **NB** and **3,4** into **B** [6]. The classifier is then applied for break index labeling as described in Section VII-B1 for the pitch accent prediction. We assume knowledge of sentence boundary through the means of punctuation in all the reported experiments.

2) Acoustic–Prosodic Model—Prosodic break index prediction is typically used in text-to-speech systems and syntactic parse disambiguation. Hence, the lexical and syntactic features are crucial in the automatic modeling of these prosodic events. Further, they are defined at the word level and do not demonstrate a high degree of correlation with specific pitch patterns. We thus use only the maximum entropy acoustic–prosodic model described in Section VII-B2. The combined maximum entropy acoustic-syntactic model is then similar to (2), where the prosodic label sequence is conditioned on the words, POS tags, supertags, and quantized acoustic–prosodic features. A binary flag indicating the presence or absence of a pause before and after the current word was also included as a feature. The results of the maximum entropy syntactic, acoustic, and acoustic-syntactic model for break index prediction are presented in Table X. The maxent syntactic–prosodic model achieves break index detection accuracies of 83.95% and 87.18% on the BU and BDC corpora. The addition of acoustics to the lexical and syntactic features does not result in a significant improvement in detection accuracy. In these experiments, we used only pitch and energy features and did not use duration features such as rhyme duration, duration of final syllable, etc., used in [2]. Such features require both phonetic alignment and syllabification and therefore are difficult to obtain in speech applications that require automatic prosody detection to be performed in lockstep. Additionally, in the context of TTS systems and parsers, the proposed maximum entropy syntactic–prosodic model for break index prediction performs with high accuracy compared to previous work.

IX. Discussion

The automatic prosody labeling presented in this paper is based on ToBI-based categorical prosody labels but is extendable to other prosodic representation schemes such as IViE [9] or INTSINT [10]. The experiments are performed on decompositions of the original ToBI labels into binary classes. However, with the availability of sufficient training data, we can overcome data sparsity and provide more detailed prosodic event detection (refer to Table I). We use acoustic features only in the form of pitch and energy contour for pitch accent and boundary tone detection. Durational features, which are typically obtained through forced alignment of the speech signal at the phone level in typical prosody detection tasks have not been considered in this paper. We concentrate only on the energy and pitch contour that can be robustly obtained from the speech signal. However, our framework is readily amenable to the addition of new features. We provide discussions on the prominence and phrase structure detection presented in Sections VII and VIII below.

A. Prominence Prediction

The baseline experiment with lexical stress obtained from a pronunciation lexicon for prediction of pitch accent yields substantially higher accuracy than chance. This could be particularly useful in resource-limited languages where prosody labels are usually not available but one has access to a reasonable lexicon with lexical stress markers. Off-the-shelf speech synthesizers like Festival and AT&T speech synthesizer have utilities that perform reasonably well in pitch accent and boundary tone prediction. The AT&T speech synthesizer performs

better than Festival in pitch accent prediction while the latter performs better in boundary tone prediction. This can be attributed to better rules in the AT&T synthesizer for pitch accent prediction. Boundary tones are usually highly correlated with punctuation and Festival seems to capture this well. However, both these synthesizers generate a high degree of false alarms.

The maximum entropy model syntactic–prosodic proposed in Section VII-B1 outperforms previously reported results on pitch accent and boundary tone classification. Much of the gain comes from the strength of the maximum entropy modeling in capturing the uncertainty in the classification task. Considering the inter-annotator agreement for ToBI labels is only about 81 % for pitch accents and 93% for boundary tones, the maximum entropy framework is able to capture the uncertainty present in manual annotation. The supertag feature offers additional discriminative information over the part-of-speech tags (also demonstrated by Rainbow and Hirschberg [59]).

The maximum entropy acoustic–prosodic model discussed in Section VII-B2 performs well in isolation compared to the traditional HMM acoustic–prosodic model. This is a simple method, and the quantization resolution can be adjusted based on the amount of data available for training. However, the model performs with slightly lower accuracy when combined with the syntactic features compared to the combined maxent syntactic–prosodic and HMM acoustic–prosodic model. We conjecture that the generalization provided by the acoustic HMM model is complementary to that provided by the maximum entropy acoustic model, resulting in slightly better accuracy when combined with the maxent syntactic model compared the maxent acoustic–syntactic model. We attribute this behavior to better smoothing offered by the HMM compared to the maxent acoustic model. We also expect this slight difference would not be noticeable with a larger data set.

The weighted maximum entropy syntactic–prosodic model and HMM acoustic–prosodic model performs the best in pitch accent and boundary tone classification. The classification accuracies are comparable to the inter-annotator agreement for the ToBI labels. Our HMM acoustic–prosodic model is a generative model and does not assume the knowledge of word boundaries in predicting the prosodic labels as in previous approaches [2], [15], [20]. This makes it possible to have true parallel prosody prediction during speech recognition. However, the incorporation of word boundary knowledge, when available, can aid in improved detection accuracies [63]. This is also true in the case of our maxent acoustic–prosodic model that assumes word segmentation information. The weighted approach also offers flexibility in prosody labeling for either speech synthesis or speech recognition. While the syntactic–prosodic model would be more discriminative for speech synthesis, the acoustic–prosodic model is more appropriate for speech recognition.

B. Phrase Structure Prediction

The baseline results from Festival speech synthesizer are relatively modest for the break index prediction and only slightly better than chance. The break index prediction module in the synthesizer is mainly based on punctuation and parts-of-speech tag information and hence does not provide a rich set of discriminative features. The accuracies reported on the BU corpus are substantially higher compared to chance than those reported on the BDC corpus. We found that the distribution of break indices was highly skewed in the BDC corpus, and the corpus also does not contain any punctuation markers. Our proposed maximum entropy break index labeling with lexical and syntactic information alone achieves 83.95% and 87.18% accuracy on the BU and BDC corpora. The syntactic model can be used in text-to-speech synthesis and sentence disambiguation (for parsing) applications. We also envision the use of prosodic breaks in speech translation by aiding in the construction of improved phrase translation tables.

X. Summary, Conclusions, and Future Work

In this paper, we described a maximum entropy discriminative modeling framework for automatic prosody labeling. We applied the proposed scheme to both prominence and phrase structure detection within the ToBI annotation scheme. The proposed maximum entropy syntactic–prosodic model alone resulted in pitch accent and boundary tone accuracies of 85.2% and 91.5% on training and test sets identical to [27]. As far as we know, these are the best results on the BU and BDC corpus using syntactic information alone and a train-test split that does not contain the same speakers. We have also demonstrated the significance of our approach by setting reasonable baseline from out-of-the-box speech synthesizers and by comparing our results with prior work. Our combined maximum entropy syntactic–prosodic model and HMM acoustic–prosodic model performs the best with pitch accent and boundary tone labeling accuracies of 86.0% and 93.1%, respectively. The results of collectively using both syntax and acoustic within the maximum entropy framework are not far behind at 85.2% and 92%, respectively. The break index detection with the proposed scheme is also promising with detection accuracies ranging from 84% to 87%. The inter-annotator agreement for pitch accent, boundary tone and break index labeling on the BU corpus [30] are 81%–84%, 93%, and 95%, respectively. The accuracies of 80–86%, 90–93.1%, and 84–87% achieved with the proposed framework for the three prosody detection tasks are comparable to the inter-labeler agreements. In summary, the experiments of this paper demonstrate the strength of using a maximum entropy discriminative model for prosody prediction. Our framework is also suitable for integration into state-of-the-art speech applications.

The supertag features in this work were used as categorical labels. The tags can be unfolded, and the syntactic dependencies and structural relationship between the nodes of the supertags can be exploited further as demonstrated in [59]. We plan to use these more refined features in future work. As a continuation of our work, we have integrated our prosody labeler in a dialog act tagging scenario, and we have been able to achieve modest improvements [64]. We are also working on incorporating our automatic prosody labeler in a speech-to-speech translation framework. Typically, state-of-the-art speech translation systems have a source language recognizer followed by a machine translation system. The translated text is then synthesized in the target language with prosody predicted from text. In this process, some of the critical prosodic information present in the source data is lost during translation. With reliable prosody labeling in the source language, one can transfer the prosody to the target language (this is feasible for languages with phrase level correspondence). The prosody labels by themselves may or may not improve the translation accuracy but they provide a framework where one can obtain prosody labels in the target language from the speech signal rather than depending only on a lexical prosody prediction module in the target language.

References

1. Lehiste, I. Suprasegmentals. Cambridge, MA: MIT Press; 1970.
2. Wightman CW, Ostendorf M. Automatic labeling of prosodic patterns. IEEE Trans Speech Audio Process Oct;1994 2(4):469–481.
3. Koehn P, Abney S, Hirschberg J, Collins M. Improving into-national phrasing with syntactic information. Proc ICASSP 2000:1289–1290.
4. Steedman M. Information structure and the syntax-phonology interface. Linguist Inquiry 2000;31(4): 649–689.
5. Liberman M, Prince A. On stress and linguistic rhythm. Linguist Inquiry 1977;8(2):249–336.
6. Silverman K, Beckman M, Pitrelli J, Ostendorf M, Wightman C, Price P, Pierrehumbert J, Hirschberg J. ToBI: A standard for labeling english prosody. Proc ICSLP 1992:867–870.
7. Taylor P. The TILT intonation model. Proc ICSLP 1998;4:1383–1386.

8. Fujisaki H, Hirose K. Modelling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation. Proc 13th Int Congr Linguists 1982:57–70.
9. Grabe E, Nolan F, Farrar K. IViE—A comparative transcription system for international variation in english. Proc ICSLP, Sydney, Australia 1998:1259–1262.
10. Hirst DJ, Ide N, Vronis J. Coding fundamental frequency patterns for multilingual synthesis with INTSINT in the MULTEXT project. Proc 2nd ESCA/IEEE Workshop Speech Synth Sep;1994 :77–81.
11. Shriberg EE, Bates RA, Stolcke A. A prosody-only decision-tree model for disfluency detection. Proc Eurospeech'97, Rhodes, Greece 1997:2383–2386.
12. Liu Y, Shriberg E, Stolcke A, Hillard D, Ostendorf M, Harper M. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. IEEE Trans Audio, Speech, Lang Process Sep;2006 14(5):1526–1540.
13. Kahn JG, Lease M, Charniak E, Johnson M, Ostendorf M. Effective use of prosody in parsing conversational speech. Proc HLT/EMNLP 2005:233–240.
14. Shriberg E, Bates R, Stolcke A, Taylor P, Jurafsky D, Ries K, Coccaro N, Martin R, Meteer M, Van Ess-Dykema C. Can prosody aid the automatic classification of dialog acts in conversational speech? Lang Speech 1998;41(3–4):439–487.
15. Hirschberg J. Pitch accent in context: Predicting intonational prominence from text. Artif Intell 1993;63(1–2):305–340.
16. Shimei P, McKeown K. Word informativeness and automatic pitch accent modeling. Proc EMNLP/VLC, College Park, MD 1999:148–157.
17. Sun X. Pitch accent prediction using ensemble machine learning. Proc ICSLP 2002:561–564.
18. Conkie A, Riccardi G, Rose RC. Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events. Proc Eurospeech, Budapest, Hungary 1999:523–526.
19. Ananthakrishnan, S.; Narayanan, S. Proc ICASSP. Philadelphia, PA: Mar. 2005 An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model; p. 269-272.
20. Hasegawa-Johnson M, Chen K, Cole J, Borys S, Kim SS, Cohen A, Zhang T, Choi JY, Kim H, Yoon TJ, Chavara S. Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus. Speech Commun 2005;46:418–439.
21. Gregory M, Altun Y. Using conditional random fields to predict pitch accent in conversational speech. Proc 42nd Annu Meeting Assoc Comput Linguist (ACL) 2004:677–704.
22. Bulyko I, Ostendorf M. Joint prosody prediction and unit selection for concatenative speech synthesis. Proc ICASSP 2001:781–784.
23. Ma X, Zhang W, Shi Q, Zhu W, Shen L. Automatic prosody labeling using both text and acoustic information. Proc ICASSP Apr;2003 1:516–519.
24. Taylor P, King S, Isard S, Wright H. Intonation and dialogue context as constraints for speech recognition. Lang Speech 2000;41(34):493–512. [PubMed: 10746367]
25. Stolcke A, Ries K, Coccaro N, Shriberg E, Bates R, Jurafsky D, Taylor P, Martin R, Van Ess-Dykema C, Meteer M. Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput Linguist Sep;2000 26:339–373.
26. Hirschberg J, Prieto P. Training intonational phrasing rules automatically for English and Spanish text-to-speech. Speech Commun 1996;18(3):281–290.
27. Chen K, Hasegawa-Johnson M, Cohen A. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model. Proc ICASSP 2004:I-509–I-512.
28. Beckman ME, Diaz-Campos M, McGory JT, Morgan TA. Intonation across Spanish, in the tones and break indices framework. Probus 14:9–36.
29. Beckman ME, Pierrehumbert JB. International structure in Japanese and English. Phonology Yearbook 3:255–309.
30. Ostendorf, M.; Price, PJ.; Shattuck-Hufnagel, S. The Boston University Radio News Corpus. Boston Univ; Boston, MA: Mar. 1995 Tech. Rep. ECS-95-001

31. Hirschberg J, Nakatani C. A prosodic analysis of discourse segments in direction-giving monologues. Proc 34th Conf Assoc Computat Linguist 1996:286–293.
32. Price PJ, Ostendorf M, Shattuck-Hufnagel S, Fong C. The use of prosody in syntactic disambiguation. J Acoust Soc Amer 1991;90(6):2956–2970. [PubMed: 1787237]
33. Wightman CW, Shattuck-Hufnagel S, Ostendorf M, Price PJ. Segmental durations in the vicinity of prosodic phrase boundaries. J Acoust Soc Amer 1992;91(3):1707–1717. [PubMed: 1564206]
34. Ross K, Ostendorf M. Prediction of abstract prosodic labels for speech synthesis. Comput Speech Lang Oct;1996 10:155–185.
35. Nenkova A, Brenier J, Kothari A, Calhoun S, Whitton L, Beaver D, Jurafsky D. To memorize or to predict: Prominence labeling in conversational speech. Proc NAACL-HLT'07 2007:9–16.
36. Harper M, Dorr B, Roark B, Hale J, Shafran Z, Liu Y, Lease M, Snover M, Young L, Stewart R, Krasnyanskaya A. Parsing speech and structural event detection. Proc JHU Summer Workshop 2005:1–116.Tech. Rep
37. Wang MQ, Hirschberg J. Automatic classification of international phrase boundaries. Comput Speech Lang 6:175–196. 1992.
38. Pitrelli JF, Beckman ME, Hirschberg J. Evaluation of prosodic transcription labeling reliability in the ToBI framework. Proc ICSLP 1994:123–126.
39. Nöth E, Batliner A, KieBling A, Kompe R, Niemann H. VERBMOBIL: The use of prosody in the linguistic components of a speech understanding system. IEEE Trans Speech Audio Process Sep; 2000 8(5):519–532.
40. Yoon, TJ. Proc ICSA Int Conf Speech Prosody. Dresden; Germany: 2006. Predicting prosodic boundaries using linguistic features. CD-ROM
41. Ostendorf M, Shafran I, Shattuck-Hufnagel S, Carmichael L, Byrne W. A prosodically labeled database of spontaneous speech. Proc ISCA Workshop Prosody in Speech Recognition and Understanding 2001:119–121.
42. Black AW, Taylor P. Assigning phrase breaks from part-of-speech sequences. Proc Eurospeech, Rhodes, Greece 1997;2:995–998.
43. Sun, X.; Applebaum, TH. Intonational Phrase Break Prediction Using Decision Tree and N-Gram Model. Vol. 1. Aalborg; Denmark: 2001. p. 537-540.
44. Veilleux NM, Ostendorf M, Wightman CW. Parse scoring with prosodic information. Proc Int Conf Spoken Lang Process 1992:1605–1608.
45. Veilleux, NM.; Ostendorf, M. HLT'93: Proc Workshop Human Lang Technol. Morristown, NJ: 1993. Prosody/parse scoring and its application in atis; p. 335-340.Assoc. Comput. Linguist
46. Rangarajan Sridhar VK, Bangalore S, Narayanan S. Acoustic–syntactic maximum entropy model for automatic prosody labeling. Proc IEEE/ACL Spoken Lang Technol, Aruba Dec;2006 :74–77.
47. Rangarajan Sridhar VK, Bangalore S, Narayanan S. Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. Proc NAACL-HLT 2007:1–8.
48. Berger A, Pietra SD, Pietra VD. A maximum entropy approach to natural language processing. Comput Linguist 1996;22(1):39–71.
49. Dudik, M.; Phillips, S.; Schapire, RE. Proc COLT. Banff, Canada: Springer-Verlag; 2004. Performance guarantees for regularized maximum entropy density estimation; p. 472-486.
50. Haffner P. Scaling large margin classifiers for spoken language understanding. Speech Commun 2006;48(IV):239–261.
51. Bangalore S, Joshi AK. Supertagging: An approach to almost parsing. Computat Linguist 1999;25 (2):237–265.
52. Joshi, A.; Schabes, Y. Tree-adjoining grammars. In: Salomaa, A.; Rozenberg, G., editors. Handbook of Formal Lanaguages and Automata. Berlin: Springer-Verlag; 1996.
53. "A Lexicalized Tree-Adjoining Grammar for English," Univ. of Pennsylvania, Philadelphia, Tech. Rep., 2001 [Online]. Available: http://www.cis.upenn.edu/xtag/gramrelease.html, XTAG
54. Chen J, Vijay-Shanker K. Automated extraction of tags from the penn treebank. Proc 6th Int Workshop Parsing Technologies, Trento, Italy 2000:73–89.
55. Xia F, Palmer M, Joshi A. A uniform method of grammar extraction and its applications. Proc Empirical Methods in Natural Lang Process 2000:53–62.

56. Bangalore S, Emami A, Haffner P. Factoring Global Inference by Enriching Local Representations. AT&T Labs-Research, Tech Rep. 2005
57. Bangalore S, Haffner P. Classification of large label sets. Proc Snowbird Learning Workshop. 2005CD-ROM
58. Hasegawa-Johnson M, Cole J, Shih C, Chen K, Cohen A, Chavarria S, Kim H, Yoon T, Borys S, Choi JY. Speech recognition models of the interdependence among syntax, prosody, and segmental acoustics. Proc HLT/NAACL, Workshop Higher-Level Knowledge in Automatic Speech Recognition and Understanding May;2004 :56–63.
59. Hirschberg J, Rambow O. Learning prosodic features using a tree representation. Proc Eurospeech, Aalborg, Denmark 2001:1175–1180.
60. Shriberg E, Ferrer L, Kajarekar S, Venkataraman A, Stolcke A. Modeling prosodic feature sequences for speaker recognition. Speech Commun 2005;46:455–472.
61. "AT&T Natural Voices Speech Synthesizer." [Online]. Available: http://www.naturalvoices.att.com.
62. Black, AW.; Taylor, P.; Caley, R. The Festival Speech Synthesis System. 1998. [Online]. Available: http://festvox.org/festival
63. Chen K, Hasegawa-Johnson M, Cohen A, Borys S, Kim SS, Cole J, Choi JY. Prosody dependent speech recognition on radio news corpus of American English. IEEE Trans Audio, Speech, Lang Process Jan;2006 14(1):232–245.
64. Rangarajan Sridhar VK, Bangalore S, Narayanan S. Exploiting prosodic features for dialog act tagging in a discriminative modeling framework. Proc ICSLP. 2007

Biographies

Vivek Kumar Rangarajan Sridhar (S'06) received the B.E. (honors) degree in electrical and electronics engineering from the Birla Institute of Technology and Science, Pilani, India, in 2002, and the M.S. degree in electrical engineering from University of Southern California, Los Angeles, in 2004, where he is currently pursuing the Ph.D. degree in electrical engineering. The focus of his Ph.D. dissertation is on the enrichment of spoken language processing using suprasegmental information.

His general research interests include speech recognition, spoken language understanding, spontaneous speech processing, disfluency detection, speech-to-speech translation, and articulatory modeling.

Mr. Rangarajan Sridhar is a recipient of the Best Teaching Assistant award from the USC Electrical Engineering Department (2003–2004).

Srinivas Bangalore received the Ph.D. degree in computer science from the University of Pennsylvania, Philadelphia, in 1997.

He is a Principal Technical Staff Member in the Voice and IP Services unit of AT&T Labs-Research and an Adjunct Associate Professor at Columbia University, New York. His thesis on was awarded the Since the Ph.D. degree, he has lead innovative research efforts at AT&T Labs-Research on topics ranging from machine translation, multimodal language processing, language generation, dialog modeling and finite-state language processing techniques. He has authored over 100 peer-reviewed journal and conference papers and has taught tutorials at ACL and ICASSP conferences. He is an active member in the technical review committees for conferences and journals related to speech and natural language processing.

Dr. Bangalore received the Morris and Dorothy Rubinoff award for outstanding dissertation for his thesis on “Supertagging” that has resulted in or could lead to innovative applications of computer technology.

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 1995–2000. 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 260 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, MMSP’06, and MMSP’07. 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.

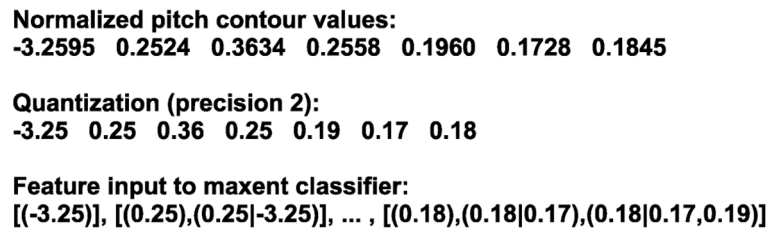


Fig. 1.
Illustration of the quantized feature input to the maxent classifier. "|" denotes feature input conditioned on preceding values in the acoustic–prosodic sequence.

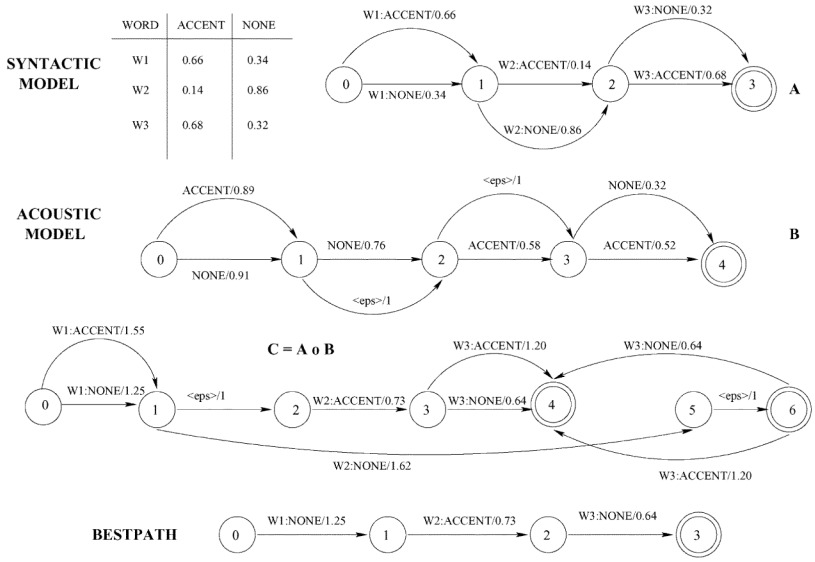


Fig. 2.

Illustration of the FST composition of the syntactic and acoustic lattices and resulting best path selection. The syntactic–prosodic maxent model produces the syntactic lattice and the HMM acoustic–prosodic model produces the acoustic lattice.

TABLE I

ToBI Label Mapping Used in Experiments. The Decomposition of Labels is Illustrated for Pitch Accents, Phrasal Tones, and Break Indices

ToBI Labels	Intermediate Mapping	Coarse Mapping
H* L+H*	High	**accent**
!H*, H+!H* L+!H*,L*+!H	Downstepped	
L* L*+H	Low	
,?,X&?	Unresolved	
L-L%,!H-L%,H-L% H-H% L-H% %?,X%?,%H	Final Boundary tone	**btone**
L-,H-,!H--X?,-?	Intermediate Phrase (IP) boundary	
<,>, no label	**none**	**none**
0	0	**NB**
1,l-,1p	1	
2,2-,2p	2	
3,3-,3p	3	**B**
4,4-	4	

TABLE II

Summary of Previous Work on Pitch Accent and Boundary Tone Detection (Coarse Mapping). Level Denotes the Orthographic Level (Word or Syllable) at Which the Experiments Were Performed. The Results of Hasegawa-Johnson *et al*. and Our Work Are Directly Comparable as the Experiments are Performed on Identical Dataset

Authors	Algorithm	Corpus	Level	Accuracy (%)	
				Pitch accent	Boundary tone
Wightman and Ostendorf [2]	HMM/CART	BU	syllable	83.0	77.0
Ross and Ostendorf [34]	HMM/CART	BU	syllable	87.7	66.9
Ananthakrishnan et al. [19]	Coupled HMM	BU	syllable	75.0	88.0
Gregory and Alton [21]	Conditional random fields	Switchboard	word	76.4	-
Nenkova et al. [35]	Decision Tree	Switchboard	word	76.6	-
Harper et al. (JHU Workshop) [36]	Decision Trees/Random Forest	Switchboard	word	80.4	-
Hirschberg [15]	CART	BU	word	82.4	-
Wang and Hirschberg [37]	CART	ATIS	word	-	90.0
Ananthakrishnan et al. [19]	Coupled HMM	BU	word	79.5	82.1
Hasegawa Johnson et al. [20]	Neural networks/GMM	BU	word	84.2	93.0
Proposed work	Maximum entropy model	BU and BDC	word	86.0	93.1

TABLE III

Summary of Previous Work on Break Index Detection (Coarse Mapping). Detection Is Performed at Word Level for All Experiments

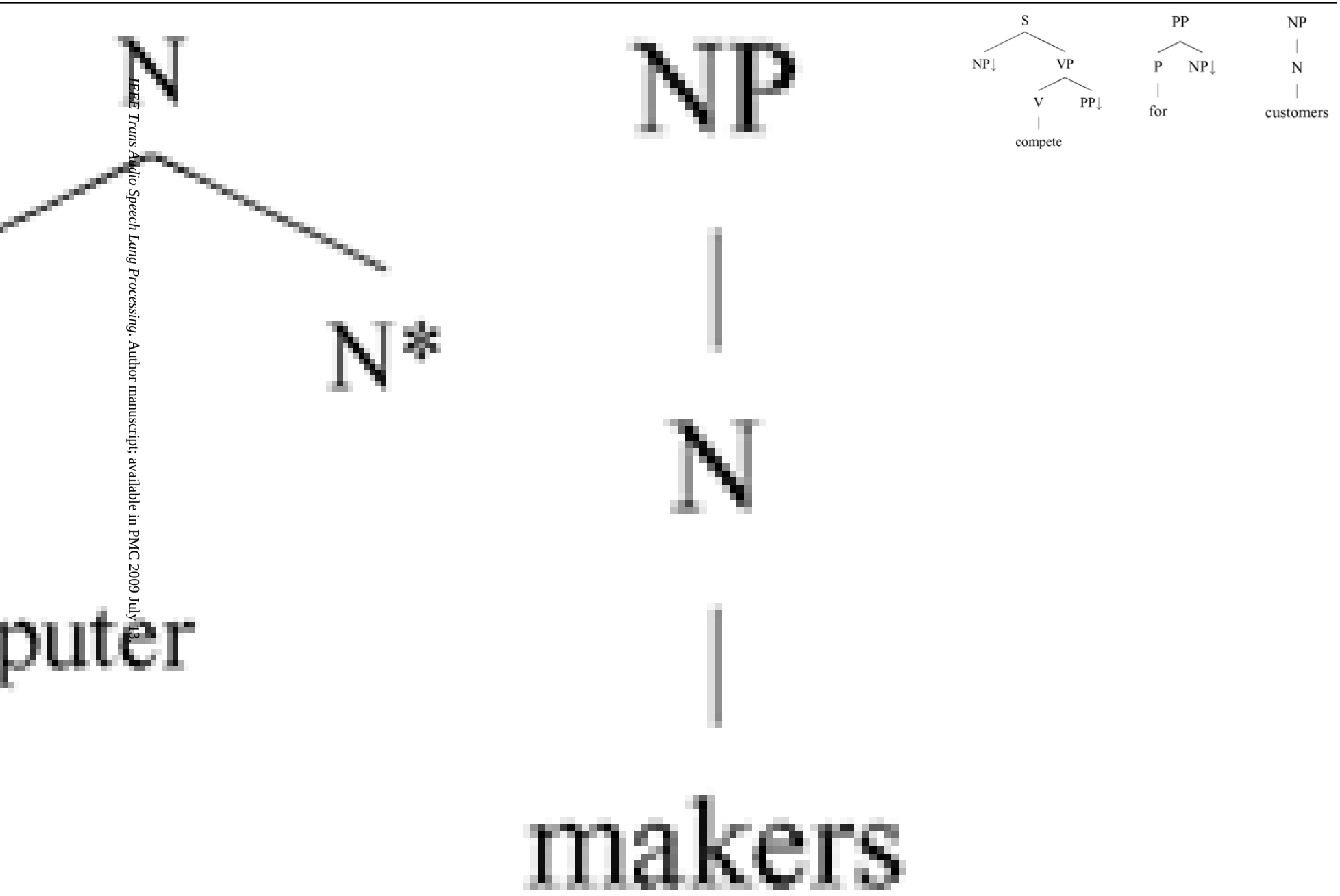
Authors	Algorithm	Corpus	Accuracy (%) Break index
Wightman and Ostendorf [2]	HMM/CART	BU	84.0
Ostendorf and Veilleux [45]	HMM/CART	ATIS	70.0
Wang and Hirschberg [37]	CART	ATTS	81.7
Taylor and Black [42]	HMM	Spoken English corpus	79.2
Sun and Applebaum [43]	CART	BU	85.2
Harper et al. (JHU Workshop) [36]	Decision Trees/Random Forest	Switchboard	83.2
Proposed work	Maximum entropy model	BU and BDC	84.0–87.5

TABLE IV

Lexical, Syntactic, and Acoustic Features Used in the Experiments. The Acoustic Features Were Obtained Over 10-ms Frame Intervals

Category	**Features used**
Lexical features	Word identity (3 previous and next words)
Syntactic features	POS tags (3 previous and next words) Supertags (3 previous and next words) function/content word distinction (3 previous and next words)
Acoustic features	Speaker normalized f0 contour (+delta+acceleration) Speaker normalized energy contour (+delta+acceleration)

TABLE V



customers for compete makers minicomputer

TABLE VI

Statistics of Boston University Radio News and Boston Directions Corpora Used in Experiments

	BU				BDC			
Corpus statistics	**f2b**	**f1a**	**m1b**	**m2b**	**h1**	**h2**	**h3**	**h4**
# Utterances	165	69	72	51	10	9	9	9
# words (w/o punc)	12608	3681	5058	3608	2234	4127	1456	3008
# pitch accents	6874	2099	2706	2016	1006	1573	678	1333
# boundary tones (w IP)	3916	1059	1282	1023	498	727	361	333
# boundary tones (w/o IP)	2793	684	771	652	308	428	245	216
# breaks (level 3 & above)	3710	1034	11721	1016	434	747	197	542

TABLE VII

Baseline Classification Results of Pitch Accents and Boundary Tones (in %) Using Festival and AT&T Natural Voices Speech Synthesizer

Corpus	Speaker Set	Prediction Module	Accuracy	
			Pitch accent	Boundary tone
BU	Entire Set	**Chance**	54.33	81.14
		Lexical stress	72.64	-
		AT&T Natural Voices	81.51	89.10
		Festival	69.55	89.54
	Hasegawa-Johnson et al. set	**Chance**	56.53	82.88
		Lexical stress	74.10	-
		AT&T Natural Voices	81.73	89.67
		Festival	68.65	90.21
BDC	Entire Set	**Chance**	57.60	88.90
		Lexical stress	67.42	-
		AT&T Natural Voices	68.49	84.90
		Festival	64.94	85.17

TABLE VIII

Classification Results (%) of Pitch Accents and Boundary Tones for Different Syntactic Representations. Classifiers With Cardinality V = 2 Learned Either Accent or Btone Classification, Classifiers With Cardinality V = 4 Classified Accent, and Btone Simultaneously. The Variable (*k*) Controlling the Length of the Local Context Was Set to *k* = 3

Corpus	Speaker Set	Syntactic features	V=2		V=4	
			Pitch accent	Boundary tone	Pitch accent	Boundary tone
BU	Entire Set	correct POS tags	84.75	91.39	84.60	91.34
		POS tags	83.71	90.52	83.50	90.36
		POS + supertags	84.59	91.34	84.48	91.22
	Hasegawa-Johnson et al. set	correct POS tags	85.22	91.33	85.03	91.29
		POS tags	83.91	90.14	83.72	90.04
		POS + supertags	84.95	91.21	84.85	91.24
BDC	Entire Set	POS + supertags	79.81	90.28	79.57	89.76

TABLE IX

Classification Results of Pitch Accents and Boundary Tones (in %) With Acoustics Only, Syntax Only, and Acoustics+Syntax Using Both Our Models. The Syntax-Based Results From Our Maximum Entropy Syntactic–Prosodic Classifier Are Presented Again to View the Results Cohesively. In the Table, A=Acoustics, S=Syntax

			Pitch accent			Boundary tone		
Corpus	**Speaker Set**	**Model**	**A**	**S**	**A+S**	**A**	**S**	**A+S**
BU	Entire Set	Maxent acoustic model	80.09	84.60	84.63	84.10	91.36	91.76
		HMM acoustic model	70.58	84.60	85.13	71.28	91.36	92.91
	Hasegawa-Johnson et al. set	Maxent acoustic model	80.12	84.95	85.16	82.70	91.54	91.94
		HMM acoustic model	71.42	84.95	86.01	73.43	91.54	93.09
BDC	Entire Set	Maxent acoustic model	74.51	79.81	79.97	83.53	90.28	90.49
		HMM acoustic model	68.57	79.81	80.01	74.28	90.28	90.58

TABLE X

Classification Results of Break Indices (in %) With Syntax Only, Acoustics Only, and Acoustics + Syntax Using the Maximum Entropy Classifier. In the Table, A=Acoustics, S=Syntax

	ToBI break indices					B/NB				
Corpus	**Chance**	**Festival**	**A**	**S**	**A+S**	**Chance**	**Festival**	**A**	**S**	**A+S**
BU	61.25	64.22	64.73	72.32	72.90	71.91	77.58	73.98	83.95	84.01
BDC	60.01	66.56	58.95	69.25	69.81	82.26	82.31	75.94	87.18	87.58