Automatic Labeling of Prosodic Patterns

Colin W. Wightman, *Member, IEEE*, and Mari Ostendorf, *Member, IEEE*

***Abstract*—This paper describes a general algorithm for labeling prosodic patterns in speech, which provides a mechanism for mapping sequences of observations (vectors of acoustic correlates) to prosodic labels using decision trees and a Markov sequence model. Important and novel features of the approach are that it allows many dissimilar correlates to be treated in a unified manner to provide more robust labeling, and that it is designed to be a post-word-recognition processing step. Application of the algorithm is illustrated with experimental results for labeling prosodic phrasing and phrasal prominence in two corpora of professionally read speech. The labels produced by the automatic algorithm exhibit agreement with hand-labeled prominence and phrasing that is close to the agreement between different human labelers.**

I. Introduction

PROSODY refers to aspects of the speech signal other than the actual words spoken, such as timing and fundamental frequency (f0) patterns. Since prosodic information usually cannot be associated with a single phone-sized segment, it is often referred to as suprasegmental information. Speakers use prosody to convey emphasis, intent, attitude, and to provide cues to aid the listener in the interpretation of their speech. Because English, like all natural languages, is highly ambiguous, the prosodic cues can be an important part of selecting the correct interpretation of an utterance.

The problem of determining the speaker's intent from a (possibly ambiguous) utterance is no longer one unique to human listeners; it is one that is increasingly faced by designers of machines to understand speech. Early systems could avoid this problem by restricting the user to a small vocabulary or constraining the application domain so that ambiguous commands could not occur. As systems for more complex tasks and with more "natural" speech interfaces are developed, however, prosodic information becomes important as a source of potential improvements in accuracy and speed. Of course, speech understanding systems will be able to utilize automatically detected prosodic cues only to the extent that the relationships between prosodic markings and intended meaning are understood. Exploration of these relationships requires the careful linguistic analysis of large corpora of speech in which the prosodic structure has been annotated. Thus, a second application for automatic labeling of prosodic patterns is transcription of large corpora. In this paper, we report the development of a method for automatically labeling prosodic information in continuous speech. Enabled by the continuing evolution of linguistic thought and convergence towards a consistent theory of the phonology of prosody, as well as advances in statistical pattern recognition, the algorithm proposed represents a critical step towards the goal of using prosody to aid automatic speech understanding.

The main prosodic attributes with which we are concerned are **phrasing** and **prominence** or, borrowing from the terminology of Cutler [1], "segmentation" and "salience." Prosodic phrasing refers to the perceived groupings of words in an utterance, and phrasal prominence refers to the greater perceived strength or emphasis of some syllables in a phrase. A more detailed representation of prosody would also include a characterization of intonation in terms of pitch accent types (as prominence markers) and boundary tone types (as prosodic phrase markers), where the different tone types have different discourse roles (see, e.g., [2]). Thus, the prosodic units of interest in spoken language systems include: prosodic phrase structure, which is related to syntax; phrasal prominence, which is mainly related to semantic and discourse structure but also to phrase structure; and intonation markers, for their indication of phrase boundaries and syllable prominence as well as their discourse function.

The principal acoustic cues which mark prominence and prosodic phrase boundaries have been known to linguists for some time, but only recently has linguistic work focused on developing more consistent taxonomies of prosodic units and an understanding of the relationships between these units. Indeed, the lack of a general consensus in these areas may be one reason why prosody has been under-utilized in spoken language systems and why the work on automatic detection of prosodic cues is difficult to assess. In addition, the recent promising research on prosodic pattern recognition, summarized below, is somewhat fragmented by the different approaches taken by those interested in only one aspect of prosody.

Initial attempts at automatically recognizing prosody [3] were not sufficiently accurate and were limited in their usefulness, in part because neither recognition technology nor phonological theories of prosody were as advanced as they are today. More recent prosodic pattern detection work has addressed a variety of problems—phrase structure recognition, tune recognition, and prominence or stress detection—with a variety of different algorithms. These algorithms will be reviewed briefly to provide a perspective on our approach. All share the same common problem of not utilizing the full range of acoustic cues available.

Manuscript received February 2, 1993; revised November 2, 1993. This work was jointly funded by NSF and DARPA under NSF Grant number IRI-8905249. The associate editor coordinating the review of this paper and approving it for publication was Dr. Amro El-Jaroudi.

C. W. Wightman is with the Electrical Engineering Department, New Mexico Institute of Mining and Technology, Socorro, NM 87801 USA.

M. Ostendorf is with the Department of Electrical, Computer, and Systems Engineering, Boston University, Boston, MA 02215 USA.

IEEE Log Number 9403980.

Previous approaches to recognition of the prosodic phrase structure have basically all been based on analyses of F0 contours. Huber [4] developed an algorithm for detecting phrases using linear regression fits of baselines and toplines, assuming that F0 contour declination describes the overall downward trend of F0 over a phrase and that a jump upwards to start a new declination line signals a phrasal boundary. Komatsu *et al.* [5] developed a method of using a F0-based algorithm similar to Lea's [3] to obtain a tree-like representation of the prosodic phrasal structure. Their work is focused principally on Japanese, like that of Shimodaira and Kimura [6], who use a dynamic programming search to find the best phrase segmentation given a set of F0 phrase templates found via clustering. A disadvantage of all these approaches is that they ignore preboundary lengthening cues, which are among the most powerful (as discussed in the next section).

Automatic recognition of intonation markers encompasses work in boundary tone classification (e.g., statement vs. question recognition) as well as classification of different pitch accent types. Methods for detecting yes/no questions from F0 contours have achieved high accuracy using rule-based techniques [7], [8], though these algorithms are limited to this simple classification problem. HMM-based algorithms provide a more general approach for classifying tune types, which could be used for recognizing boundary tones or pitch accent types. HMM's have been used extensively for several related problems, including contour classification [9], [10], boundary tone spotting [11], and Chinese tone classification (e.g., [12]). Again, these methods share the disadvantage of ignoring duration cues, and are therefore more suitable for intonation pattern classification than prominence or phrase detection.

Stress, accent, and/or emphasis detection all deal with detection of the relative prominence of a syllable. A discussion of the different detection methods is difficult, because "stress" has been used ambiguously to refer to several types of prominence, including strong vs. weak syllable distinctions as indicated by lexical stress marking, as well as phrasal prominence as indicated by a pitch accent.[1] This confusion may account for the wide variety of algorithms proposed, since different levels of prominence are cued differently in the acoustic signal. Waibel [7] investigated detection of stress using Gaussian classifiers with syllable-level F0, duration, and energy features, with the goal of reducing lexical ambiguity for word recognition. More recently, lexical stress has been modeled directly in recognition systems by using separate models for stressed and unstressed vowels. At the phrasal level, prominence recognition algorithms include hidden Markov model spotting of emphasis using frame-based energy and duration features [13], and linear discriminant functions based on syllable-level features [14], [15], or frame-level features [16]. The only detection rates quoted are those of Campbell [16], who achieves 72–92% correct detection and 4–7.5% false detections using only energy and duration cues on a speaker-dependent corpus, where higher rates are obtained for more exaggerated speaking styles.

[1] This is a problem in the linguistics literature, as well as in the automatic detection literature.

In addition to using only one or two of the acoustic correlates available, most of the previous methods for detecting prosodic features attempt to work directly from the speech signal itself without utilizing a speech recognizer, and generally use an approach tailored to the specific type of prosodic information being recognized. Our approach to automatic labeling is unusual in that we use the results of speech recognition to obtain many of the features which are analyzed. The postrecognition approach has the advantages that the word recognizer produces a phonetic segmentation, which provides a means of quantitatively analyzing phrase-final lengthening and other durational cues, and that the prosodic structures which are detected are aligned with the hypothesized word sequence. In addition, because the structure of our method is quite general, it is applicable to all of the types of prosodic information discussed above. Further, we can investigate the performance of the prosodic labeling algorithm without confounding it with recognition performance by constraining the recognizer search to the utterance transcription.

In the next section, we will discuss the prosodic structure of spoken English, both in terms of the phonological units which have been posited by theoretical linguists, and in terms of the acoustic correlates which have been reported to demarcate these units. Section III then describes the prosodic labeling algorithm. In Section IV, we describe two corpora of professionally read speech and report the results of several labeling experiments using the proposed algorithm. We then conclude in Section V with a discussion of the significant results and possible extensions and applications of the algorithm.

II. Representation of Prosodic Patterns

In this paper, we focus on the detection of boundaries between prosodic constituents (phrasing) and phrasal prominence, though the methods described here are extensible to more detailed prosodic representations. We choose this focus primarily because these cues are simplest to hand label and will likely offer the greatest source of knowledge for speech understanding: prosodic phrasing plays an important role in the disambiguation of sentences and prominences are frequently used to mark discourse structure. More detailed pitch accent labels are likely to convey more subtle information associated with speaker attitude. The following section reflects this emphasis in its discussion of abstract prosodic pattern representation and acoustic correlates.

A. Prosodic Phrasing and Prominence

Researchers have noticed that fluent spoken English is not produced in a smooth, unvarying stream. Rather, speech has perceptible breaks and clumps, as well as relatively stronger and weaker syllables. For example, some words seem to be more closely grouped with adjacent words than others; we call these groups prosodic phrases. These phrases can be grouped together to form larger phrases, which may be grouped to form sentences, paragraphs, and complete discourses. In addition, some words or syllables are perceived as stronger, sometimes because the speaker has chosen to emphasize a word but strong

syllables also occur in more typical utterances that contain no emphatic or contrastive words. These observations raise questions about how many such phrase and prominence "units" there are and how they are best defined.

The domain of linguistic theory most appropriate to address these questions is phonology, traditionally defined as the study of sound units and their structural interrelationships in spoken language. Work in this field has been reported for at least seven centuries (c.f. Jones [17]). However, until the late 1970's, if grammars addressed the question of intonational, rhythmic, and pausing patterns at all, it was in terms of the familiar constituents of syntactic surface structure trees. In addition, research on prominence was confounded by terminology that did not clearly distinguish between lexical and phrasal stress.

More recently, researchers have noted that the **prosodic phrase structure** is not isomorphic to the syntactic structure [18]–[20], and phonologists have begun to develop theoretical frameworks that include a separate hierarchy of prosodic constituents. Within the linguistic literature, numerous researchers have posited a number of prosodic phrasing constituents with hierarchical relationships such as *word, minor phrase, major phrase, utterance* (see [21] for a more complete review). However, while many phonologists are in agreement on the need for some types of prosodic constituents, there are still substantial differences in how they choose to define those constituents. Moreover, there are several types of constituents that have been suggested by some, but not widely adopted. Nonetheless, if we consider all the constituents that have been suggested, eliminating notational variants, we arrive at a superset (a set union of all the theories) of prosodic constituents.

The superset of prosodic constituents is represented in the perceptual labeling system developed by Price *et al.* in [22]. Under this system, the degree of perceived disjuncture between each pair of words is expressed by a break index between 0 and 6, which is assigned by human labelers. The labeling system is described briefly here; the reader is referred to [21] for a discussion of the mapping to prosodic constituents proposed in the literature. A 0 is assigned between two orthographic words between which there is obvious phonetic reduction (as in a deleted /h/ in *did he*). A break index of 1 is assigned as a default to unmarked word boundaries. A 2 is assigned to boundaries to denote a perceived grouping of words that is not intonationally marked. The boundaries labeled with 3 and 4 are "intermediate" and "intonational" phrase boundaries, respectively, as described in [23]. Perceived groupings of intonational phrases within a (typically long) sentence are marked with a 5, and sentence boundaries are labeled with a 6.[2] These break indices represent an extension of the set embodied in the ToBI prosodic transcription system, which only uses indices between 0 and 4 [24]. The difference is that ToBI treats the intonational phrase as the highest level constituent, whereas we perceived (and hence labeled) higher levels of phrasing. Whether these higher levels of phrasing are evidence of larger phonological constituents or reflections of discourse structure is still an unresolved issue. Nonetheless, we choose to use the larger set of break indices because it more accurately reflects our perceptions, and because the ToBI labels can readily be obtained by merging break indices of 4, 5, and 6.

[2] Although the "sentence" is not recognized as a prosodic constituent and "utterance" might be a more appropriate term, all the utterances used in this work comprise syntactically well-formed sentences.

Many phonologists represent a hierarchical metrical structure of relatively weaker and stronger syllables (e.g., [20] and [25]), under the hypothesis that syllable prominence can only be thought of in relational terms. In this hierarchy, **phrasal prominence** is used to refer to relatively stronger syllables above the level of the word. Theories differ as to whether pitch accent placement either influences or is influenced by the metrical structure of an utterance. For example, Bolinger [26] proposes that phrasal prominence is simply a matter of accent placement. In addition, the notion of relative levels of phrasal prominence is not well understood; although the metrical hierarchy theoretically implies an unrestricted number of levels of prominence, most studies describe only nuclear and nonnuclear prominence. We will avoid theoretical controversy by simply transcribing our corpus according to presence vs. absence of perceived phrasal prominence. However, recent experiments [27], [28] support the view that perception of prominence is influenced strongly by the placement of pitch accents, and we believe that the prominent syllables correspond to pitch accents.

Assuming that prominences are often (if not always) associated with pitch accents, then it is likely that these may be confused with the other type of intonation marker: the boundary tone. To help reduce confusion between pitch accents and unaccented phrase-final syllables, we chose to also model boundary tones. In addition, automatic detection of boundary tones will be useful for prosodic phrase recognition. The set of intonation markers used in this study are then simply "P" for prominent syllables, "s" for unmarked syllables, "BT" for a syllable marked with an intonational phrase (break levels 4–6) boundary tone, and "P-BT" for syllables marked with both a prominence and a boundary tone. These classes represent a subset of the labels embodied in the ToBI prosodic transcription system, which uses more detailed classifications of prominences and boundary tones and which in addition uses a boundary marker for intermediate prosodic phrases (level 3). This simple four-class labeling system allows us to avoid the cost of more detailed hand-labeling and labeling larger training sets, as well as avoiding controversies in linguistic theory. As a consequence, the detection system described here may suffer in performance from lumping together classes that are very different acoustically, such as the falling and rising F0 patterns for different types of boundary tones. However, the general method is extensible to more detailed prosodic labels.

B. Acoustic Correlates

While phonological research has investigated abstract prosodic structure, research in phonetics has analyzed the acoustic cues that mark these units. The different F0, duration, and energy correlates described in the literature are discussed below, as they will provide the basis for feature extraction in automatic detection.

The cues which have been claimed to mark **prosodic phrase boundaries** are numerous: boundary foot lengthening, preboundary lengthening, pauses, breaths, boundary tones, and speaking rate changes. Of these, we will consider all but foot lengthening. As shown by Wightman, *et al.* [21], lengthening of the interstress intervals (feet) is purely a consequence of preboundary lengthening. Thus, if we include preboundary lengthening as a feature, including foot lengthening as a feature would be redundant.

Preboundary lengthening describes the tendency for segments prior to a prosodic boundary to be longer than they would be in other contexts. References to this phenomenon are ubiquitous in the literature, and it has been observed in all the romance languages, Russian and Swedish, in addition to English [29]. Preboundary lengthening is a powerful cue, which O'Malley *et al.* [30] were able to exploit to recover the syntactic structure of algebraic expressions. As shown by Wightman, *et al.* [21], preboundary lengthening is strongly related to the perceived size of a boundary (break index) and occurs even at smaller breaks that do not include a pause.

The investigation reported in [21], however, also indicated that several perceptually distinct types of breaks were indistinguishable on the basis of preboundary lengthening. Two cues that appear to contribute to the perception of different levels above the intonational phrase are unfilled pauses and breaths. Unfilled pauses, or silences between words not associated with stop closures, are more frequent at larger breaks and, when present, tend to have longer duration for larger breaks. Several researchers have noted the importance of pauses in signaling major phrase boundaries [30]–[32], but results in [21] and [33] indicate that the duration of the pause, not just the fact of its occurrence, contains salient information. A pause, however, is not a completely reliable indicator of a major prosodic break. Many intonational phrase boundaries do not coincide with a pause and conversely, in spontaneous speech, a pause may be an indicator of hesitation and not a prosodic break.

Inhalation is a necessary part of speaking, and speakers appear to be quite particular about where they breathe. Work by Fodor *et al.* [34] suggests that breathing is sensitive to syntax and Sugito [35] has shown that it is sensitive to discourse. Price *et al.* [36] report that breaths in professional radio speech occur at 85% of the sentence boundaries, 53% of the other major phrase breaks, and almost nowhere else, although this may be quite different in spontaneous speech. In fact, as recognition researchers take on the challenges of longer utterances and spontaneous speech, explicit models of breaths are being incorporated in recognition (e.g., [37]). Breaths can also be detected independently with high accuracy. In the corpus used here, a three-class Gaussian classifier with a Markov sequence model (an extension of [38]) detected all 364 breaths in 63 paragraphs of speech, with a 3% false insertion rate.

The one cue which is always present at a major boundary is the boundary tone. Boundary tones are distinctive f0 features that signal major phrase boundaries. The characteristic low f0 at the end of a declarative sentence and the high f0 at the end of a yes/no question are both examples of boundary tones. In the phonology of intonation described by Beckman and Pierrehumbert [23], there are two (H and L) F0 boundary markers for intermediate phrases and two (H% and L%) additional markers at major intonational phrases, which combine to give four possible tunes at major boundaries. Other theories of intonation (e.g., [39] and [40]) also include the notion of major phrase boundary tones, though the number and classification systems generally differ.

The final cues which we will consider are speaking rate changes. While hand-labeling speech, we have often noticed that major boundaries, particularly those around parenthetical clauses, are often associated with abrupt changes in speaking rate. We can use the normalized duration measure developed in [21] to estimate speaking rate: by averaging over several syllables, we get an expression of how much longer or shorter the phones are, on average, than they would be in speech at the standard rate. Comparing the average normalized duration for several syllables before vs. after a word boundary should thus provide an estimate of the change in speaking rate.

We now turn our attention from phrasal boundaries to prominence. The cues which have been claimed to mark **phrasal prominence** include duration lengthening, pitch accents, and increased energy, all three of which have proven useful in the automatic prominence detection work of others. The relative importance of these cues and their interaction is not well understood, however, in part because of the confusion in the literature in terminology relating to stress and prominence.

Duration lengthening has been described as both a correlate distinguishing between lexically stressed and unstressed vowels (e.g., [41]), as well as an indicator of phrasal prominence (e.g., [42]). (Of course, it is possible that the lengthening of lexically stressed vowels is simply a consequence of the fact that only stressed vowels receive phrasal prominence.) Although lengthening can also be observed at prosodic phrase boundaries, we can distinguish such effects from prominence-related lengthening in two ways. First, by using separate acoustic models for stressed and unstressed vowels, we can reduce the chance of mistaking a stressed vowel for a lengthened version of its unstressed counterpart. We can also make use of Campbell's results which show that the lengthening in stressed syllables is uniformly distributed over the syllable, whereas boundary-related lengthening is concentrated in the syllable rhyme [42]. Another indicator of phrasal prominence is a pitch accent. As for boundary tones, different theories of intonation characterize different classes of pitch accents. For practical reasons, we shall not attempt to distinguish different phonological classes of accents, and will simply label syllables as prominent or not prominent, assuming that there is always some type of pitch accent associated with a prominent syllable. However, the algorithms that we shall describe are general enough to handle classification of different pitch accent types, given appropriate features. Finally, energy (or the energy integral over a syllable) appears to be an important cue for automatic detection of prominence. Little discussion of energy cues can be found in the linguistics literature, however, probably because energy is less important than F0 and duration in human perception of prominence.

It should be clear from the above discussion that, while the linguistics literature can provide guidance for the selection of

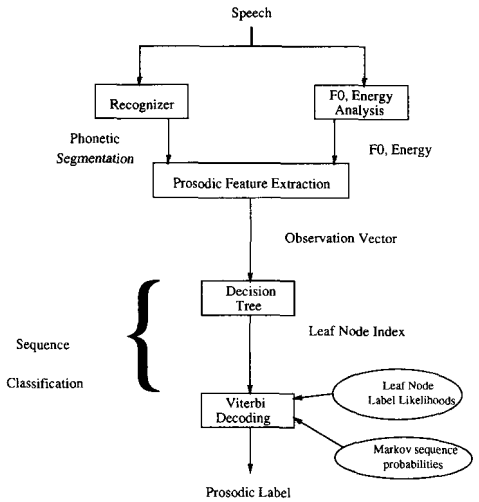

Fig. 1. Block diagram for the basic labeling system.

acoustic correlates of phrasal boundaries, there is insufficient quantitative data to provide a basis for classification rules. In addition, there are very few studies that have addressed the interaction of different cues. Our goal was thus to develop an algorithm that, given features based on those that have been posited in the literature, could automatically find a mapping between the correlates and the perceptually labeled prosodic patterns.

III. The Labeling Algorithm

Having explored the linguistic foundations of our labeling algorithm, we can now describe the basic architecture of the labeling system, illustrated in Fig. 1. Prosodic labeling can be posed as a standard pattern recognition problem, which involves first feature extraction and then classification. The feature extraction component maps the raw acoustic signal and associated phonetic alignments to a time sequence of feature vectors that ideally are good discriminators of the different classes (in this case, prosodic breaks or intonation markers). Because the two types of prosodic markers have different acoustic correlates and are associated with different-sized units, the feature extraction component is tailored to the specific class of prosodic markers to be labeled. The classification component, on the other hand, involves a more general mapping problem that can be solved with the same technique for the different classes of prosodic information. As a consequence, the discussion of the labeling algorithm will begin with a presentation of the general classification algorithm used here, based on decision trees, and then proceed with separate discussions of prominence and break labeling focusing on the feature extraction components.

A. General Prosodic Labeling Algorithm

Our approach to labeling prosodic information involves extracting syllable- or word-level features, and classifying these to a set of (syllable- or word-level) prosodic labels using a decision tree. Of the many different classification techniques available, we have chosen decision trees for prosodic labeling because they allow classification with nonhomogeneous features, without requiring assumptions about the independence of the features. In addition, it is simple to examine the structure of the tree (i.e., the sequence of questions found automatically in decision tree design) to gain insight into the relative importance of the various cues in the feature vector. A second component of the classification module is a Markov model of the label sequence. This component builds on results in speech recognition where it has been shown that the use of a language model (e.g., n-gram Markov model) can improve word recognition accuracy [43]. Similarly, for prosodic labels, we can take advantage of sequence modeling. For example, it is very unlikely that a break index of 6 would occur twice in a row; this would be a one-word sentence. These two components, the basic decision tree classifier and the Markov sequence model, will be described next with the label recognition (decoding) algorithm, followed by a discussion of parameter estimation issues associated with these modeling assumptions.

Probabilistic Model and Label Decoding: A decision tree [44] classifies a feature vector $\mathbf{x}$ by asking a series of questions about the elements of $\mathbf{x}$, the answers to which provide directions for descending the branches of the tree. The specific question asked at each node in the tree is determined in training, based on a prespecified set of allowable questions for the specific application. When a terminal node (or, leaf node) $T(\mathbf{x})$ is reached, the vector is given the classification label assigned to that leaf. Alternatively, the decision tree can provide the conditional probability distribution of the labels at the leaf node $\mathrm{p}(\alpha|T(\mathbf{x}))$, which is in fact what we need given that the model also incorporates the probability of the label sequence, as described below. In essence, the tree is used to classify each vector to a cell in the feature space that is associated with a label probability distribution, so the tree provides a piecewise constant mapping from feature space to label likelihoods.

Our goal is not simply to classify independent features, but to map a *sequence* of feature vectors $\mathbf{x}_1^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ to a *sequence* of prosodic labels $\alpha_1^n = \{\alpha_1, \ldots, \alpha_n\}$. Therefore, the classification component includes a model of the label sequence $\mathrm{p}(\alpha_1, \ldots, \alpha_n)$. One option would be to specify a grammar for the labels, and optionally to estimate the probabilities of grammar rules. A simpler option is taken here: assume a Markov (or n-gram "Markov") model that allows all possible label sequences, but represents their relative likelihood. That is, for a Markov or simple bigram model,

$$\mathrm{p}(\alpha_1, \ldots, \alpha_n) = \mathrm{p}(\alpha_1) \prod_{i=2}^{n} \mathrm{p}(\alpha_i|\alpha_{i-1}).$$

Combining the decision tree "acoustic model" with the Markov sequence assumption gives an overall probabilistic model of the form:

$$\mathrm{p}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \alpha_1, \ldots, \alpha_n)$$

$$= \mathrm{p}(\mathbf{x}_1|\alpha_1)\mathrm{p}(\alpha_1)\prod_{i=2}^{n}\mathrm{p}(\mathbf{x}_i|\alpha_i)\mathrm{p}(\alpha_i|\alpha_{i-1}) \quad (1)$$

$$= [\prod_{j=1}^{n}\mathrm{p}(\mathbf{x}_j)]L(\alpha_1|\mathbf{x}_1)\prod_{i=2}^{n}L(\alpha_i|\mathbf{x}_i)\mathrm{p}(\alpha_i|\alpha_{i-1}) \quad (2)$$

where

$$L(\alpha|\mathbf{x}) = \frac{\mathrm{p}(\alpha|\mathbf{x})}{\mathrm{p}(\alpha)} = \frac{\mathrm{p}(\mathbf{x}|\alpha)}{\mathrm{p}(\mathbf{x})}$$

$\mathrm{p}(\alpha|\mathbf{x})$ is given by the decision tree, $\mathrm{p}(\alpha)$ is the marginal probability of α, and $\mathrm{p}(\alpha_i|\alpha_{i-1})$ is given by the Markov model. The term $\prod_{j=1}^{n}\mathrm{p}(\mathbf{x}_j)$ can be ignored in recognition, since the goal is to chose the most likely label sequence:

$$\hat{\alpha}_1^n = \arg\max_{\alpha_1^n} p(\alpha_1^n, \mathbf{x}_1^n) \quad (3)$$

$$= \arg\max_{\alpha_1^n} L(\alpha_1|\mathbf{x}_1)\mathrm{p}(\alpha_1) \cdot \prod_{i=2}^{n} L(\alpha_i|\mathbf{x}_i)\mathrm{p}(\alpha_i|\alpha_{i-1}). \quad (4)$$

The maximization problem (4) can be solved with a dynamic programming algorithm, as follows.

1) For all α, compute $\mathcal{L}(\alpha, 1) = \log[L(\alpha|\mathbf{x}_1)\mathrm{p}(\alpha)]$
2) For each time $i = 2, \ldots, n$ and all α, compute

$$\mathcal{L}(\alpha, i) = \max_{\alpha'}\{\log[L(\alpha|\mathbf{x}_i)\mathrm{p}(\alpha|\alpha')] + \mathcal{L}(\alpha', i-1)\}$$

$$prev(\alpha, i) = \arg\max_{\alpha'}\{\mathcal{L}(\alpha', i-1) + \log \mathrm{p}(\alpha|\alpha')\}$$

3) $\hat{\alpha}_n = \arg\max_\alpha = \mathcal{L}(\alpha, n)$
4) For each time $i = n, \ldots, 2$

$$\hat{\alpha}_{i-1} = prev(\hat{\alpha}_i, i).$$

From (1), it is easy to see that this model is similar to a hidden Markov model (HMM) [43], [45]. The prosodic labels correspond to the states in an HMM, and the decision tree plays the role of the vector quantizer. In addition, the form of (2) and (4) are the same as those given for the reformulated HMM described in [46]. In recognition, the problem of choosing the most likely label sequence is the same as that of choosing the most likely HMM state sequence, so our labeling recognition algorithm is the same as HMM Viterbi decoding in speech recognition. In training, however, the prosodic labels are generally given and therefore the states/labels are not hidden, and as a consequence the iterative Baum-Welch algorithm used for training HMM's is not needed.

Training Considerations: We can use the HMM analogy to simplify the parameter estimation discussion. In [46], it was shown that the maximum likelihood estimate for the reformulated HMM was given by the relative frequency estimate for the Markov state transitions and by designing the quantizer (e.g., decision tree) to maximize the mutual information between the quantizer output and the HMM state. For the HMM, computing the maximum likelihood estimate is one step of a two-step iterative algorithm. For the prosodic labeling model, however, the iterative algorithm is not needed when labeled training data is available. Thus, maximum likelihood parameter estimation for this model simply involves standard decision tree design [44] using a maximum mutual information (minimum conditional entropy) criterion, as reviewed below, and relative frequency estimates for the Markov label sequence models.

The standard approach to decision tree design is a greedy growing algorithm, which successively splits nodes according to which node optimizes the tree design criterion over all possible questions at that node. The design criterion used here is maximum mutual information between prosodic labels α and tree leaves t:

$$I(T; A) = \sum_{\alpha,t}\mathrm{p}(\alpha, t)\log\frac{\mathrm{p}(\alpha, t)}{\mathrm{p}(\alpha)\mathrm{p}(t)}$$

which is equivalent to minimizing the conditional entropy

$$H(A|T) = -\sum_{\alpha,t}\mathrm{p}(\alpha, t)\log\mathrm{p}(\alpha|t)$$

since $\mathrm{p}(\alpha)$ is fixed. Thus, the node splitting criterion is

$$\max_{t,q}\mathrm{p}(t)H(A|t) - \mathrm{p}(t_l(q))H(A|t_l(q)) - \mathrm{p}(t_r(q))H(A|t_r(q)) \quad (5)$$

where the node probabilities can be substituted with the observation count for each node. The specific questions allowed q are determined by the type of prosodic labels being classified, e.g., breaks vs. intonation markers, and are described in the next sections. The parameters of the questions (e.g., threshold for testing a continuous variable) are chosen as part of the maximization in (5).

Since the tree design algorithm is greedy, it is possible to obtain better results (lower conditional entropy) by growing a large tree and pruning back to the desired size. In this study, the pruning is accomplished by using the optimal pruning method of Chou *et al.* [47], which prunes off the subtree which minimizes

$$\frac{\mathrm{H}(\text{subtree root}) - \mathrm{H}(\text{subtree leaves})}{(\text{number of nodes in subtree})}.$$

The pruning algorithm may be based on entropy estimates given by the training data, or estimates given by an independent data set, as in this work.

Applying the HMM analogy further, we can consider the case where prosodic labels are not given and introduce an algorithm for unsupervised training. We frequently encounter situations in which a relatively small amount of the training data is labeled, but a much larger unlabeled data set is also available. In these cases, we can use the labeled data to train a model according to the algorithm given for known label sequences, which we will refer to as "supervised training." This model can then be used as the starting point for refinement via "unsupervised learning," using the iterative Baum-Welch HMM estimation algorithm with the unlabeled data. The main complication associated with iterative training involves the decision tree design, since labels are no longer deterministically

assigned to observations. The joint quantizer/HMM design algorithm described in [46] addresses this problem by associating vectors of state likelihoods with each observation vector, and incorporating these in the conditional label distribution estimates in tree design. Alternatively, the Viterbi algorithm could be used to recover the most likely state sequence, which could then be used as labels in the standard tree design algorithm. The supplementary unsupervised training step could be an important tool in training models for prosodic labeling, since hand labeling data with prosodic labels is very costly. Using the original hand-labeled data (with the known labels) both for the initial estimates and in the corpus for iterative training may provide some insurance against the problems of model divergence associated with unsupervised training.

B. Labeling Intonation Markers

The prosodic labeling algorithm can be used to classify intonation markers, pitch accents, and boundary tones, which can be categorized at the syllable level.[3] As mentioned earlier, the set of classification labels in our experiments will include: unmarked syllable, accented syllable, boundary tone, and a syllable carrying both an accent and a boundary tone. (More detailed labels would be easily accommodated, but the prosodic labels were not available for much of the data at the time of this study.) Although one can describe a grammar of these units, e.g., two syllables within the same word cannot both contain a boundary tone, we chose to allow all possible label sequences for simplicity, under the assumption that the estimated relative frequencies of transitions between labels would eliminate these impossible events.

Since the labels occur at the syllable level, feature extraction will be at the syllable level as well. Thus, our feature extraction involves determining the boundaries between syllables and extracting a feature vector for each syllable based on durations of the phones in that syllable and the F0 and energy contours over the syllable. (F0 contours were estimated by an algorithm in Waves+, software from Entropic Research Laboratory, which is similar to that described by Secrest and Doddington [48].) The sequence of feature vectors will then be quantized and decoded to a sequence of syllable labels. Syllable boundaries can be stored in the recognition lexicon, but sufficient accuracy for this application is also possible with a simple set of rules.

In determining what features to extract and what questions to allow in the decision tree, we elected to use a relatively large number of features since there are no widely accepted tests for boundary tones. This approach allows the training algorithm to select the best features and, as a result, provides insight into the important features. Making use of the linguistic results reviewed earlier, we developed a set of 12 features which are extracted for each syllable. Three of these features are categorical; i.e., they assume values from a finite set.

- The shape of the next syllable's contour. This is a four-way classification of the syllable's F0 contour as a rise, a fall, a rise-fall (maximum), or a fall-rise (minimum), found by comparing the initial, final, and mean F0.
- A flag indicating if this syllable contains a stressed vowel, because theoretically prominence can only occur in these cases.
- A flag indicating if this syllable is word-final. Since boundary tones occur only on word-final syllables, this feature should prevent many false recognition errors. Incorrect labeling of word-internal pitch accents as boundary tones was one of the principal reasons for the high false alarm rates experienced by Butzberger [11] (his spotting algorithm did not have knowledge of the word boundary locations).

For these features, we use a simple value test as a node-splitting question. The best value to test for is selected as part of the tree design algorithm.

The eight remaining features are continuous-valued. To reduce speaker dependence, all F0-related features are expressed as ratios, comparing one F0 value with another. That is, we consider only F0 changes, thus extracting features that do not depend on particular F0 values. We also include several features that are useful in break detection, since these help distinguish boundary tones from prominences.

- Ratio of maximum F0 in this syllable to the mean F0 of the next syllable.
- Ratio of the maximum F0 in this syllable to the maximum F0 in the previous syllable.
- Ratio of the maximum F0 to the mean F0 within this syllable.
- Ratio of the minimum F0 to the mean F0 within this syllable.
- Preboundary lengthening, measured by the mean normalized duration of the syllable rhyme.
- The difference between the mean normalized duration of the syllable rhyme and the syllable onset. This feature is motivated by Campbell's finding that lengthening due to prominence affects the entire syllable whereas lengthening due to boundary phenomena affects principally the syllable rhyme [42].
- Pause duration.
- Ratio of the final F0 to the mean F0 within this sentence. This is the feature used by Daly and Zue to distinguish between boundary tones in statements and yes/no questions [8].
- The mean energy in the syllable, which several researchers have found useful for "stress" detection. (We also investigated normalizing this feature by the mean energy in the previous syllable and the mean energy in the utterance, but found the simple unnormalized energy feature to yield the best performance.)

The node-splitting questions for these features simply compare them to threshold values. Again, the best thresholds to use in these questions are determined as part of the tree training algorithm.

C. Labeling Prosodic Breaks

For the prosodic break labeling problem, there are seven different labels: break levels 0 through 6. As for the intonation

[3] In fact, the intermediate phrase boundary tones described by Beckman and Pierrehumbert [23] cannot be isolated to the syllable level in all cases, but the approximation is reasonable in many instances.

labels, we will allow all possible break label sequences, particularly since we have no information to suggest that any are not allowed. Unlike the intonation labels, prosodic breaks are associated with each word in the utterance, so the feature extraction algorithm will operate at the word level for this application. As shown in [21], however, the durational marking of prosodic boundaries is limited to the final syllable of the word, and any following pause.

As discussed in Section II, a number of cues mark prosodic phrase boundaries: breaths, pauses, preboundary lengthening, boundary tones, and speaking rate changes. Features motivated from these results are extracted for break detection.

- A flag indicating whether or not the word was followed by an audible breath. (The breaths marked in this corpus were hand-corrected, but automatic breath detection is very accurate, so experimental results should not be significantly different from a fully automatic implementation.)
- The duration of the pause following the word. If the word was followed by an audible breath, the duration of the breath was added to the durations of any silence. That is, the "pause duration" is the time between the syllable offset and the onset of the next word.
- The mean normalized duration of the rhyme of the syllable, as defined in [21].
- The difference between the mean normalized duration of the syllable rhyme and the syllable onset.
- The difference between the mean normalized duration of the three syllables prior to the word-final syllable and the mean normalized duration of the three syllables following the word-final syllable. This feature is a measure of the change in speaking rate that occurs at the boundary.
- A flag indicating whether or not the word contains any stressed vowels. As was shown in [21], the preboundary lengthening displayed prior to a break labeled with 0 or 1 is significantly different for words that contain stressed vowels than for words that do not. Inclusion of this feature allows the quantizer design to take this into account.
- The probability of a boundary tone in the final syllable of the word, estimated using the intonation labeling system described in the previous section. Specifically, the forward-backward algorithm is used to compute $\gamma_i(t) = \Pr(\text{state} = i \text{ at time } t | \text{ the observation sequence})$, and the terms that correspond to boundary tone states, in this case BT and P-BT, are summed to get the final feature value. We have found the probability of a boundary tone to be more useful than a simple flag indicating that the word-final syllable had been labeled with a boundary tone, because the threshold for rejecting boundary tone labels can be set in decision tree design.

We also investigated the inclusion of Pr(P) as a feature but found that it did not improve performance.

IV. Experimental Evaluation

To evaluate our labeling algorithm, and demonstrate its application, we have conducted a number of experiments. In this section, we describe those experiments and the performance achieved, and discuss the acoustic features found to be most important for automatic detection.

A. Corpora

We have tested our labeling algorithm using two corpora of professionally read speech, chosen based on the availability of hand-labeled prosodic markers. The first is the *ambiguous sentence corpus*, developed by Price *et al.* [22] for perceptual studies. The ambiguous sentence corpus contains 35 pairs of phonetically-similar, but syntactically ambiguous, sentences. These 70 sentences were each read by four professional FM radio news announcers (one male and three female) who are native speakers of American English, yielding a total of 280 sentences containing 2140 words (3091 syllables) and 7560 phone segments. Each sentence corresponds to a separate utterance, so sentence boundaries are known. Since there is a relatively small amount of data from any one speaker, this corpus is used in speaker-independent experiments, where three speakers are used in training and the fourth is used in testing. Performance estimates are obtained by repeating this four times, leaving out a different speaker each time. This corpus has the disadvantages that the vocabulary is the same for both testing and training and that the sentences were designed to be fully voiced. Although these factors might contribute to an optimistic performance estimate, they are counteracted to some degree by the small number of speakers available. Further details on this corpus are given in [22].

The second corpus is a collection of *radio news stories* read by a professional FM public radio announcer (female), recorded in the studio during actual broadcast. This corpus is a subset of the radio news corpus collected at Boston University [49]. The radio news stories used contain 8568 words (14 095 syllables) in 457 sentences. The sentences occur in paragraph-sized utterances (116 paragraphs), and sentence boundaries are not assumed to be known. This corpus also differs from the ambiguous sentence corpus in that the sentences are longer on average and there are audible breaths in the speech. The experiments on this corpus are speaker-dependent. A rotation algorithm is again used to obtain performance estimates. That is, two-thirds of the corpus is for training, the remaining third is used for testing, and the process is repeated for each third of the corpus.

Both corpora are hand-labeled with the seven-level break index and binary prominence labels described previously. Prosodic labeling typically involved two or three labelers who discussed disagreements before assigning the label. Inter-labeler agreement is fairly high using this transcription system. In an analysis of transcriber agreement for the ToBI system on a variety of speaking styles [24], expert break index labelers agreed within ± 1 index 95% of the time (for the 5-level ToBI break labels) and expert tone labelers agreed on presence vs. absence of accent in 86-88% of the words. Our own analyses on a small set of radio-style utterances found transcriber agreement to be slightly higher: 98% and 94% agreement for 7-level breaks and prominences, respectively.

In addition, the utterances in both corpora were automatically marked with phonetic labels and their corresponding

TABLE I
CONFUSION MATRIX FOR AUTOMATICALLY LABELED INTONATIONAL FEATURES USING A SPEAKER-INDEPENDENT ALGORITHM ON THE AMBIGUOUS SENTENCE CORPUS

hand labels	*automatic labels*				*total observed*
	S	P	P-BT	BT	
S	1457	223	8	23	1711
P	179	685	34	1	899
P-BT	11	41	191	9	252
BT	61	9	45	154	269

TABLE II
CONFUSION MATRIX FOR AUTOMATICALLY LABELED INTONATIONAL FEATURES USING A SPEAKER-DEPENDENT ALGORITHM ON THE RADIO NEWS STORIES

hand labels	*automatic labels*				*total observed*
	S	P	P-BT	BT	
S	7081	1075	51	141	8348
P	605	3191	82	17	3895
P-BT	43	121	400	80	644
BT	340	30	59	779	1208

time alignments using a recognizer constrained to the known word sequence. For the ambiguous sentence corpus, phonetic alignments were obtained from the SRI Decipher speech recognition system [50], which is based on hidden Markov models. For the radio news corpus, the segmentations were derived using a recognition system based on the stochastic segment model [51]. Both systems used a lexicon that supports multiple pronunciations of words and separate models for lexically stressed vs. unstressed vowels. The audible breaths in the radio news data were hand transcribed, automatically detected and then hand-edited to correct the small number of missed breath detections.

B. Prominence and Boundary Tone Labeling

Speaker-independent prominence and boundary tone labeling experiments were conducted on the ambiguous sentence corpus. The codebook rate was set at 45 after examining the trade-off between codebook rate and the conditional entropy of the class labels given the tree leaves. Table I contains a confusion matrix for the automatic labels vs. the hand labels. For the four classes, the overall accuracy of the algorithm is 79%. In interpreting performance, however, it is probably more useful to look separately at accuracy in recognizing prominences and boundary tones. Prominences (combining P and P-BT) are correctly detected at a rate of 83% and falsely detected at a rate of 14%. Boundary tones (combining BT and P-BT) are correctly detected at a rate of 77%, but have only a 3% false detection rate.[4] The boundary tone detection rate is considerably better than the HMM spotting result obtained by Butzberger [11], demonstrating the advantages of a post-recognition algorithm (e.g., having access to word boundary information reduces false detections) and the advantages of using syllable level features (e.g., normalized duration and F0 extrema) rather than frame-based features. Furthermore, a closer examination of the false rejection errors revealed that most were continuation rises. This is an indication that we have not yet extracted features appropriate for the detection of these rises and further research will be necessary to identify them and incorporate them into the system.

Speaker-dependent prominence and boundary tone labeling experiments were also conducted on the radio news corpus. The codebook rate was chosen to be the same as for the ambiguous sentence corpus. Table II contains a confusion matrix for the automatic labels vs. the hand labels. For the four classes, the overall accuracy of the algorithm is 81%. Prominence and boundary tone detection rates in the radio news are similar to those for the ambiguous sentence data: 84% correct vs. 13% false prominence detection and 71% correct vs. 2% false boundary tone detection. The more complex sentence and paragraph structure of the radio news corpus may account for the lower accuracy observed here than in Campbell's stress detection experiments [16], which were based on a small corpus of 27 sentences (100 utterances) simulating conference registration dialogue. The overall accuracy of presence vs. absence of prominence in our two corpora is 85–86%, which is fairly close to the 86–94% agreement of human labelers for this task.

Fig. 2 shows the top of the quantizer tree used for labeling intonation features in the ambiguous sentence corpus. (The tree for the radio news corpus is similar, with the exception that it makes use of breaths as a cue to boundary tones.) Notice that the first split is on the flag indicating whether the syllable nucleus is stressed. This separates almost all the prominent syllables, which confirms both our observations and linguistic theory. The syllables are then split on the basis of a pause which, as suggested above, is found only in major breaks which have boundary tones. Examining the top splits in the tree, where the largest reduction in conditional entropy are obtained, we find that the six most important features are (in order of decreasing information): the stress flag, pause duration, preboundary lengthening, the ratio of maximum F0 in this syllable to that of the previous syllable, the end-of-word flag, and the ratio of the final F0 to the mean F0 of the sentence.

C. Break Index Labeling

Speaker-independent break index labeling experiments were also conducted on the ambiguous sentence corpus. Here, the codebook rate was set at 16, again after examining the trade-off between codebook rate and the conditional entropy of the class labels, given the tree leaves. Table III gives the confusion matrix for hand-labeled vs. recognized breaks. For this experiment, sentence boundaries were given, so we present both results excluding these points and, in parentheses, corresponding figures including the sentence boundaries. The overall accuracy of the algorithm is 55% (60%) exact identification and 88% (90%) identification within ± 1. Defining break levels 4–6 as major phrase breaks, correct detection of major phrases is 64% (84%), with a false detection rate of 6%.

One problem with the detection algorithm is that very few breaks are labeled with indices of 0, 3, or 5. A likely

[4]The correct detection/false detection trade-off can be controlled by adjusting the Markov chain parameters to make boundary tones and prominences more or less likely. The trade-off point here reflects the use of the empirical distribution, which is appropriate for minimizing overall classification error.

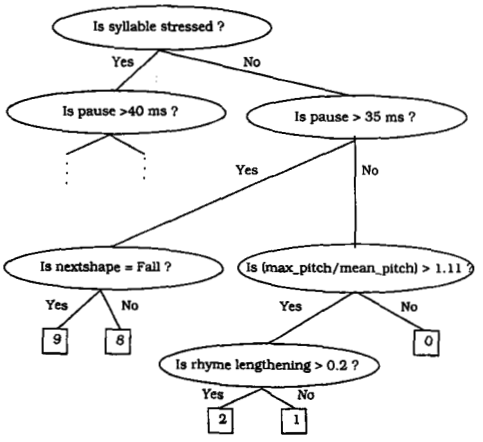

Fig. 2. Top nodes of quantization tree used to label intonational features.

TABLE III
CONFUSION MATRIX FOR AUTOMATICALLY LABELED BREAK INDICES USING A SPEAKER-INDEPENDENT SYSTEM ON THE AMBIGUOUS SENTENCE CORPUS

hand	*automatic labels*							*total*
labels	0	1	2	3	4	5	6	*observed*
0	3	189	26	1	2	0	0	221
1	7	704	178	0	20	3	0	912
2	0	145	178	3	36	0	0	362
3	0	48	44	1	36	0	0	129
4	0	23	57	2	128	0	0	210
5	0	1	1	0	24	0	0	26
6	0	0	0	0	0	0	280	280

TABLE IV
CONFUSION MATRIX FOR AUTOMATICALLY LABELED BREAK INDICES USING A SPEAKER-DEPENDENT SYSTEM ON THE RADIO NEWS CORPUS

hand	*automatic labels*							*total*
labels	0	1	2	3	4	5	6	*observed*
0	0	147	2	2	3	0	0	154
1	0	4355	96	163	131	0	0	4745
2	0	849	64	87	103	0	0	1103
3	0	311	40	162	201	0	0	714
4	0	177	45	160	564	55	5	1006
5	0	7	3	5	45	309	28	397
6	0	8	0	0	10	157	274	449

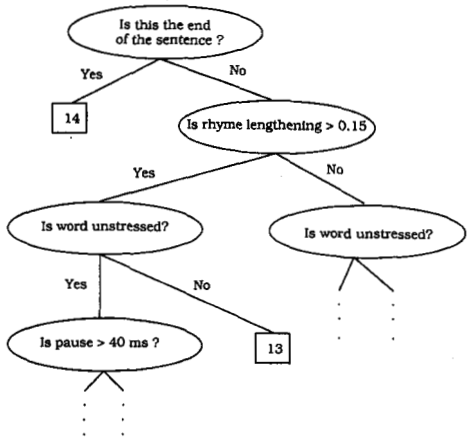

Fig. 3. Top nodes of the quantization tree used to label break indices in the ambiguous sentence data.

explanation for this behavior is that the corpus contains relatively few occurrences of the indices 0, 3, and 5, so they are less likely to be recognized because of their lower prior probability and because there are insufficient examples to estimate robust models. A good smoothing algorithm for the Markov transition probabilities might help address this problem.

Table IV shows the confusion matrix for the speaker-dependent radio news break index detection experiment. In this case, sentence boundaries are not generally given, and we include these points in the accuracy figures. The average accuracy for exact identification is 67%, with 89% correct identification within ± 1. Major phrases are correctly detected at a rate of 78%, with a 7% false detection rate. The different performance figures are similar to those achieved on the speaker-independent ambiguous sentence corpus. The main difference in performance is that break level 5 is much more reliably identified in the radio news corpus, because detected breaths almost always coincide with break levels 5 and 6. We continue to see difficulties in recognizing 0-boundaries because they are so rare relative to 1-boundaries; here, no boundaries were labeled with a 0. This could be corrected by including phonetic information in the acoustic feature vector so that evidence of clitization (*e.g.*, a flapped t) could be used. Overall, the performance is reasonably close to that of human transcriber agreement (although not as close as the performance of the prominence labeling): 89% agreement within ± 1 for the automatic algorithms vs. 95-98% for human labelers.

The top nodes of the quantization trees for the ambiguous sentence corpus and the radio news corpus are illustrated in Figs. 3 and 4, respectively. One difference between the two trees is that the top node in the ambiguous sentence tree is split on the flag, indicating that the boundary is utterance final, which provides all sentence boundaries (break level 6) in the ambiguous sentence corpus but not in the radio news corpus. In addition, the probability of a boundary tone is a more important feature in the radio news corpus, where there are many more utterance-internal major phrases than in the ambiguous sentence corpus. The break tree for the radio news also makes use of the breath cue, which is not available in the ambiguous sentences.

V. CONCLUSIONS

In summary, we have described an algorithm for labeling prosodic patterns in speech, based on a model that uses decision trees to map observations to probability distributions and Markov assumptions about the label sequence, similar to a discrete hidden Markov model. The model is trained using maximum likelihood estimation: recognition of prosodic pat-

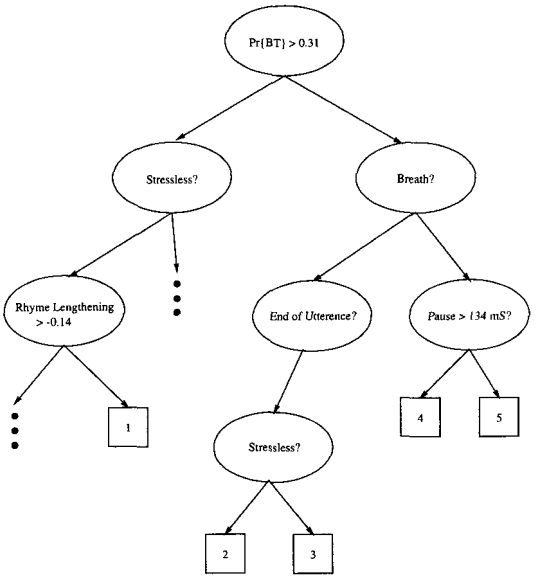

Fig. 4. Top nodes of the quantization tree used to label break indices in the radio news data.

terns involves dynamic programming, or Viterbi decoding, to find the most likely label sequence given acoustic observations. The algorithm is unique in that it assumes postrecognition processing, i.e., that automatically recognized phonetic labels and durations are available, and in that it is general enough to handle a variety of different prosodic patterns that may require a mix of categorical and continuous acoustic features. (In fact, the algorithm can be used to classify any sequence of patterns associated with vectors of nonhomogeneous features, making it applicable to other problems such as phone classification.) The specific types of prosodic patterns explored here are prosodic breaks, phrasal prominences or pitch accents, and boundary tones. Features appropriate to these applications are proposed based on results from linguistics, and though the features are application dependent, their use in the decision tree is determined through automatic training and the algorithm can easily be retrained in different domains. The algorithm was evaluated for prosodic labeling on two corpora of radio-style speech, including one set of speaker-independent experiments and one set of speaker-dependent experiments. Performance under the two conditions was similar, and in both cases the agreement between hand-marked labels and automatically detected labels is only slightly less than the inter-transcriber agreement reported for human labelers.

Although the performance of the prosody labeling algorithm is already quite good, there are several extensions that might yield improved performance. For example, we noted that improvements in boundary tone labeling (e.g., for continuation rises) might be achieved through more sophisticated representation of the types of intonation patterns in a syllable, which are now only crudely categorized as fall, rise, fall-rise, and rise-fall. Improvement in boundary tone detection would probably lead to improvement in prosodic break detection. It is also likely that there are other features that might be added to improve performance. Another option for obtaining better results is the use of more sophisticated questions in the decision tree, including categorical set membership questions and questions about multiple variables. Research on smoothing algorithms for both the observation distributions and the Markov transition probabilities would provide more robust parameter estimates and higher accuracy for the less frequent label classes. Finally, specification of a grammar for the label sequence, or the use of a higher order n-gram sequence model, would probably improve recognition accuracy. Of course, in addition to the extensions aimed at improving performance, a second area for future research is the use of (partially) unsupervised training for labeling large corpora and learning new prosodic categories.

As mentioned previously, the prosodic labeling algorithm described here is general enough to be applicable to other types of prosodic labels. Some examples include: breaks marked with the "p" diacritic according to the ToBI transcription system [24], which indicates a hesitation associated with excessive lengthening or pause duration; more detailed pitch accent and boundary tone labels, such as those used by the ToBI system; and labels for different levels of phrasal prominence. Application to different types of labels would require the development of new feature sets, though many of the features proposed here would still be useful. In addition to extending the algorithm to new classes of prosodic labels, an important related area for future work is extension to different speaking styles (e.g., dialogues) and to spontaneous, nonprofessional speech.

The experiments described here have focused on automatic labeling of prosodic information, that is, labeling an utterance where the word sequence is known. Automatic prosodic labeling will facilitate annotation of large corpora, which is essential for developing computational models of prosody for both speech synthesis and speech understanding applications. The algorithm described here is also applicable to speech understanding problems, where the word sequence is not known. In this case, we would still make use of recognizer output, but would consider multiple word or sentence hypotheses. Given the prosodic labeling model described here, there are many ways to incorporate prosody in speech understanding. Detection of phrasal prominence may be useful for spotting focused or important information, as suggested in [13]. Recognition of prosodic patterns can also be useful in resolving syntactic ambiguity. Mechanisms that have been proposed for using this particular model of prosody to resolve ambiguity include: using detected prosodic breaks directly in a parser through constraints on grammar rules [52], [53], using automatically detected breaks together with a model of the prosody/syntax relationship to compute a parse score (or a score of "prosodic consistency") [54], and using the decision tree component of the prosodic labeling model to provide the probability of prosodic patterns (rather than the detected patterns) for computing a parse score [55]. Though computational prosody research is still in its early stages, it is clear that the introduction of a general model for prosodic

pattern recognition represents an important step towards the use of prosody in spoken language systems.

ACKNOWLEDGMENT

The authors gratefully acknowledge J. Butzberger and H. Murveit at SRI and O. Kimball and F. Richardson at Boston University for their help in obtaining the phonetic alignments. Thanks also to N. Veilleux, K. Ross, G. Baldwin, and E. Shriberg for their work in labeling the speech corpora.

REFERENCES

[1] A. Cutler, "Prosody in situations of communication: Salience and segmentation," in *Proc. Int. Conf. Phonetic Sci.*, 1991, pp. 264–270.
[2] J. Pierrehumbert and J. Hirschber, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, P. Cohen *et al.*, Eds. Cambridge, MA: MIT Press, 1990.
[3] W. Lea, "Prosidic aids to speech recognition," in *Trends in Speech Recognition*, W. Lea, Ed. Englewood, NJ: Prentice-Hall, 1980, pp. 166–205.
[4] D. Huber, "A statistical approach to the segmentation and broad classification of continuous speech into phrase-sized information units," in *Proc. Int. Conf. ASSP*, 1989, pp. 600–603.
[5] A. Komatsu, E. Ohira, and A. Ichikawa, "Prosodical sentence structure inference for natural conversational speech understanding," in *Proc. Eurospeech 89*, 1989, vol. 2, pp. 400–403.
[6] H. Shimodaira and M. Kimura, "Accent phrase segmentation using pitch pattern clustering," in *Proc. Int. Conf. ASSP*, 1992, pp. 1217–1220.
[7] A. Waibel, *Prosody and Speech Recognition*. San Mateo, CA: Morgan Kaufmann, 1988.
[8] N. Daly and V. Zue, "Acoustic, perceptual, and linguistic analyses of intonation contours in human/machine dialogues," in *Proc. Int. Conf. Spoken Lang. Processing* (Kobe, Japan), 1990, pp. 497–500.
[9] A. Ljolje and F. Fallside, "Recognition of isolated prosodic patterns using hidden Markov models," *Comput., Speech, Lang.*, vol. 2, pp. 27–33, 1987.
[10] J. Butzberger Jr. *et al.*, "Isolated word intonation recognition using hidden Markov models," in *Proc. Int. Conf. ASSP*, 1990, vol. S-2, pp. 773–776.
[11] J. Butzberger Jr., "Statistical methods for analysis and recognition of intonation patterns in speech," Master's thesis, Boston Univ., Jan. 1990.
[12] X. Chen *et al.*, "A hidden Markov model applied to Chinese four-tone recognition," in *Proc. Int. Conf. ASSP, 1987*, vol. 2, pp. 787-800.
[13] F. Chen and M. Withgott, "The use of emphasis to automatically summarize a spoken discourse" in *Proc. Int. Conf. ASSP*, 1992, vol. 1, p. 229–232.
[14] E. Nöth *et al.*, "Intensity as a predictor of focal accent," in *Proc. 13th Int. Cong. Phonetic Sci.*, 1991, pp. 230–233.
[15] J. Hieronymous, D. McKelvie, and F. McInnes, "Use of acoustic sentence level and lexical stress in HSSM speech recognition," in *Proc. Int. Conf. ASSP*, 1992, vol. 1, pp. 225–227.
[16] W. N. Campbell, "Prosodic encoding of speech," in *Int. Conf. Spoken Lang. Processing* (Banff, Canada), 1992, pp. 663–66.
[17] C. Jones, *A History of English Phonology*. New York: Longman, 1989.
[18] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York: Harper and Row, 1968.
[19] J. P. Gee and F. Grosjean, "Performance structures: A psycholinguistic and linguistic appraisal," *Cognitive Psychology*, vol. 15, pp. 411–458, 1983.
[20] M. Y. Liberman and A. Prince, "On stress and linguistic rhythm," *Linguistic Inquiry*, vol. 8, pp. 249–336, 1977.
[21] C. Wightman *et al.*, "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Amer.*, Mar. 1992.
[22] P. Price *et al.*, "The use of prosody in syntactic disambiguation," *J. Acoust. Soc. Amer.*, vol. 90, pp. 2956–2970, 1991.
[23] M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," in *Phonology Yearbook 3*, pp. 255–309, 1986.
[24] K. Silverman *et al.*, "Break indices: A standard for prosodic transcription," in *Proc. Int. Conf. Spoken Lang. Processing* (Banff, Canada), 1992.
[25] E. Selkirk, *Phonology and Syntax: The Relation Between Sound and Structure*. Cambridge, MA: MIT Press, 1984.
[26] D. Bolinger, "Pitch accent and sentence rhythm," in *Forms of English: Accent, Morpheme, Order*, D. Bolinger, Ed.. Cambridge, MA: Harvard Univ. Press, 1965.
[27] S. Shattuck-Hufnagel, M. Ostendorf, and K. Ross. "Pitch accent placement within lexical items in American English," to appear in *J. Phonetics*.
[28] M. Beckman *et al.*, "Stress shift, stress clash and polysyllabic shortening in a prosodically annotated discourse," in *Proc. Int. Conf. Spoken Lang. Processing*, 1990, pp. 5–8.
[29] J. Vassière, "Language-independent prosodic features," in *Prosody: Models and Measurements*, A. Cutler and D. R. Ladd, Eds.. New York: Springer-Verlag, 1983, pp. 53–66.
[30] M. H. O'Malley, D. R. Kloker, and B. Dara-Abrams, "Recovering parentheses from spoken algebraic expressions," *IEEE Trans. Audio, Electroacoust.*, vol. AU-21, pp. 217–220, 1973.
[31] N. Macdonald, "Duration as a syntactic boundary cue in ambiguous sentences," in *Proc. IEEE Int. Conf. ASSP* (Philadelphia), 1976.
[32] I. Lehiste, J. Olive, and L. Streeter, "Role of duration in disambiguating syntactically ambiguous sentences," *J. Acoust. Soc. Amer.*, vol. 60, no. 5, pp. 1199–1202, 1976.
[33] N. Kaiki and Y. Sagisaka, "Pause characteristics and local phrase-dependency structure in Japanese, in *Proc. Int. Conf. Spoken Lang. Processing* (Banff, Canada), 1992, pp. 357–360.
[34] J. A. Fodor, T. G. Bever, and M. F. Garrett, *The Psychology of Language*. New York: McGraw-Hill, 1974.
[35] M. Sugito, "On the role of pauses in production and perception of discourse," in *Proc. Int. Conf. Spoken Lang. Processing* (Kobe, Japan), 1990.
[36] P. J. Price, M. Ostendorf, and C. W. Wightman, "Prosody and parsing." in *Proc. DARPA Speech, Natural Lang. Workshop*, October 1989.
[37] J. Butzberger *et al.*, "Spontaneous speech effects in large vocabulary speech recognition applications," in*Proc. DARPA Speech, Natural Lang. Workshop*, Feb. 1992, pp. 339–343.
[38] C. W. Wightman and M. Ostendorf, "Automatic recognition of prosodic phrases," in *Proc. IEEE Int. Conf. ASSP* (Toronto), 1991.
[39] D. Bolinger, *Intonation and Its Parts*. Stanford, CA: Stanford Univ. Press, 1986.
[40] J. 't Hart, R. Collier, and A. Cohen, *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge, UK: Cambridge Univ. Press, 1990.
[41] T. H. Crystal and A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech," *J. Acoust. Soc. Amer.*, vol. 88, no. 1, pp. 101–112, 1990.
[42] W. N. Campbell, "Evidence for a syllable-based model of speech timing," in *Proc. Int. Conf. Spoken Lang. Processing* (Kobe, Japan), 1990, pp. 9–12.
[43] L. Bahl, F. Jelinek, and R. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, no. 5, pp. 179–190, 1983.
[44] L. Breiman *et al.*, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks/Cole Advanced Books and Software, 1984.
[45] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[46] M. Ostendorf and R. Rohlicek, "Joint quantizer design and parameter estimation for discrete hidden Markov models," in *Proc. IEEE Int. Conf. ASSP* (Albuquerque, NM), 1990.
[47] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Optimal pruning with applications to tree-structured source coding and modeling," *IEEE Trans. Inform. Theory*, vol. 35, no. 2, pp. 299–336, 1989.
[48] B. Secrest and G. Dodington, "An integrated pitch tracking algorithm for speech systems," in *Proc. Int. Conf. ASSP*, 1983, pp. 1352–1355.
[49] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," in manuscript.
[50] M. Weintraub *et al.*, "Linguistic constraints in hidden Markov model based speech recognition," in *Proc. IEEE Int. Conf. ASSP* (Glasgow), pp. 699–702, 1989.
[51] M. Ostendorf *et al.*, "Continuous word recognition based on the stochastic segment model," in *Proc. DARPA Workshop Continuous Speech Recognit.*, 1992.
[52] J. Bear and P. Price, "Prosody, syntax and parsing," in *Proc. Assoc. Computational Linguistics*, 1990.
[53] M. Ostendorf *et al.*, "The use of relative duration in syntactic disambiguation," in *Proc. 3rd DARPA Workshop Speech, Natural Lang.*, 1990, pp. 26–31. A shorter version appears in *Proc. Int. Conf. Spoken Lang. Processing-90* (Kobe, Japan), pp. 13–16.
[54] M. Ostendorf, C. Wightman, and N. Veilleux, "Parse scoring with prosodic information: An analysis/synthesis approach, *Comput. Speech, Language*, pp. 193–210, July 1993.
[55] N. Veilleux and M. Ostendorf, "Probabilistic parse scoring with prosodic information, in *Proc. Int. ASSP*, 1993.

Colin W. Wightman (M'92) received the B.S. degree in engineering from Swarthmore College in 1982, the M.S.E.E. degree from the State University of New York at Stony Brook in 1985, and the Ph.D. degree in electrical engineering from Boston University in 1992.

In 1983, he joined the AIL division of Eaton Corporation, where he worked on digital signal processors for very low frequency radio communications. In 1985, he joined the Equipment Division of Raytheon Company, where he was involved in several aspects of developing ground and airborne terminals for satellite communication networks. He is currently an Assistant Professor in the Electrical Engineering Department at the New Mexico Institute of Mining and Technology, where he has been since 1991. His research interests include machine intelligence and pattern recognition, particularly the understanding of spoken language by machines.

Dr. Wightman is a member of the Acoustical Society of America, the Association for Engineering Education, and Sigma Xi.

Mari Ostendorf (M '85) received the B.S., M.S. and Ph.D. degrees in 1980, 1981 and 1985, respectively, all in electrical engineering, from Stanford University, Stanford, CA.

In 1985 she joined the Speech Signal Processing Group at BBN Laboratories, where she worked on low-rate coding and acoustic modeling for continuous speech recognition. She is currently an Associate Professor in the Department of Electrical, Computer and Systems Engineering at Boston University, which she joined in 1987. Her research interests include data compression and statistical pattern recognition, particularly in speech processing applications. Her recent work involves investigation of segment-based models for continuous speech recognition and stochastic models of prosody for both recognition and synthesis.

Dr. Ostendorf has served on the Speech Processing Committee of the IEEE Signal Processing Society and is a member of Sigma Xi.