USING TEXT ANALYSIS TO PREDICT INTONATIONAL BOUNDARIES

Julia Hirschberg
AT&T Bell Laboratories
2D-450 – 600 Mountain Avenue
Murray Hill NJ 07974 USA
julia@research.att.com

Abstract

Relating the intonational characteristics of an utterance to features inferable from its orthographic transcription is important both for speech recognition and for speech synthesis. Results are presented for predicting the location of intonational phrase boundaries in a corpus of spontaneous (elicited) speech from syntactic, temporal and other features inferred from simple text analysis of its transcription. Classification and Regression Tree (CART) techniques are employed to model the relationship between hand-labeled boundary phenomena and textual features. Results from an additional experiment using these prediction trees to distinguish correct strings from those incorrectly recognized by a speech recognizer are also reported.[1]

1. Introduction

Relating the intonational phrasing of an utterance to other features which can be inferred from its transcription is important for speech synthesis and for speech recognition. For synthesis, more natural intonational phrasing can be assigned if accurate text predictors of phrase boundaries are known. For recognition, durations can be calculated more accurately if likely phrase boundaries can be predicted from recognition hypotheses and hypotheses themselves can be filtered by matching text-based boundary predictions with acoustic evidence. To date, however, most attempts to predict boundary locations have failed to provide general, robust results.

Intuitively, intonational phrasing divides an utterance into meaningful 'chunks' of information (Bolinger, 1989). Variation in phrasing can change the meaning hearers assign to tokens of a given sentence. For example, interpretation of a sentence like '*Bill doesn't drink because he's unhappy.*' will vary, if it is uttered as one phrase (Bill does indeed drink — but the cause of his drinking is not his unhappiness) or as two (Bill does *not* drink — and the reason for his abstinence is his unhappiness).

To characterize this phenomenon phonologically, we adopt Pierrehumbert's theory of intonational description for English (Pierrehumbert, 1980). In this theory, two levels of phrasing are significant in English intonational structure. Both types 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.

While phrase boundaries are perceptual categories, they are generally 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, and lengthening of the final syllable in the phrase (sometimes accompanied by glottalization or devoicing of that 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.

Previous research on the location of intonational boundaries has largely focussed on the relationship between these boundaries and syntactic constituency. While current work acknowledges the role that semantic and discourse-level information play in boundary assignment, most authors assume that syntactic configuration provides the basis for prosodic 'defaults', which may be 'overridden' by semantic or discourse considerations. While most interest in boundary prediction has been focussed on synthesis (Gee and Grosjean, 1983; Bachenko and Fitzpatrick, 1990), currently there is considerable interest in predicting boundaries to aid recognition (Ostendorf et al., 1990; Steedman, 1990).

The most successful empirical studies in boundary location have investigated how phrasing can disambiguate potentially syntactically ambiguous utterances in read speech (Lehiste, 1973; Ostendorf et al., 1990). Analysis based on corpora of natural speech (Altenberg, 1987) have so far reported limited success — even when the availability of syntactic, semantic, and discourse-level information well beyond the capabilities of current NL systems is assumed to be available.[2] This paper reports results of recent experiments on the automatic prediction of intonational boundaries from transcriptions of spontaneous (elicited) speech; initial results were presented in (Wang and Hirschberg, 1991a; Wang and Hirschberg, 1991b). Syntactic, distance, and other variables obtained from simple text analysis of utterance transcriptions were correlated with hand-labeled prosodic information for the training set. Classification and Regression Tree (CART) techniques were employed to model the relationship between intonational boundary phenomena and features of the text. Success rates of just over 90% in predicting presence or absence of boundary were achieved, representing a major improvement over other attempts at boundary prediction from unrestricted text. Resulting decision trees have been used to assign intonational boundaries in synthetic speech and to distinguish correct from incorrectly recognized strings in a recognition experiment.

[1] Thanks to Chin-Hui Lee, Michael Riley, David Roe, David Talkin, and Jan van Santen for helpful discussions. Code implementing CART was written by Michael Riley and Daryl Pregibon. David Talkin provided alignment software for labeling the ATIS June 1990 test set, which was done using Talkin's WAVES software by Jill Burstein and Nava Shaked. Part-of-speech tagging employed Ken Church's tagger, and syntactic analysis used Don Hindle's parser, Fidditch. This work is based on earlier work done jointly with Michelle Q. Wang.

[2] Bachenko and Fitzpatrick (1990; 1991) classify 83.5-86.2% of boundary/null boundary data points correctly for a test set of 35 citation-form sentences; Ostendorf et al (1990) report 80-83% correct prediction of boundaries only on another test set of 35 citation-form sentences, and an average success rate of 87.4% correct (averaging percent correct for boundaries and for null boundaries) on a 23-sentence (386) word FM radio news story. (1987) developes rules based on hand-labeled syntactic and semantic information which correctly classify an average of 72% of boundaries (tone units) in 48 minutes of partly-read, partly spontaneous speech from a single speaker. (These figures are derived from Altenberg's report of 95% coverage, 93% success at the sentence level; 94% coverage, 70% success at the basic clause level; and 78% coverage, 80% success in expanded phrases. Testing and training were done on the same utterances.)

2. Corpus and Features Analyzed

The training corpus used in this analysis consisted of 298 utterances (24 minutes of speech from 26 speakers) from the speech data collected by Texas Instruments for the DARPA Air Travel Information System (ATIS) Spoken Language System evaluation task. In a Wizard-of-Oz simulation, subjects were asked to make travel plans for an assigned task, providing spoken input and receiving teletype responses. The test set was the DARPA June 1990 test data, consisting of 138 utterances from five subjects.

To prepare the data for analysis and testing, the speech was labeled by hand, for location and type of intonational boundary and presence or absence of pitch accent. Labeling was done from both the waveform and pitchtracks of each utterance. Although major and minor boundaries were distinguished during labeling, in the analysis presented below, these have been collapsed to a single category.

Data points included 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. There are 3677 such potential boundary sites in the training corpus. Feature values obtainable via automatic text analysis were considered, as well as phonological features (observed pitch accent and distance from previous observed boundary, inter alia) currently available only through hand labeling, to see whether performance improved when the decision procedure was given richer data sets.

Variables considered in training included temporal variables, such as utterance and phrase duration, and distance of the potential boundary from beginning and end of utterance. These distances were measured in seconds, as well as words. Phrase length has also been proposed (Gee and Grosjean, 1983; Bachenko and Fitzpatrick, 1990) as a determiner of boundary location, such that prosodic phrases may have roughly equal length. To capture this, elapsed distance from the last (actual) boundary to the potential boundary site was calculated and divided by the length of the last phrase encountered (again, measured in seconds as well as in words).

Syntactic structural variables were also considered, including simple part-of-speech information as well as higher-level syntactic constituency. The latter in particular, as noted above, has generally been considered a good predictor of prosodic phrasing (Gee and Grosjean, 1983; Selkirk, 1984; Marcus and Hindle, 1985; Steedman, 1990). It has been proposed that some constituents may be more likely than others to be internally separated by intonational boundaries, and that some syntactic constituent boundaries may be more or less likely to coincide with intonational boundaries. To test the former hypothesis, the class of the lowest node in the parse tree to dominate both w_i and w_j, was determined, using Hindle's parser, Fidditch (1989). To test the latter, 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 were identified. Word class has also been used often to predict boundary location, particularly in text-to-speech systems. 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. These possibilities were tested by examining part-of-speech in a window of four words surrounding each potential phrase break, using Church's part-of-speech tagger (1988).

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. Informal observation suggests that phrase boundaries are more likely to occur in some accent contexts than in others. For example, phrase boundaries seem to occur more often between accented words than between deaccented words. To test the correlation between accent and phrasing, observed pitch accent values of w_i and w_j for each $< w_i, w_j >$ were examined. In some experiments, predicted pitch accent values (obtained via procedures described in (Hirschberg, 1990)) were substituted for observed values, to see if performance degraded. Classification and Regression Tree (CART) techniques were then used to generate decision trees automatically from resulting feature vectors of values for these variables.

3. Classification and Regression Trees

CART (Brieman et al., 1984) techniques can be used to generate decision trees from sets of continuous and discrete variables by using sets of splitting rules, stopping rules, and prediction rules. These rules affect the internal nodes, subtree height, and terminal nodes, respectively. At each internal node, CART determines which factor should govern the forking of two paths from that node. Furthermore, CART must decide which values of the factor to associate with each path. Ideally, the splitting rules should choose the factor and value split which minimizes the prediction error rate. The splitting rules in the implementation employed for this study (Riley, 1989) approximate optimality by choosing at each node the split which minimizes the prediction error rate on the training data. In this implementation, all these decisions are binary, based upon consideration of each possible binary partition of values of categorical variables and consideration of different cut-points for values of continuous variables.

Stopping rules terminate the splitting process at each internal node. 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, which usually lacks the generality necessary to account for data outside of the training set. To compensate, the second rule set forms a sequence of subtrees. 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. The stopping rules thus have access to cross-validated error rates for each subtree. The subtree with the lowest rate then defines the stopping point for each path in the full tree. Trees described below all represent cross-validated data.

The prediction rules work in a straightforward manner to add the necessary labels to the terminal nodes. For continuous variables, the rules calculate the mean of the data points classified together at that node. For categorical variables, the rules 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. Results of Analysis

(Wang and Hirschberg, 1991a; Wang and Hirschberg, 1991b) reported results of initial boundary classification experiments on the ATIS sample. Approximately 96% of the training data was modeled by the best prediction trees. Cross-validated classification rates of just over 90% were achieved for trees grown using hand-labeled information, such as observed pitch accent values and distance from prior boundary. And the same cross-validated success rate was achieved when only automatically obtainable feature values were employed, indicating that variables inferrable from text performed just as well without the additional acoustic information – an encouraging result. We also found that the same level of performance could be obtained even without the automatically available but more resource-intensive syntactic constituency information available from the parser employed (Hindle, 1989).

The generalizability of these results was further tested by manually separating the data into training and test sets, training new decision trees on the training data and testing on the reserved data. Results were then compared to the cross-validated results for the original trees with corresponding feature sets. In no case was the difference between success rates for the hand-separated data and the cross-validated results greater than two percentage points. So it appears that the cross-validated predictions are reliable.

4.1 Boundaries vs. Null Boundaries

Since approximately 80% of the data points represent actual 'null boundaries', it was important to look at whether these data points were being predicted more successfully than data points which represented actual boundaries. That is, if 80% of data points are correctly classified as 'null boundary', then one can achieve 80% success simply by classifying every data point as such.[3] The decision tree most successful in classifying observed intonational boundaries does so correctly in about 80% of cases (see Table 1); this tree classifies 'null boundary‘ cases correctly in around 93% of cases. However, the tree which performs best at classifying null boundaries, with 99.2% successfully classified, classifies observed boundaries correctly only 62% of the time (see Table 2).

Table 1: Confusion Matrix for Best Boundary Classification

	Boundary	*NoBoundary*	*% Correct*
Boundary	895	231	79.5%
NoBoundary	187	2364	92.7%

Table 2: Confusion Matrix for Best Null Boundary Classification

	Boundary	*NoBoundary*	*% Correct*
Boundary	435	267	62.0%
NoBoundary	25	2950	99.2%

So, to correct for the imbalance in the data — and in prediction performance on boundaries vs. null boundaries — the average of the two success rates was taken to represent the overall success of a prediction tree. Thus, for the tree whose classification performance is given in Table 1, this score would be 86.1%;[4] in fact, this represents the best performance of the trees considered in the study under this new metric. So, the tree which classifies boundary data points most accurately also performs best overall. This tree uses some observed acoustic features (in particular, observed pitch accents values rather than predicted), as well as automatically inferrable feature values, and also includes disfluencies as boundary data points;[5] thus this tree was trained on more boundary data points than other trees. The best overall performance from automatically inferrable information alone is 81.7%, obtained when syntactic constituency is considered along with other variables; however, similar performance (over 80% average correct) can be obtained when constituency information is omitted as well. In sum, while cross-validated results of around 90% were obtained from the original analysis (Wang and Hirschberg, 1991a; Wang and Hirschberg, 1991b), normalized scores of ten percent less are probably more representative of the actual performance of the predictor trees.

As an alternative correction for the skewedness of the original training data, the null boundary data points were sampled to roughly the same size as the boundary data, bringing the total sample size to approximately 40% of the original. New trees were then trained on this balanced sample, using nine of the feature sets previously test, including sets with only automatically-inferrable features and other sets with acoustic features as well. In every case, prediction of observed boundaries improved while prediction of observed null boundaries declined, when compared to predictions made with the same set of features values on the full training corpus. The best mean score obtained was 91.1% correct (Table 3); however, it should be noted that the cross-validated score for this tree is only 82.2%.

Table 3: Confusion Matrix for Best Classification, Balanced Sample

	Boundary	*NoBoundary*	*% Correct*
Boundary	621	81	88.5%
NoBoundary	46	697	93.8%

4.2 Boundary Prediction in Recognition

To evaluate the potential usefulness of intonational boundary detection for speech recognition, an additional experiment was performed on a test set, the ATIS JUNE 1990 test set. For this task, only trees trained on the full 298-utterance sample from the ATIS TI training set were employed. Essentially, the goal was to determine whether, assuming that accurate acoustic information about boundary location is obtainable,[6] candidate strings can be ranked with respect to how closely boundary predictions made from a recognized or self-scoring string correspond to boundaries predicted from acoustic evidence.

To this end, preliminary recognition results were obtained for 98 of the June 1990 test sentences, such that recognized strings differed from correct strings in each case.[7] Predictions for location of prosodic phrasing were made for both the (mis)recognized string and the correct string, again using decision trees grown only on the original training data.

Several trees were tested, all of which made use only of automatically obtained feature values. Temporal location of each boundary for the predicted locations was determined from the word durations of the recognition stage for the recognized string and the self-scoring stage for the correct string; i.e., if a boundary was predicted from the string to lie between word w_i and w_j, the temporal location of that boundary was determined from the durations of words w_1-w_i. These sequences of temporal boundary locations were then compared to sequences of observed boundary locations from the hand-labeled utterances in two ways. First, the actual number of boundaries observed was compared to the number of boundaries predicted for recognized vs. correct string, under the assumption that the string for which number of predicted boundaries was closest to number of observed boundaries should be preferred. Second, the location of predicted vs. recognized boundaries was compared to observed boundaries so that in each case cumulative temporal distance of observed from predicted boundary was minimized for each string. The assumption here was that strings for which predicted boundaries appeared temporally close to observed boundaries should be preferred. These two measures — minimal difference in number of boundaries predicted vs. number of boundaries observed and minimal cumulative distance of predicted boundaries from observed boundaries — were then employed to select between the two candidate strings (recognized and correct) in each of the 98 cases.

Results for this experiment were quite encouraging: The first metric tested, minimal difference in number of predicted vs. number of observed boundaries, preferred the correct over the incorrect string in 64-92% of the comparisons, depending upon which of the prediction trees were employed. However, the second metric, smallest cumulative distance of predicted from observed boundaries, preferred the correct over the (mis)recognized string in all of the 98 cases — for each of the prediction trees tested.

5. Discussion

Initial results on the prediction of boundary locations on a 298-utterance sample from the ATIS TI corpus indicated that the presence or absence of intonational boundaries can be predicted with over 90% accuracy using only feature values obtainable automatically from text analysis. In this paper, the problem of null boundary/boundary distinction has been examined and a scoring mechanism proposed to accommodate it. Mean percent correct scores represent a simple alternative to traditional boundary location metrics, which commonly ignore insertions, and to standard insertion/deletion measures used, for example, in speech recognition. Under this metric the best trees grown on the full data set score 86.1% (using some observed as well

[3] The confusion matrices presented here and below over-estimate success rates slightly. CART cross-validated error is averaged over multiple trees. These figures are calculated from a tree whose cross-validated length is chosen on this basis. The tree itself varies a few percentage points from the average. So, percentages should be considered a best approximation.

[4] Again, this is calculated on an actual subtree whose cross-validated length is minimal in terms of classification error. The cross-validated average may vary by a few percentage points.

[5] It is not clear whether disfluencies should be taken to represent true intonational boundaries.

[6] An earlier experiment in locating boundaries from recognizer output alone on a different data set was successful in 92% of cases, using word durations and pausal duration. Other results reported in the literature are equally encouraging (Ostendorf et al., 1990).

[7] For this output, the recognizer (Lee et al., 1990) used 47 context independent models, the LL/BBN bigram model used for the DARPA February 1991 evaluation, and a 1065 word lexicon also used in the February 1991 evaluation.

as automatically inferrable data) and 81.7% correct (using only automatically inferrable data).

An alternative to this approach was also tested — downsampling the data set to balance occurrence of boundary and null boundary data points. As expected, trees grown on the balanced sample performed better at balancing boundary with null boundary predictions. The most successful of the nine trees trained on the balanced sample scored 91.1 mean percent correct, predicting 88.5% of boundary data points and 93.8% of null boundary sites correctly. This tree was grown using only automatically available information, including syntactic constituency. While general boundary/null boundary location is not improved by the availability of syntactic constituency information (Wang and Hirschberg, 1991b), boundary location may indeed be. Future analysis of the full ATIS TI training set will test this hypothesis.

Finally, potential applicability of prediction of intonational boundaries from text analysis for recognition, in ranking candidate sentences given acoustic information and N-best strings, was tested on the ATIS June 1990 test set. Results of two simple distance metrics used with recognition results for the June 1990 test set were extremely encouraging: in particular, minimal cumulative distance of predicted boundaries from boundaries observed from acoustic information served to distinguish correct from (mis)recognized string in all the 98 cases tested. Considering that the trees used in this experiment had not performed remarkably well at direct boundary prediction for the same utterances in the previous experiment, these results are even more surprising. Apparently even relatively poor predictive performance can still be useful in distinguishing incorrect from correct strings — assuming, of course, that good acoustic information concerning boundary location is obtainable. One suspects, however, that the more difficult task of ranking a set of hypotheses will require improved accuracy of boundary/ null boundary location. Next steps in this work thus include training the decision trees on the remainder of the ATIS TI utterances, which is currently being labeled, and testing acoustic indicators of boundary location on the ATIS database.

References

Bengt Altenberg. 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.

J. Bachenko and E. Fitzpatrick. 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*. To appear.

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

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

K. W. Church. 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.

E. M. Fitzpatrick.1991. Personal communication.

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

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

Julia Hirschberg. 1990. Assigning pitch accent in synthetic speech: The given/new distinction and deaccentability. In *Proceedings of the Seventh National Conference*, pages 952–957, Boston. American Association for Artificial Intelligence.

C.-H. Lee, L. R. Rabiner, R. Pieraccini, and J. Wilpon. 1990. Acoustic modeling for arge vocabulary speech recognition. *Computer Speech and Language*, 4:127–165, April.

I. Lehiste. 1973. Phonetic disambiguation of syntactic ambiguity. *Glossa*, 7:197–222.

Mitchell P. Marcus and Donald Hindle. 1985. A computational account of extra categorial elements in japanese. In *Papers presented at the First SDF Workshop in Japanese Syntax*. System Development Foundation.

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

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

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

E. Selkirk. 1984. *Phonology and Syntax*. MIT Press, Cambridge MA.

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

Michelle Q. Wang and Julia Hirschberg. 1991a. Predicting intonational boundaries automatically from text: The ATIS domain. In *Proceedings*. DARPA Speech and Natural Language Workshop, February.

Michelle Q. Wang and Julia Hirschberg. 1991b. Predicting intonational phrasing from text. In *Proceedings of ACL-91*, Berkeley.