Automatic Classification of Intonational Phrase Boundaries

Michelle Q. Wang
Churchill College
Cambridge University
Cambridge UK

Julia Hirschberg
AT&T Bell Laboratories
2D-450
600 Mountain Avenue
Murray Hill NJ 07974

June 21, 1995

Abstract

The relationship between the intonational characteristics of an utterance and other features inferable from its text represents an important source of information both for speech recognition, to constrain the set of allowable hypotheses, and for speech synthesis, to assign intonational features appropriately from text. This work investigates the usefulness of a number of textual features and additional intonational features in predicting the location of one particular intonational feature — intonational phrase boundaries — in natural speech. The corpus for this investigation is 298 utterances from the 774 in the DARPA-collected Air Travel Information Service (ATIS) database. For statistical modeling, we employ classification and regression tree (CART) techniques. We achieve success rates of just over 90%, representing a major improvement over other attempts at boundary prediction. (This paper appeared in *Computer Speech and Language*, 6 (1992): 175–196.)

1 Introduction

Understanding the relationship between an utterance's intonational features and features which can be inferred from the text uttered provides important information both for speech recognition and for speech synthesis. In recognition, the association of intonational features with syntactic and acoustic information can be used to reduce the number of candidates under consideration. In synthesis, better understanding of the relationship between intonational features and syntactic and discourse-level characteristics of the text in natural speech provides the basis for algorithms to assign intonational features more naturally.

In this study, we investigate the usefulness of a number of textual features and additional intonational features in predicting the location of one particular intonational feature — intonational phrase boundaries — in natural speech. We select potential boundary predictors based upon hypotheses derived from linguistic theory (Section 2), as well as examination of previous studies of phrase boundaries (Section 3). Our corpus for this investigation is 298 sentences from the 774 sentences of the DARPA-collected Air Travel Information Service (ATIS) database(DARPA, 1990), described in Section 4.1. For statistical modeling, we employ classification and regression tree techniques (CART), described in Section 4.3. Our experiments and results are described in Sections 4.2-4.4. In Section 5, we discuss the implications of our findings and future directions for this research.

2 Intonational Theory

Intuitively, the intonational phrasing of an utterance divides it into meaningful 'chunks' of information (Bolinger, 1989). For example, variation in a sentence's intonational phrasing can change the meaning hearers are likely to assign to an individual utterance of the sentence. Consider the utterances illustrated in (1), where the presence of a comma denotes a phrase boundary:

(1) a. Bill doesn't drink because he's unhappy.

b. Bill doesn't drink because he's unhappy. (He drinks because he's an alcoholic).

c. Bill doesn't drink, because he's unhappy. (He believes that alcohol will amplify his depression).

If a speaker utters the sentence (1a) as a single phrase, as in (1b), hearers are likely to infer that Bill does indeed drink — but not from unhappiness. If the speaker instead utters the sentence as two phrases, as in (1c), hearers are likely to understand that Bill does *not* drink and that the reason for his abstaining is his unhappiness.

In describing intonation phrasing in particular and intonational features in general, we adopt Pierrehumbert's theory of intonational description for English (Pierrehumbert, 1980). According to this view, two levels of phrasing are significant in English intonational structure. Both types of phrases are composed of sequences of high and low tones in the FUNDAMENTAL FREQUENCY (f0) contour. An INTERMEDIATE (or minor) PHRASE consists of one or more PITCH ACCENTS (local f0 minima or maxima) plus a PHRASE ACCENT (a simple high or low tone which controls the pitch from the last pitch accent of one intermediate phrase to the beginning of the next intermediate phrase or the end of the utterance). INTONATIONAL (or major) PHRASES consist of one or more intermediate phrases plus a final BOUNDARY TONE, which may also be high or low, and which occurs at the end of the phrase. Thus, an intonational phrase boundary necessarily coincides with an intermediate phrase boundary, but not vice versa. Differences in phrasing are illustrated in the f0 contours depicted in Figures 1, illustrating the utterance of (1b), and 2, illustrating (1c).[1]

Figure 1: *Bill doesn't drink because he's unhappy* as One Phrase

While phrase boundaries are perceptual categories, they are associated with certain physical characteristics of the speech signal. In addition to the tonal features described above, phrases may be identified by one of more of the following features: pauses (which may be filled or not), changes in amplitude,

[1] These contours are produced by the Bell Labs Text-to-Speech System (Olive and Liberman, 1985).

Figure 2: *Bill doesn't drink because he's unhappy* as Two Phrases

and lengthening of the final syllable in the phrase (sometimes accompanied by glottalization of that syllable and perhaps preceding syllables). In general, major phrase boundaries tend to be associated with longer pauses, greater tonal changes, and more final lengthening than minor boundaries.

3 Previous Research

The variety of previous research on the prediction of intonational boundaries classification is a tribute to the complexity of the problem. In this section, we review three of the more recent studies. Each of these employs analysis of varying amounts of natural speech to develop boundary prediction procedures.

Altenberg's (Altenberg, 1987) analysis of a portion of the London-Lund corpus is unusual among studies of boundary prediction both in the large size of the corpus analyzed and in its reliance upon semantic and discourse-level as well as syntactic information for prediction purposes. In proposing a partial algorithm for assigning intonational features in speech synthesis, Altenberg examines intonational phrasing and pitch accent location in a 45-minute prepared talk by a single speaker. Altenberg presents frequency distributions for factors such as part-of-speech and syntactic constituency as well as surface position. Throughout, he assumes that the data is both correctly tagged and parsed, and that subtle semantic and discourse-level distinctions (e.g. distinguishing restrictive from non-restrictive relative clauses and determining 'focus of attention') are available. However, he proposes no automatic method for acquiring either discourse-level information or the syntactic information he requires.

The results of Altenberg's analysis vary widely depending on the syntactic domain of analysis, which he divides into sentence, clause, and phrase level boundaries. For the sentence level, Altenberg's syntactico-semantic rules appear to perform fairly well, successfully separating units characterized as 'direct speech', adverbial infinitivals, prepositional complements, and nominal *that*-clauses) into intonational phrases with 88.6% success (of 554 designated sentence-level structural junctures). However, for boundaries associated with smaller domains, Altenberg concludes that "prosodic boundaries are generally grammatically unpredictable" (Altenberg, 1987, p.121), citing performance of 66% on 'expanded clauses' and 62% on 'expanded (multiple) phrases', which represent only a portion of his data. Thus, although Altenberg's rules perform well on certain types of major boundaries, they do not seem suitable for general boundary prediction. Even where his rules are successful, the nature of the input required makes his algorithm impossible to automate. Also, the algorithm has only been tested on the training corpus.

Also intended to generate phrasing rules for speech synthesis, the system proposed by Bachenko and Fitzpatrick (Bachenko and Fitzpatrick, 1990) considers only variables that can be inferred from text, including constituent structure, lexical information, and surface position. They extract this information using a modified version of Hindle's parser (Hindle, 1989). The motivation for their rules comes from Gee and Grosjean (Gee and Grosjean, 1983). These rules come in two different flavors — location rules and boundary salience rules. The location rules effectively derive phonological words using both tree position and lexical information. The boundary salience rules use a combination of syntactic constituency, verb balancing and verb adjacency, length rules, and final bundling. The rules work together to assign relative values of boundary salience. First, the syntactic rules apply and produce low-level phrases. Next, the verb balancing and verb adjacency rules measure the length of constituents to the left and right of the verb and ensure that the verb becomes the intonational pivot of the sentence. Then, length rules work to divide the post-verbal material into phrases. Finally, the bundling rules use the boundary salience values to collect all the phrases into a binary tree. This tree then represents the phrasing assigned to the utterance. Three levels of phrasing are assigned by the system, although the authors evaluate system performance only on primary and secondary boundary assignment.

System performance is tested on a corpus of 35 sentences of read speech, 21 chosen from the training corpus and 14 used for evaluation by Grosjean and Gee.[2] A total of 84 primary and secondary boundaries appear in these 35 sentences.[3] If we assess boundary assignment by type, Bachenko and Fitzpatrick predict approximately 60% of 45 observed primary boundaries correctly, overgenerating by only five, and also predict 44% (17 of 39) of secondary boundaries

[2] While two versions of the 21 training sentence were originally collected, only the one with a greater number of boundaries was actually used in evaluation.

[3] Summarizing results presented in Bachenko and Fitzpatrick (Bachenko and Fitzpatrick, 1990). are here augmented by analysis of the predictions, which are presented in an appendix.

correctly, over-generating here by 24. If distinctions between primary and secondary phrasing are ignored, then this system correctly predicts the location of 56% of boundaries (47 of 84) and predicts only 5 spurious boundaries — although 2 of the Grosjean and Gee sentences are reworded slightly to obtain a prosodic analysis. While this success rate appears relatively low, testing on a larger corpus might provide a more balanced evaluation.

In contrast to the above studies, Ostendorf et al. (Ostendorf et al., 1990) attempt to incorporate prosodic boundaries into a syntactic parser to reduce syntactic ambiguity in in speech recognition. To this end, they collected a corpus of 35 pairs of structurally ambiguous sentences (ambiguous in seven different ways) from three radio announcers, and hand labeled the speech with prosodic 'break indices', assigning boundary levels from 0 (no boundary) to 6 (utterance final). They then trained a Hidden Markov Model to compute break indices from the observed relative durations of phonetic segments.[4] The model was trained on the speech of two speakers and tested on the third. (Because of the similarity in speech patterns of the announcers, speaker independence is difficult to assess.) For the training data, this technique predicted two levels of observed boundary with better than 86% success. For the test data, the correlation between predicted and observed boundaries was between .80 and .83 (variation due to differences in hand-labeling). Finally, observed break indices were incorporated into some of the grammar rules of the SRI Core Language

[4] Numeric indices later interpreted as boundary likelihoods were computed by adding together a pause factor and the average of the normalized duration over the rhyme of the final syllable of each word.

Engine grammar (Alshawi et al., 1988), so that observed prosodic boundaries could be used in syntactic analysis. When this modified grammar was tested on the seven sentence pairs representing preposition/particle ambiguities, the new grammar was found to reduce the number of parses found by about 25%, without ruling out any correct parses. Further data for testing is currently being collected.

While the preliminary results of this experiment are intriguing, the question remains whether or not these results will generalize beyond the current genre and beyond the current set of syntactic ambiguities being examined. It is widely recognized that radio speech differs from natural, spontaneous speech in both phrasing and accent(Bolinger, 1989), so it is not clear that ordinary speakers will disambiguate intonationally potentially ambiguous utterances as do professional speakers. Also, the limitation of this study to seven particular types of structurally ambiguous sentence, while interesting for the study of whether speakers use intonational to disambiguate such sentences, does not tell us much about how intonational phrasing in sentences not perceived to be structurally ambiguous. And finally, as the authors note, these findings must be tested on much more data than the seven sentence pairs tested, before the method's usefulness for predicting relationships between intonational boundaries and syntactic information can be established.

The studies described above provide some indication of current goals in boundary prediction as well as some of the problems of achieving these goals.

Clearly, testing on larger corpora will be necessary before approaches such as Bachenko and Fitzpatrick's or Ostendorf et al.'s can be fully evaluated. In addition, confining testing to syntactically ambiguous sentences or indeed to read sentences at all is at best an indirect index into the uses of intonational phrasing in natural speech. While Altenberg's corpus is both larger and, arguably, more 'natural', his phrasing rules cannot be used for automatic classification, since they assume input not currently available from parsers or other means of text analysis — and not likely to be available in the near future.

As a step toward addressing both the problems of corpus size and type and of the need for classifying boundaries from variables that can be extracted automatically, we decided to examine a corpus of spontaneous multi-speaker speech, to determine how well information currently obtainable via simple techniques of text analysis can predict intonational boundaries in natural speech, and how such predictions compare to those made when observed additional intonational information is also available.

4 The Study

4.1 Data

The corpus used in this analysis consists of 298 utterances (24 minutes of speech from 26 different speakers) collected by Texas Instruments for use in the DARPA ATIS (Air Travel Information System) spoken language system evaluation task. These 298 tokens represent all of the (syntactic) questions currently in the ATIS database.[5] In a Wizard-of-Oz simulation, subjects were given a travel scenario and asked to make travel plans accordingly, providing spoken input and receiving output at a terminal. Speakers were told that their utterances were being recognized by machine; vocabulary constraints were enforced by error messages. The full ATIS corpus currently consists of 774 sentences, culled from these experiments. The quality of the ATIS corpus is extremely diverse. Speakers range in fluency from close to isolated-word speech to exceptional fluency. Many utterances contain hesitations and other disfluencies, as well as long pauses (greater than 3 sec. in some cases). The number of such pauses may reflect the complexity of the task confronted by speakers. A scenario typical of those presented to the subjects appears below:

> A small gymnastic team from overseas will be touring several cities in this country. Pick the cities they will visit and plan their itinerary. Please be sure: 1) the team will not miss meals, 2) there will be ground transportation to and from downtown, and 3) they do not travel on weekends, and 4) they fly on large (presumably safe) aircraft.

In addition, the novelty of communicating with a computer appears to have influenced the verbal styles chosen by some speakers.

[5] Although the corpus is currently limited to questions, we plan to test our hypotheses on the declarative sentences in the near future. While it does not seem likely that questions and declaratives will require separate phrase boundary prediction rules, the ease of automatically distinguishing questions from declaratives means that the current analysis would remain useful even in the event that separate rule sets prove to be necessary.

To prepare this data for analysis, we labeled the speech prosodically by hand, noting location and type of intonational boundaries and presence or absence of pitch accents. Labeling was done from both the waveform and pitchtracks of each utterance. Each label file was checked by 2-3 labelers. We labeled two levels of boundary, intonational phrase boundaries and intermediate phrase boundaries; in the analysis presented below, however, these are collapsed to a single category. More data will be needed for analysis before this distinction can be productively analyzed.

4.2 Selection of Variables for Analysis

For the current analysis of boundary location, we define our data points to consist of all potential boundary locations in an utterance, defined as each pair of adjacent words in the utterance $< w_i, w_j >$, where w_i represents the word to the left of the potential boundary site and w_j represents the word to the right. We selected these potential variables for analysis based upon their theoretical potential as boundary predictors and the feasibility of acquiring values for these variables automatically.

Given the variability in performance we observed among speakers, an obvious variable to include in our analysis is speaker identity. While for applications to speaker-independent recognition this variable would be uninstantiable, we nonetheless need to determine how important speaker idiosyncracy may be in boundary location. We have found no significant increase in predictive power when this variable is used; so, results presented below are speaker-independent.

One class of variable which is readily obtainable involves temporal information. Temporal variables include utterance and phrase duration, and distance of potential boundary from various strategic points in the utterance. Although it is tempting to assume that phrase boundaries represent a purely intonational phenomenon, we must allow for the possibility that processing constraints help govern their occurrence. That is, longer utterances may tend to be produced with more boundaries for mechanical production reasons. Accordingly, we measure the length of each utterance both in seconds and in words. The distance of the boundary site from the beginning and end of the utterance is another variable which appears likely to be correlated with boundary location. We speculate that the tendency to end a phrase might be affected by the position of the potential boundary site in the utterance. For example, it seems likely that positions very close to the beginning or end of an utterance might be less likely positions for intonational boundaries — again, perhaps, on grounds of production constraints. We measure this variable too, both in seconds and in words.

Gee and Grosjean (Gee and Grosjean, 1983) and Bachenko and Fitzpatrick (Bachenko and Fitzpatrick, 1990) inter alia have noted the importance of phrase length in determining boundary location. Simply put, it seems possible that consecutive phrases often have roughly equal lengths. To capture this notion, we calculate the following fraction: the elapsed distance from the last boundary to the potential boundary site, divided by the length of the last phrase encountered.

Unfortunately, to obtain this information from text analysis alone would require us to factor prior boundary predictions into subsequent predictions. As a first step, therefore, to test the utility of this information, we have used observed boundary locations in our current analysis. However, using prior boundary prediction would be feasible, if we discover that this information has predictive power. We employ both temporal and word count prediction for this variable.

Previous researchers have long considered syntactic information to be a good predictor of phrasing information (Gee and Grosjean, 1983; Selkirk, 1984; Marcus and Hindle, 1985; Steedman, 1990). To investigate such possibilities, we consider first the sentence type to which the potential boundary site belongs. It is possible that certain classes of sentences, e.g., indirect questions, will contain more phrase boundaries than others, possibly due to some disambiguating function that phrasing might serve. We accommodate this possibility by dividing questions into three main categories — *wh*-questions, direct yes/no questions, and indirect yes/no questions. Certainly if question type turns out to be an important predictor for presence or absence of intonational boundary, results of our future examination of declaratives should prove even more interesting.

Part-of-speech information is another factor widely believed to predict boundary location, particularly in text-to-speech. For example, the belief that phrase boundaries rarely occur after function words forms the basis for most algorithms used to assign intonational phrasing for text-to-speech. Furthermore, we might expect that heads of phrases (e.g. prepositions and determiners) do not constitute the typical end to an intonational phrase. We explore these ideas by examining a window of four words surrounding each potential phrase break. We obtain part-of-speech information via Church's part-of-speech tagger (Church, 1988), whose output has been modified slightly to predict preposition/ particle distinctions.

In addition to part-of-speech information, we also investigate the importance of syntactic constituency to the prediction of boundary location. Intuitively, we want to test the notion that some constituents may be more or less likely than others to be internally separated by intonational boundaries. To test which constituents are *less* likely to be intonationally separated, we examine the class of the lowest node in the parse tree to dominate both w_i and w_j. To test which constituents are *more* likely to be separated, we determine the class of the highest node in the parse tree to dominate w_i, but not w_j, and the class of the highest node in the tree to dominate w_j but not w_i. In this way, we can determine whether constituency relations are important to intonational phrasing. We use Hindle's parser, Fidditch (Hindle, 1989), to provide information on syntactic constituency for this analysis.

Recall that each intermediate phrase is composed of one or more pitch accents plus a phrase accent, and each intonational phrase is composed of one or more intermediate phrases plus a boundary tone. Clearly, then, pitch accents form the main building blocks for phrases. In addition, informal observation suggests that phrase boundaries are more likely to occur in some accent contexts than in others. For example, phrase boundaries between words that are deaccented seem to occur much less frequently than boundaries between two accented words. To test this observation, we look at the pitch accent values of w_i and w_j for each $< w_i, w_j >$. Since our prosodic labeling of the speech under analysis includes pitch accent location, we can use observed data values for the accent variables. However, since we would prefer to predict boundary location simply from information collected from text, we also substituted predicted pitch accent information for the observed data. We obtain accent predictions from text analysis procedures described in (Hirschberg, 1990). We use predicted accent as a binary feature (accented or not) or as a four-valued features (cliticized, deaccented, accented, or 'NA', for end of sentence), to see whether or not finer-grained information will prove useful in predicted boundary location. On the one hand, we want to discover from actual accent information how useful accent location can be in predicting boundary location. On the other hand, we want to determine whether information available from text analysis can help to automate our analysis.

In the analyses described below, we employ varying combinations of these variables to predict intonational boundaries. We use classification and regression tree techniques to generate decision trees automatically by selecting among variables provided to minimize error in a step-wise procedure.

4.3 Classification and Regression Tree Analysis

Classification and regression trees (CART) techniques (Breiman et al., 1984) permit the automatic generation of decision trees from sets of continuous or discrete variables. At each node in the generated tree, CART selects the statistically most significant variable to minimize prediction error at that point. In the implementation of CART used in this study (Riley, 1989)), all of these decisions are binary, based upon consideration of each possible binary split of values of categorical variables and consideration of different cut-points for values of continuous variables.

Generation of CART decision trees depends on a set of splitting rules, stopping rules, and prediction rules. These rules affect the internal nodes, subtree height, and terminal nodes, respectively. At each internal node, the program must determine which factor should govern the forking of two paths from that node. Furthermore, the program must decide which values of the factor to associate with each path. Ideally, the splitting rules ought to choose the factor and value split which minimize the prediction error rate. The splitting rules described in (Riley, 1989) use a heuristic which approximates optimality by choosing at each node the split which minimizes the prediction error rate on the training set of data.

Stopping rules are necessary for terminating the splitting process at each internal node. Thus, these rules govern the height of each subtree. To determine the best tree, this implementation uses two sets of stopping rules. The first set is extremely conservative, resulting in an overly large tree. The problem with this type of tree is that it usually lacks the generality necessary to account for data outside of the training set. To compensate for this difficulty, the implementation uses a second set of rules. First, a sequence of subtrees is formed. Then, each tree is grown on a sizable fraction of the training data and tested on the remaining portion. This step is repeated until the tree has been grown and tested on all of the data. By performing this step, the stopping rules have access to cross-validated error rates for each subtree. The subtree with the lowest rates then defines the stopping points for each path in the full tree. Trees presented below all represent cross-validated data.

The prediction rules work in a straightforward manner to add the necessary labels to the terminal nodes. In the case of a continuous variable, the rules calculate the mean of the data points that have been classified together at that node. In the case of a categorical variable, the rules simply choose the class that occurs most frequently among the data points. The success of these rules can be measured through estimates of deviation. In this implementation, the deviation for continuous variables is the sum of the squared error for the observations. The deviation for categorical variables is simply the number of misclassified observations.

4.4 Procedure and Results

In analyzing boundary locations in our data, we have two goals in mind. First, we want to discover the extent to which boundaries can be predicted, given information which can be generated automatically from the text of an utterance. Second, we want to learn how much predictive power can be gained by including additional sources of information which, at least currently, cannot be generated automatically from text. That is, how well can automatically inferable information be used to predict intonational boundaries — and how much better can we do using additional information available currently only through human observation, such as information about other intonational features. So, for example, while we can predict pitch accent information with reasonable success from text (Hirschberg, 1990), what is the relationship between the use of such approximated information and the actual observations in permitting inference of boundary location? In effect, how good a substitute is approximate information for these purposes? In discussion of our results below, we will compare predictions based upon automatically inferable information with those based upon observed data.

One decision which must be made in labeling intonational boundaries is how to treat speaker hesitations, false starts, and other disfluencies when determining boundary location. While such phenomena may be accompanied by many of the acoustic correlates of true prosodic boundaries, it is not clear theoretically that they should be included as true intonational boundaries for purposes of description or prediction. Although disfluencies themselves may be to some extent predictable in terms of syntactic configuration or surface position (Hindle, 1983), there is no evidence that regularities they may exhibit will be similar to the regularities observable in the location of intonational boundaries. Since this issue is difficult to resolve, we perform our analyses on two data sets: one in which disfluencies are included as boundaries and one in which they are not.

We employ four different sets of variables during the first phase of analysis. The first data set (corresponding to the results in Figure 6) includes observed phonological information on pitch accent and on location of prior intonational boundaries, as well as automatically inferred information. Including observed values for accent and prior boundary location allows us to obtain a clear look at the contributions of the automatically generated analysis information by avoiding errors introduced by estimation of the phonological parameters. The success rate of boundary prediction for this data set is extremely high. The tree illustrated in Figure 6 correctly classifies 3330 out of 3677 potential boundary sites — an overall success rate of 90%. Furthermore, the number of decisions in this tree is only five. Thus, the tree represents a clean and simple model of phrase boundary prediction rules, assuming accurate phonological information about pitch accent and prior boundary location.

Turning to the decision nodes represented in the tree itself, we immediately note the importance of length of current phrase compared with that of prior phrase in boundary location. This variable alone (assuming that the boundary site occurs before the end of the utterance) allows us to correctly classify 2403 out of 2556 potential boundary sites. We conclude that occurrence of a phrase boundary is extremely unlikely if its presence would result in a phrase with less than half the length of the preceding phrase.

The first and last decision points in the tree are easily the most trivial. The information garnered from the first split is that utterances virtually always end with a boundary — rather unsurprising news. Similarly, the last split in the tree shows the importance of distance from the beginning of the utterance in boundary location. According to this node, boundaries are more likely to occur (subject to the constraints found higher in the tree) when more than $2\,\frac{1}{2}$ seconds have elapsed from the start of the utterance.

Both internal tree splits bear much more theoretical importance than do the first and last splits. The third node in the tree indicates that when w_i and w_j are co-constituents in a nominal category (including proper and common nouns at different bar levels), they are usually both members of a single intonational phrase as well. In other words, noun phrases form a tightly bound unit intonationally.

The fourth split in the tree shows the role of accent context in determining phrase boundary location. The accenting of w_i is clearly an important factor. If w_i is not accented, then it is unlikely that a phrase boundary will occur after it.

The significance of accenting in the phrase boundary classification tree leads to the question of whether or not predicted accents will have a similar impact on the paths of the tree. In the second analysis, we substituted predicted accent values for the actual values (see Figure 6). Interestingly, the success rate of the tree remains approximately the same, at 90%. However, the number of splits in the tree increases to nine and fails to include the accenting of w_i as a node in the decision tree. A closer look at the accent predictions themselves reveals that the majority of misclassifications come from function words that precede a boundary. Although the accent prediction algorithm we used to generate accent predictions expects these words to be cliticized, they are accented in this context. This phenomenon appears to be an artifact of the corpus; such accented function words generally occur before relatively long pauses in an utterance. Nevertheless, it is interesting to note that the classification tree is able to compensate quite well for the loss of the accent information. An implication of this result is that accent values may be predicted by some of the same factors that predict phrase boundaries. Preliminary investigation indicates that both pitch accent and boundary location are sensitive to location of prior intonational boundaries as well as part-of-speech tags and syntactic constituency of context.

In the third analysis, we eliminate the dynamic boundary percentage measure. The result remains nearly as good as before, with a success rate of 89%. The proposed decision tree (see Figure 6) confirms the usefulness of information about the accent status of w_i in boundary prediction. By itself (again assuming that the potential boundary site occurs before the end of the utterance), this factor can account for 1590 out of 1638 potential boundary site classifications. Similarly, this tree confirms the strength of the intonational ties among the components of noun phrases. In this tree, 536 out of 606 potential boundary sites receive final classification from this feature.

We conclude this phase of our analysis by producing a classification tree that uses text-based information alone. For this analysis we use predicted accent values instead of observed values and omit boundary distance percentage measures. Using binary-valued accented predictions (are $< w_i, w_j >$ accented or not), we obtain a success rate for boundary prediction of 89%, and using a four-valued distinction for predicted accented (cliticized, deaccented, accented, 'NA') we increased this slightly to 90%. The tree in Figure 6) presents results of the latter analysis.

The tree in Figure 6 contains more nodes than those appearing in previous figures; more sources of information are used to obtain a similar classification percentage. Interestingly, predicted accent is used only trivially here, to indicate sentence-final boundaries (*ra*='NA'), a function performed in figures 6-6 by the variable indicating distance of potential boundary site from end of utterance (*et*). The second split in the tree is also based upon temporal distance — of boundary site from beginning of utterance. Together these measurements correctly predict nearly forty percent of the data (38.2%). This tree next makes use of a variable which has not appeared in earlier trees — the part-of-speech of w_j. The tree shows that, where w_j is a function word other than 'to,' 'in,' or a conjunction (true for about half of potential boundary sites), in the majority of cases (88%) a boundary does not occur. Part-of-speech of w_i and type of constituent dominating w_i but not w_j are further used to classify these items.

This portion of the classification appears to support the notion of 'function word group' used commonly in assigning prosody in text-to-speech, in which phrases are defined, roughly, from one function word to the next. Overall rate of the utterance and type of utterance are made use of in the tree, in addition to part-of-speech and constituency information and distance of potential boundary site from beginning and end of utterance. In general, results of this first stage of analysis suggest — encouragingly — that there is considerable redundancy in the features predicting boundary location; when some features are unavailable, others can be used with similar rates of success.

Above we have seen that locating phrase boundaries from which disfluencies are excluded can be accomplished fairly successfully by calculating distance measures for each potential boundary site $< w_i, w_j >$, determining the utterance's overall rate, and making reference to part-of-speech and constituence information for w_i and w_j. We now must determine whether or not these variables remain important when we include 'boundaries' which arise from disfluency.

In the second phase of this study, we repeat the previous analyses, but extend our definition of a phrase boundary to include hesitations and disfluencies (see Figures 6-6). The result is a drop in success rate. The first analysis is successful 85% of the time, as compared with 90%. Similarly, the rates for the second analysis drop from 90% to 83%; for the third analysis from 89% to 85%; and for the fourth analysis from 89% to 83%. However, the same factors which are important in the first phase of the study continue to be significant in the second phase. Major splitting nodes in the four analyses include proximity to the end of the utterance, accent value of w_i, occurrence of a function word as w_j, and membership of w_i and w_j in a noun phrase. This comparison confirms our belief that such disfluencies should not in fact be classified as true phonological boundaries, but should undergo separate analysis.

5 Discussion

The application of CART techniques to the problem of predicting and detecting phrasing boundaries not only provides a classification procedure for predicting intonational boundaries from text, but it increases our understanding of the importance of several among the numerous variables which might plausibly be related to boundary location. In order to confirm the generality of our analysis, we are extending our corpus to include all of the sentences in the ATIS database. If the differences among these trees and our current trees are slight, we can then proceed to test our predictions on sentences from other corpora. The fact that question type is a fairly insignificant factor in the classification of the present corpus' boundaries suggests that this generalization will in fact be possible.

In addition to the current set of variables we are using for boundary prediction, we will test additional factors in our extended study. Currently, we are investigating whether or not GENERALIZED MUTUAL INFORMATION scores can serve as indicators of intonational phrase boundaries (Magerman and Marcus, 1990). We will include distituency information which (Magerman and Marcus,

1990) employs to predict syntactic structure for predicting prosodic structure. We will also study whether simple NP-detection provides a viable substitute for the full syntactic constituency information we have employed to date.

We will also examine possible interactions among the statistically important variables which have emerged from our initial study. Although CART techniques have worked extremely well at classifying phrase boundaries, our results remain something of a "black box." For an initial study like ours, CART offers a simple and immediate overview of the phenomenon. However, CART's step-wise treatment of variables, optimization heuristics, and dependency on binary splits obscure the possible relationships that exist among the various factors. Now that we have discovered a set of variables which do well at predicting intonational boundary location, we must refine our understanding of just how these variables interact.

6 Acknowledgments

We thank Michael Riley for helpful discussions. Code implementing the CART techniques employed here was written by Michael Riley and Daryl Pregibon. Part-of-speech tagging employed Ken Church's tagger, and syntactic analysis used Don Hindle's parser, Fidditch.

References

Alshawi, Hiyan, David M. Carter, Jan van Eijck, Robert C. Moore, and Steve G.

type	type of question (WH, Yes/no, Indirect, Indirect and WH, WH and WH)
tt	total # seconds in utterance
tw	total # words in utterance
st	distance (in seconds) between start and w_j
et	distance (in seconds) between w_j and end
sw	distance (in words) between start and w_j
ew	distance (in words) between w_j and end
la	is w_i accented or not/ cliticized, deaccented, accented, NA
ra	is w_j accented or not/ cliticized, deaccented, accented, NA
per	[distance (in words) from last boundary] / [length (in words) of last phrase]
tper	[distance (in seconds) from last boundary] / [length (in seconds) of last phrase]
j1, j2, j3, j4	part-of-speech label for $w_{i-1}, w_i, w_j, w_{j+1}$, respectively[6]
	v = verb
	b = "to be" verb
	m = modifier
	f = function word
	n = noun
	p = preposition
	w = WH
fs, fl, fr	fidditch's label for, respectively:
	s = smallest constituent that dominates w_i and w_j
	l = largest constituent that dominates w_i but not w_j
	r = largest constituent that dominates w_j but not w_i
	m = modifier
	d = determiner
	v = verb
	p = preposition
	w = WH
	n = noun
	s = sentence
	f = function word

Table 1: Key to Phrase Boundary Tree Labels[7]

Figure 3:

Figure 4:

Figure 5:

Figure 6:

Figure 7:

Figure 8:

Figure 9:

Figure 10:

Pulman. 1988. Overview of the Core Language Engine. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, pages 1108–1115, Tokyo, Japan. Institute for New Generation Computer Technology.

Altenberg, Bengt. 1987. *Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion*, volume 76 of *Lund Studies in English*. Lund University Press, Lund.

Bachenko, J. and E. Fitzpatrick. 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3):155–170.

Bolinger, Dwight. 1989. *Intonation and Its Uses: Melody in Grammar and Discourse*. Edward Arnold, London.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Wadsworth & Brooks, Pacific Grove CA.

Church, K. W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin. Association for Computational Linguistics.

DARPA. 1990. *Proceedings of the Speech and Natural Language Workshop*, Hidden Valley PA, June. Morgan Kaufmann.

Gee, J. P. and F. Grosjean. 1983. Performance structure: A psycholinguistic and linguistic apprasial. *Cognitive Psychology*, 15:411–458.

Hindle, D. M. 1989. Acquiring disambiguation rules from text. In *Proceedings of the 27th Annual Meeting*, pages 118–125, Vancouver. Association for Computational Linguistics.

Hindle, Donald. 1983. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting*, pages 123–128, Cambridge MA. Association for Computational Linguistics.

Hirschberg, J. 1990. Accent and discourse context: Assigning pitch accent in synthetic speech. In *Proceedings of the Eighth National Conference*, pages 952–957, Boston. American Association for Artificial Intelligence.

Magerman, David M. and Mitchel P. Marcus. 1990. Parsing a natural language using mutual information statistics. In *Proceedings of AAAI-90*, pages 984–989. American Association for Artifical Intelligence.

Marcus, Mitchell P. and Donald Hindle. 1985. A computational account of extra categorial elements in Japanese. In S. Y. Kuroda, editor, *Papers presented at the First SDF Workshop in Japanese Syntax*, La Jolla. System Development Foundation.

Olive, J. P. and M. Y. Liberman. 1985. Text to speech – an overview. *Journal of the Acoustic Society of America, Suppl. 1*, 78(Fall):s6.

Ostendorf, M., P. Price, J. Bear, and C. W. Wightman. 1990. The use of relative duration in syntactic disambiguation. In *Proceedings of the Speech and Natural Language Workshop*, pages 26–31, Hidden Valley PA, June. DARPA, Morgan Kaufmann.

Pierrehumbert, Janet B. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, Massachusetts Institute of Technology, September. Distributed by the Indiana University Linguistics Club.

Riley, Michael D. 1989. Some applications of tree-based modelling to speech and language. In *Proceedings of the Speech and Natural Language Workshop*, Cape Cod MA, October. DARPA, Morgan Kaufmann.

Selkirk, E. O. 1984. Phonology and syntax: The relation between sound and structure. In T. Freyjeim, editor, *Nordic Prosody II: Proceedings of the Second Symposium on Prosody in the Nordic language*, pages 111–140, Trondheim. TAPIR.

Steedman, M. 1990. Structure and intonation in spoken language understanding. In *Proceedings of the 28th Annual Meeting*, pages 9–16, Pittsburgh. Association for Computational Linguistics.