# Probabilistic Parse Scoring Based on Prosodic Phrasing

*N. M. Veilleux and M. Ostendorf*

Boston University
44 Cummington St.
Boston, MA 02215

## ABSTRACT

The relative size and location of prosodic phrase boundaries provides an important cue for resolving syntactic ambiguity. In previous work, we have introduced an analysis/synthesis formalism for scoring parses in terms of the similarity between prosodic patterns recognized from a given utterance and synthesized for the hypothesized parse. This paper describes a new approach to the synthesis problem, as well as an alternative scoring criterion. Specifically, a decision tree is designed to predict prosodic phrase structure for a given syntactic parse, and the tree is used to compute a parse score, which now is the probability of the recognized break sequence. Unlike the rule-based synthesis algorithm used in the previous work, the decision tree can be automatically trained and can therefore be designed specifically for different speaking styles or task domains. In experiments with a corpus of ambiguous sentences spoken by FM radio announcers, we have achieved disambiguation performance similar to the rule-based algorithm, which is close to the performance of human subjects in perceptual experiments using the scoring algorithm with hand labeled breaks.

## 1. Introduction

Spoken language understanding is a difficult problem, in part because of the many ambiguities inherent in natural language. Syntactic ambiguity arises when a given expression can be described by more than one syntactic structure, and contributes substantially to the difficulty of the natural language processing problem. Several factors may be involved in resolving such ambiguities, including semantics, discourse and bias toward a specific syntactic structure. In spoken language, prosody, or the suprasegmental information in an utterance, is an important cue. Prosody is especially useful in automatic speech understanding, since computer representations of semantics and discourse are not as sophisticated as human knowledge.

Experiments have shown that listeners can resolve several types of syntactic ambiguities by using prosodic information [3, 6]. The results of Price *et al.* [6] indicated that human listeners could reliably select the intended meaning of two target syntactic structures (86% correct identification for six out of seven types of structural ambiguity). Of the prosodic patterns studied in that work, the relative size and location of phrase boundaries seemed to provide the principal cue for resolving ambiguities. Thus, it seems likely that automatically detected prosodic phrase breaks could be used in speech understanding systems to reduce syntactic ambiguity.

Assuming that prosodic cues can be detected automatically, there are many different ways in which prosody might be used in syntactic disambiguation for speech understanding. In earlier work [9], we proposed a scoring algorithm to rank candidate parses based on an analysis/synthesis method. In this approach prosodic patterns recognized from a given utterance are compared to those synthesized using a hypothesized syntactic parse from a set of possible parses. Specifically the method involves: (1) automatically predicting prosodic break locations for each candidate syntactic structure (synthesis); (2) automatically recognizing prosodic breaks in the spoken utterance (analysis); and (3) scoring each hypothesized parse according to a measure of the similarity between predicted and observed prosodic structure. The score can then be used to rank competing hypotheses, as in the experiments here, or used in combination with other knowledge sources to choose the correct sentence interpretation.

This algorithm originally used a rule-based synthesis algorithm together with a correlation measure of similarity between predicted and observed prosodic phrase structure. Here, we expand on this approach by presenting an alternative prediction and scoring method. Specifically, we replace the rule-based synthesis component with a stochastic model which uses a decision tree to predict prosodic phrase structure. The probability distributions at the leaves of the tree can be used to find the probability of an observed prosodic structure given a syntactic parse, and this probability is then the prosodic score for the hypothesized parse.

In the following section, we briefly describe the speech corpus and break index representation of prosodic phrase structure. Next, we review the synthesis and scoring components used in our previous parse scoring system

[9]. As an alternative, we introduce a probabilistic scoring algorithm based on a stochastic model of phrase structure. We then present experimental results for the different synthesis/scoring techniques using the task of automatically disambiguating confusable sentence pairs, comparing results to those of human listeners. Finally, we discuss the implications of the results for future work.

## 2. Corpora and Labeling

The experiments presented here are primarily based on the corpus of ambiguous sentences described in more detail in (Price *et al.*,1991)[6]. Four professional FM radio announcers were asked to read 35 pairs of sentences, where members of a pair were phonetically similar but associated with different syntactic structures and therefore different meanings. The sentences included five examples of each of seven types of structural ambiguity: parenthetical clauses vs. non-parenthetical subordinate clauses, appositions vs. attached noun (or prepositional) phrases, main clauses linked by coordinating conjunctions vs. a main clause and a subordinate clause, tag questions vs. attached noun phrases, far vs. near attachment of final phrase[1], left vs. right attachment of middle phrase, and particles vs. prepositions.

In addition to the ambiguous sentence corpus, we also used a corpus of radio news speech for training the new stochastic synthesis algorithm. The data consists of 14 news stories, six from one announcer and eight from a second announcer, both female, for a total of 211 sentences (4210 words). These news stories were used only for training the synthesis algorithm, so just the word transcriptions were used (no acoustic information). These sentences differ from the ambiguous sentence pairs (the test data) in several ways. The ambiguous sentences are, on the average, shorter (7.6 vs. 19.6 words) and have a flatter syntactic structure (4 vs. 7 levels). In addition, the ambiguous sentence pairs were designed to cover specific syntactic structures, some of which are not generally found in the FM radio news stories. For example, the fourteen radio stories contain no sentences with tag questions and only five examples of embedded sentences, although both of these structures are common in the ambiguous sentences.

The parse scoring algorithms described here are based on an integer "break index" representation of prosodic constituent structure. Each word boundary in an utterance is labeled with a break index from 0-6 that corresponds to a level in a constituent hierarchy, or equivalently to the amount of prosodic decoupling between words. A "0" represents the most tightly bound words, such as in clitic groups, while a "6" represents the prosodic break between sentences. The correspondence between this seven-level system of indices and various hierarchies proposed in the literature is discussed in depth in Wightman *et al.* [10]. For training and evaluating our algorithms, utterances have been hand-labeled according to this break indexing system. Sentences were also annotated with skeletal parses as part of a preliminary version of the University of Pennsylvania Treebank project [4].

[1] High vs. low attachment is probably more accurate syntactic terminology, but "far" vs. "near" is used in [6] as more descriptive.

## 3. Rule-Based Synthesis and Scoring

The focus of this paper is on the synthesis and scoring components of our prosodic parse scoring system. The rule-based synthesis and correlation scoring algorithms used in previous work are described below for reference. The analysis component, prosodic break recognition, is also based on classification trees and is described in [8].

### 3.1. Performance Structure Synthesis

Prosodic phrase break prediction algorithms have typically been rule-based. In our previous work, we investigated a variety of rule-based algorithms designed to predict performance structures, based on Gee and Grosjean's Phi algorithm [2]. Since results for these algorithms were similar [9], only the performance of the Phi algorithm will be used for comparison here.

Given the syntactic structure, the Phi algorithm iteratively groups successively larger prosodic constituents together, beginning at the word level, to form a binary tree. A break index, which indicates the relative coherence between constituents, is deterministically assigned after each word in the sentence according to node count in the tree. The absolute value of the breaks is not linguistically meaningful and, in addition, has no theoretical upper bound. We refer to these indices as $\phi$-breaks to distinguish them from the seven-level labeling system used in the analysis.

### 3.2. Correlation Score

In [9], a correlation score was used to compare the $\phi$-breaks for competing syntactic structures with observed breaks for some utterance to be disambiguated. Specifically, the score is simply an estimate of the correlation coefficient between observed and synthesized breaks.

An advantage of the correlation score is that it is invariant to linear transformations of the break indices. That is, a high-valued sequence of breaks will have the same interpretation as a low-valued sequence of breaks, if relatively higher and lower breaks are in the same position. A disadvantage is that it requires a hard decision from

the synthesis step, and therefore is limited in the amount of variability it allows.

## 4. Probabilistic Synthesis and Scoring

Although the performance-structure-based algorithms appear to be quite useful, rule-based methods have some disadvantages. First, they are difficult to implement for different styles or to tailor for a specific task domain, since the development of new rules is required. Second, they do not allow for the natural variability in phrasing observed in multiple spoken renditions of the same sentence. An alternative to the rule-based approach is to use a model that can be trained automatically, such as a stochastic model. The fact that stochastic models associate a probability with a sequence of break indices suggests a method of parse scoring based on the probability of a parse given the recognized breaks.

### 4.1. Decision Tree Break Synthesis

Existing break synthesis models that can be automatically trained are based on decision trees. Wang and Hirschberg [7] first proposed the use of decision trees to predict the presence or absence of a prosodic break, with very successful results. Although their experiments involved predicting only one type of break, the model is general and can be extended to predict an arbitrary number of levels.

Using a set of allowable questions about bracketed word sequences, binary decision trees partition the labeled training data into successively more homogeneous sets with respect to class distributions. Classes in our case are different levels of prosodic breaks (i.e., 0 - 6). Trees are designed using a greedy growing algorithm [1], which iteratively reduces the impurity of the class distributions at the leaves of the tree[2]. In this work, the size of the tree was determined based on complexity/performance trade-offs in the training set. At each terminal node of the tree, the training data defines a relative frequency estimate of the probability for each level of break represented. The tree can be used for synthesis by choosing the most probable break level at each terminal node, as in [7]. For the parse scoring application, however, the stochastic model can be used to find the probability of a break sequence given a hypothesized parse. Using the probabilities directly has potential performance advantages over making a hard decision on predicted parses before a subsequent scoring stage.

The decision tree was designed using the FM radio news stories, based on questions that used part-of-speech information, syntactic information and location in the sentence in terms of numbers of words. Part-of-speech information used was based only on capitalization and function word tables, and the questions about syntactic structure are based on Treebank skeletal parses [4]. The different syntactic units used are SBAR or SBARQ (declarative or question embedded sentences), S or SINV (declarative or inverted main clauses), NP, VP, PP, phrase beginning with Wh-question word, ADJP or ADVP. All questions are based on features derived from text information only; no acoustic information is used in the synthesis algorithm. The specific questions are detailed below; motivation for these can be found in [5], where a similar set of questions were investigated. Abbreviations in parentheses refer to labels used in the resulting tree, illustrated in Figure 1.

1. Is this a sentence boundary? (sent)
2. Is the left word is a content word and the right word a function word? (cw-fw)
3. Is the left word a function word and the right word a content word? (fw-cw)
4. What is the function word type of word to the right? I.e. conjunctions, articles, auxiliary verbs and models, pronouns, prepositions, and a default category. (fw-type)
5. Is either adjacent word capitalized, i.e. a proper name? (cap)
6. How many content words have occurred since the previous function word? (# cw)
7. How many function words have occurred since the previous function word? (# fw)
8. What is the relative location in the sentence? Specifically, what is the ratio of the number of orthographic words over the sentence length in words quantized to the nearest eighth? (s loc)
9. What is the largest syntactic unit that dominates the word preceding the potential boundary location and does not dominate the succeeding word? (dom lft)
10. What is the largest syntactic unit that dominates the word succeeding the potential boundary location and does not dominate the preceding word? (dom rt)
11. What is the smallest syntactic unit that dominates both? (dom both)
12. How many syntactic units end between the two words? (# ])
13. How many syntactic units begin between the two words? (# [)
14. What is the depth (number of levels from the top) in the syntax tree of the right word? Depth was measured as the number of open brackets minus the number of closed brackets. (depth)
15. What is the total number of initiating and terminating syntactic units between the two words? This number is roughly related to how far the juncture is from the bottom of the syntax tree. (height)

[2] Specifically, we use the Gini criterion $i(t) = \sum_{i \neq j} p(i|t)p(j|t)$ as a measure of impurity.

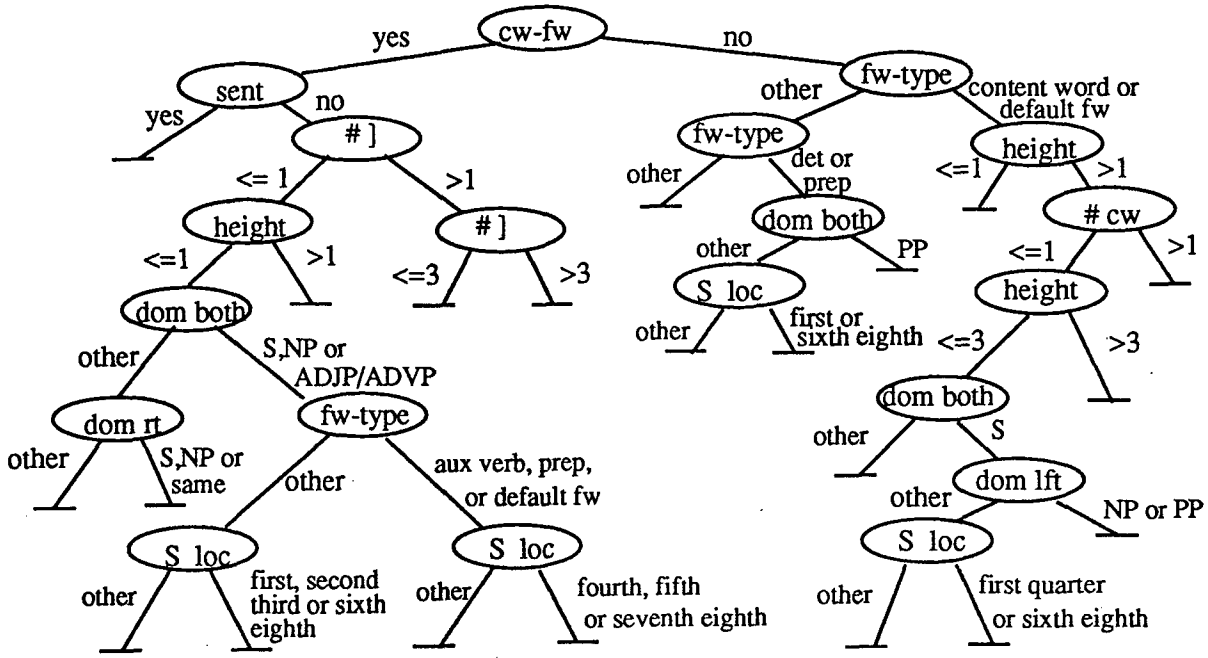


Figure 1: Synthesis decision tree used in parse scoring, designed on radio news text. The syntactic features of each word juncture determine assignment to a terminal node, which is associated with a probability distribution of breaks.

### 4.2. Probability Score

Use of the decision tree for break synthesis suggests an alternative approach to the correlation score, which is to compute the probability of the sequence of automatically labeled break indices conditioned on the hypothesized parse. The probability score is computed as follows. The text and the hypothesized parse are processed to generate a sequence of feature vectors $[x_1, \ldots, x_n]$, one for each word, which are subsequently each encoded by the tree to a node $t_i = T(x_i)$. The score of the observed breaks indices $[b_1, \ldots, b_n]$, is then

$$S_P = \frac{1}{n} \sum_{i=1}^{n} \log p(b_i|t_i),$$

where $p(b|t)$ is the distribution associated with terminal node $t$. The factor $1/n$ normalizes the score to account for differences in word length; otherwise, the score is biased to favor shorter sentences.

Rather than computing the probability of the sequence of breaks, we could have explicitly predicted a sequence of breaks from text, taking the most probable sequence, and then used the correlation scoring approach with these synthesized breaks. However, if the predicted break indices are the same for both parses, then it is impossible to distinguish them using the correlation score, though it was possible to choose between them using the probability of observed breaks. This phenomenon does occur for the ambiguous sentences and therefore correlation scoring has lower disambiguation performance than probability scoring. Of course, if two hypothesized syntactic structures result in the same node sequence, the sentence cannot be disambiguated with the tree. However, in our corpus of 35 pairs of ambiguous sentences, only two were assigned the same node sequence. The corresponding human productions for this sentence pair had similar ambiguities in break indices labeled for at least one of the four speakers.

## 5. Experiments

We have tested our analysis/synthesis approach by using it to perform the same task that the human subjects in [6] were asked to perform. Specifically, we attempt to select which of two interpretations was intended by the speaker by choosing the interpretation with the highest score. For each test utterance, we use an automatic break labeling algorithm to recognize the break indices (the algorithm used in [8] with additional acoustic features) under each of the two possible sentence hypotheses. The two break sequences are then scored according to a synthesis model using the syntactic structure of the corresponding sentence hypothesis. The candidate sentence having the highest score is selected. In the event of a tie, the first sentence in the pair is chosen. These experiments were repeated using both the the rule-based synthesis algorithm and the new decision tree algorithm. The Phi algorithm was evaluated in conjunction with the

correlation score, and the decision tree synthesis algorithm was used with a probabilistic score. In addition, we repeat the experiments using hand-labeled breaks in order to examine the performance of the synthesis module alone. The results of these experiments are summarized in Table 1 for each of the 14 types (7 pairs) of syntactic ambiguity, which list the percent of sentence correctly identified for each category out of a set of 20 sentences/category. For comparison, Table 1 also contains the results reported for the human subjects [6].

Not all syntactic differences can be disambiguated by prosodic information, and such cases obviously cannot be handled by our algorithms. For completeness, Table 1 includes results for all categories, although our analysis will focus mainly on the categories that were most reliably identified by human listeners (those for which mean response minus standard deviation was greater than chance, indicated with an asterisk in the table). In addition, this analysis will ignore the main-main vs. main-subordinate clause category, since in [6], the sentences were found to be very similar prosodically.

The results based on the hand-labeled break indices show that the decision tree synthesis algorithm in combination with a probabilistic score gave disambiguation accuracy similar to the Phi algorithm, and comparable to performance of human listeners on this test subset.

When using automatically labeled breaks rather than hand-labeled breaks, there is significant degradation in performance for both the Phi and decision tree algorithms. The biggest loss in performance was for the particle category, which was correctly identified with the hand-labeled breaks but identified at the level of chance using the automatically labeled breaks. In this case, automatically detected prominence information may prove to be useful, because particles are often prominent whereas prepositions are not [6].

When correlation is used as the similarity measure, the decision tree performance degrades about 10% in accuracy, e.g., from 74% to 64% with automatically labeled breaks and on the reliable categories. Clearly the probabilistic score is preferable. However, the fact that the accuracy of the decision tree when used with the correlation score is much lower than that for the performance structure algorithms, suggests that some improvement is possible in the tree synthesis algorithm.

## 6. Discussion

In summary, we have introduced a decision tree synthesis algorithm and probability-based scoring method for use in an analysis/synthesis formalism. We have evaluated this new probabilistic synthesis/scoring mechanism on a set of 70 ambiguous sentences, each spoken by four radio announcers, and have compared performance to the rule-based synthesis algorithm and correlation scoring previously investigated. The performance structure (rule-based) synthesis algorithm and the probabilistic decision tree approach gave similar results. Considering only eight categories of structures that could be disambiguated by humans with high reliability (out of fourteen categories investigated), the algorithms achieve disambiguation performance comparable to human listeners when scoring hand-labeled break indices (89-91% accuracy). However, as in the case of the rule-based algorithms, performance degrades to 73-74% accuracy when scoring automatically labeled break indices.

The decision tree result is very encouraging, given the significant differences between the training and test data. Since the decision tree can be easily retrained for specific applications, performance should improve with training based on a larger and more representative sample of sentences. Moreover, the decision tree synthesis method could also be improved through the use of new questions and more detailed part-of-speech labels. The question set used here was originally chosen to classify only intermediate and intonational phrases, and new questions about factors that are correlated with the lower level breaks might be particularly useful additions.

The parse scoring algorithm on hand-labeled data shows some loss in accuracy relative to human performance if we also consider the sentences that were less reliably identified by the human listeners. Several different factors probably account for this effect, including the fact that these sentences simply exhibit more variability. It is also likely that humans are using other prosodic cues in addition to phrase breaks to resolve ambiguities, such as phrasal prominence. This additional information could be incorporated using the analysis/synthesis approach with a probabilistic synthesis model that predicts both breaks and prominences.

Using the parse scoring algorithms with automatically labeled breaks incurs a significant loss in disambiguation performance. While it is possible that further improvements in the detection algorithm may be successful, using the break detection algorithm jointly with the probabilistic synthesis model in scoring a parse may also improve performance.

While these results are encouraging, there are several issues that may affect performance in a spoken language system. First, the syntactic parses were hand corrected. Second, the sentences here represent a narrow range of syntactic classes and performance outside of this set may vary. Finally, the analysis component used phone seg-

| Ambiguity | Hand Labels | | Machine Labels | | Human |
|---|---|---|---|---|---|
| | & G-G | & T-prob | & G-G | & T-prob | Perception |
| *+ Parenthetical* | 60 | 70 | 60 | 35 | 77 |
| *- Parenthetical* | 90 | 70 | 60 | 70 | 96* |
| *+ Apposition* | 90 | 100 | 95 | 90 | 92* |
| *- Apposition* | 60 | 70 | 30 | 50 | 91* |
| *Main-Main* | 55 | 60 | 85 | 85 | 88 |
| *Main-Subordinate* | 50 | 65 | 70 | 60 | 54 |
| *+ Tag* | 90 | 100 | 90 | 90 | 95* |
| *- Tag* | 70 | 80 | 55 | 55 | 81 |
| *Far Attach* | 100 | 65 | 70 | 65 | 78 |
| *Near Attach* | 40 | 70 | 40 | 55 | 63 |
| *Left Attach* | 100 | 80 | 85 | 80 | 94* |
| *Right Attach* | 100 | 95 | 90 | 75 | 95* |
| *Particle* | 100 | 100 | 55 | 55 | 82* |
| *Preposition* | 95 | 95 | 80 | 80 | 81* |
| Average | 79 | 80 | 69 | 68 | 84 |
| Average for * | 91 | 89 | 73 | 74 | 91 |

Table 1: Percent correct disambiguation as a function of syntactic ambiguity for: different synthesis algorithms compared to hand-labeled breaks (G-G: Gee/Grosjean, T-prob: decision tree synthesis); different synthesis algorithms compared to automatically labeled breaks; and human perceptual results. Those categories which were identified by human listeners with significant reliability are marked with asterisks. Percentages are based on 5 sentences from each of 4 speakers, giving 20 utterances in each category and 280 utterances total.

mentations from a recognizer constrained to the correct word sequence. While these issues need to be investigated, it is possible that use of prosodic parse scoring may help overcome and not be limited by problems in other components of a spoken language system. For example, it is possible that using a prosodic parse score would enhance the overall performance of the system because recognition errors would yield low probability break index sequences.

## 7. Acknowledgments

The authors gratefully acknowledge Colin Wightman for his contributions to the foundation of this work and for his automatic break detection results and Patti Price and Stefanie Shattuck-Hufnagel for their valuable suggestions and insights. This research was jointly funded by NSF and DARPA under NSF grant number IRI-8905249, and by NSF under grant number IRI-8805680.

## References

1. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone C. J. (1984). *Classification and Regression Trees.* Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA.
2. Gee, J. P. & Grosjean, F. (1983). Performance Structures: A Psycholinguistic and Linguistic Appraisal. *Cognitive Psychology 15,* 411–458.
3. Lehiste, I. (1973). Phonetic Disambiguation of Syntactic Ambiguity. *Glossa 7,* 2, 107–121.
4. Marcus, M. P. & Santorini, B. Building a very large natural language corpora: The Penn treebank. Submitted manuscript.
5. Ostendorf, M. & Veilleux, N. A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location. Submitted manuscript.
6. Price, P., Ostendorf, M., Shattuck-Hufnagel, S. & Fong, C. (1991). The Use of Prosody in Syntactic Disambiguation. *Journal of the Acoustical Society of America 90,* 6, 2956–2970.
7. Wang, M. & Hirschberg, J. (1992). Automatic Classification of Intonational Phrase Boundaries. *Computer Speech and Language,* to appear.
8. Wightman, C. W. & Ostendorf, M. (1991). Automatic Recognition of Prosodic Phrases. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* 321–324.
9. Wightman, C. W., Veilleux, N. M. & Ostendorf, M. (1991). Using Prosodic Phrasing in Syntactic Disambiguation: An Analysis-by-Synthesis Approach. *Proceedings of the DARPA Workshop on Speech and Natural Language,* 384–389.
10. Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. & P. Price. (1992). Segmental Durations in the Vicinity of Prosodic Phrase Boundaries. *Journal of the Acoustical Society of America* March 1992.