The Use of Relative Duration in Syntactic Disambiguation

M. Ostendorf† P. J. Price‡ J. Bear‡ C. W. Wightman†

† Boston University
44 Cummington St.
Boston, MA 02215

‡ SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025

Abstract

We describe the modification of a grammar to take advantage of prosodic information automatically extracted from speech. The work includes (1) the development of an integer "break index" representation of prosodic phrase boundary information, (2) the automatic detection of prosodic phrase breaks using a hidden Markov model on relative duration of phonetic segments, and (3) the integration of the prosodic phrase break information in SRI's Spoken Language System to rule out alternative parses in otherwise syntactically ambiguous sentences. Initial experiments using ambiguous sentences read by radio announcers achieved good results in both detection and parsing. Automatically detected phrase break indices had a correlation greater than 0.86 with hand-labeled data for speaker-dependent models; and, in a subset of sentences with preposition ambiguities, the number of parses was reduced by 25% with a simple grammar modification.

Introduction

"Prosody," the suprasegmental information in speech, i.e., information that cannot be localized to a specific sound segment, can mark lexical stress, identify phrasing breaks and provide information useful for semantic interpretation. Although all of these aspects may be useful in spoken language systems, particularly important are prosodic phrase breaks which can provide cues to syntactic structure to help select among competing hypotheses, and thus help to disambiguate otherwise ambiguous sentences. In speech understanding applications, information such as prosody that aids disambiguation, is particularly important, since speech input, as opposed to text, introduces a vast increase in the amount of ambiguity a parser must face. For example, Harrington and Johnstone [7] found that even when all phonemes are correctly identified, the indeterminancy of phoneme sequences when word boundaries are unknown yields in excess of 1000 word string parses for many of their 4 to 10 word sentences. Moreover, these estimates rise dramatically as indeterminancy is introduced in the phoneme sequences: only 2% of the same sentences had fewer than 1,000 parses when phonetically similar phonemes were clustered together (e.g., voiced stops). This indeterminancy vastly increases the work for a parser.

The work reported here focuses on the use of relative duration of phonetic segments in the assignment of syntactic structure, assuming a known word sequence. Specifically, relative duration of phonemes is estimated by a speech recognizer constrained to recognize the correct string of words. These duration values are then used to compute phrase break indices, which are in turn used to rule out alternative parses in otherwise syntactically ambiguous sentences. In this paper, we begin by providing some theoretical background on prosodic phrase breaks, and by describing a numerical representation of phrase breaks for use in speech understanding. Next we describe algorithms for automatic recognition of these break indices from speech for known text, and the modification of a grammar to use these indices in parsing. Finally, we present experimental results based on ambiguous sentences from three speakers, showing that the use of prosodic information can significantly reduce the number of candidate syntactic parses by pruning mismatches between prosody and syntax.

Prosodic Phrase Breaks

In recent years, there have been significant advances in the phonology of prosodic phrases. While this work is not yet explicit enough to provide rules for automatically determining the prosodic phrase boundaries in speech, it is useful as a foundation for our computational models. Several researchers in linguistics have proposed hierarchies of prosodic phrases [9, 12, 10]. Although not all levels of these hierarchies are universally accepted, our data appear to provide evidence for prosodic words as individual words or clitic groups, groupings of prosodic words, intermediate phrases, intonational phrases, groupings of intermediate phrases as in parenthetical phrases, and sentences. Since it is not clear how many of these levels will be useful in speech understanding, we have represented all seven possible types of boundaries, but focus initially on the information in the highest levels, sentences and intonational phrases.

In order to utilize this information in a parser, we developed a numerical representation of this hierarchy

using a sequence of break indices between each word. A *break index* encodes the degree of prosodic decoupling between neighboring words. For example, an index of 0 corresponds to cliticization, and an index of 6 represents a sentence boundary.[1] We found that these break indices could be labeled with very high consistency within and across labelers. We anticipate that the strongest boundaries (highest level of the hierarchy) will be both easiest to detect and most useful in parsing, and will refer to these boundaries (4-6) as major phrase boundaries.

The acoustic cues to prosodic breaks vary according to the different levels of the hierarchy. For example, there are phonological rules that apply across some boundaries of the hierarchy but not others (e.g., [12, 18]). Intonation cues mark intermediate and intonational phrase boundaries [15, 3]. Pauses and/or breaths may also mark major phrase boundaries. Our initial data indicate that duration lengthening is a fairly reliable cue to phrase boundaries. Prepausal lengthening is observed in English and many, but not all, other languages (see [16] for a summary). Lengthening without a pause has been observed in English in clause-final position (e.g., [5]) and in phrase-final position, though lengthening of phones at the ends of words is disputed in the literature (see [8]).

The typical representation of syntactic structure is not identical to prosodic phrase structure.[2] There is some work suggesting methods for predicting rhythmic structure from syntactic structure, but not vice versa [13, 6]. These results are complicated by the complex relationship between intonation and rhythm and the influence of semantics on intonation. Although people disagree on the precise relationship between prosody and syntax, it is generally agreed that there is some relationship. We have shown, for example, that prosody is used by listeners to choose the appropriate meaning of otherwise ambiguous sentences with an average accuracy of 86% [11]. Although listener performance varied as a function of the type of syntactic ambiguities, reliability was high for left vs. right attachment and particle/preposition ambiguities, both difficult problems in parsing. An example illustrates how the break indices resolve syntactic and word sequence ambiguities for two phonetically identical sentences:

- Marge 0 would 1 never 2 deal 0 in 2 any 0 guys 6

- Marge 1 would 0 never 0 deal 3 in 0 any 0 guise 6

Note that the break index between "deal" and "in" provides an indication of how tightly coupled the two words are. For "in" as a particle, we expect tighter connection to the preceding verb whose meaning is modified than to the following phrase which is the object of that verb. For "in" as a preposition, we expect a tighter con-

nection to the following object of the preposition than to the preceding verb.

Use of Phrase Breaks in Parsing

With the goal of using prosodic phrase breaks to reduce syntactic ambiguity in a parser, we have developed an algorithm for automatically computing break indices, and we have modified the structure of the grammar to incorporate this information. The current effort is focused on demonstrating the feasibility of this approach; therefore the problem is restricted in scope. The techniques we have developed are extensible to more general cases; the positive results encourage us to relax some of these restrictions. In the next section we describe the basic approach, which is based on the following simplifications.

- We assume knowledge of the orthographic word transcription of a sentence and the sentence boundary. Word boundary ambiguities complicate the integration architecture and were not considered in this study, although this is an important issue to address in future work.

- Only relative duration is currently used as a cue for detecting the break indices, though we expect to improve performance in later versions of the algorithm by utilizing other acoustic cues such as intonation.

- Only preposition ambiguities were investigated. The focus on preposition ambiguities was motivated by the following facts:

  1. Prepositions are very frequent: 80-90% of the sentences in our radio news database, in the resource management sentences, and in the. ATIS database contain at least one prepositional phrase.

  2. Sentences with prepositions are usually syntactically ambiguous.

  3. Our perceptual experiments suggest that prosody could be used effectively in many sentences with preposition ambiguities. We believe that the techniques developed can be adapted to more general attachment ambiguities.

- The sentences examined were read by professional FM radio announcers, so the prosody is somewhat more exaggerated and probably easier to recognize than it would be with nonprofessional speakers.

Phrase Break Detection

Using a known word sequence as a tightly constrained grammar, a speech recognizer can be used to provide time alignments for the phone sequence of a sentence. We have used the speaker-independent SRI DECIPHER system [17], which uses phonological rules to generate bushy pronunciation networks that provide a more accurate phonetic transcription and alignment.

---

[1] We have changed our scale from our previously reported six levels to the present seven levels to achieve a more natural mapping between our levels and those described in the linguistics literature.

[2] Steedman, however, [14] had made some interesting arguments concerning the appropriateness of categorial grammar for reflecting prosodic structure.

In earlier work [11], each phone duration was normalized according to speaker- and phone-dependent means and variances. Raw break features were generated by averaging the normalized duration over the final syllable coda of each word and adding a pause factor. Lexically stressed and non-stressed vowels were separated in the computation of means and variances. Finally, these features were normalized (relative to the observed phone durations in the current sentence) to obtain integer break indices with a range of 0 to 5. These indices had a high correlation with hand-labeled data (0.85) and were successfully used in a parser to reduce the number of syntactic parses by 25% [2]. However, the normalization algorithm required knowledge of the raw break features for the entire sentence, which has the disadvantages that the algorithm is non-causal and may not reflect sentence-internal changes in speaking rate. In addition, the algorithm used the full scale of break indices, so every sentence was constrained to have at least one 0 break index.

A new algorithm, using a hidden Markov model (HMM), was investigated for computing the break indices from the raw break features described above. The algorithm is not strictly causal (in the same sense that the HMM recognizer is not causal - decisions are not made until sometime after a word has been observed), but does not require any ad hoc scaling. We anticipate the time delay associated with HMM prosody decoding to be similar to delays associated with a speech recognizer. In a slight variation from previous work, the raw break indices were computed from the rhyme (vowel nucleus + coda) of the final syllable instead of the coda alone. This change did not have an effect on the correlation with hand labels. A second, more important difference is that the phone duration means are adapted according to a local speaking rate. Local speaking rate is given by the average normalized durations over the last M phones, excluding pauses, where $M = 50$ was determined experimentally. The mean duration for each phone is adjusted with each new observed phone according to

$$\hat{\mu}_\alpha = \mu_\alpha + \sigma r/N$$

where r is the speaking rate, N is a feedback coefficient that is equal to 5000 at steady state, but varies at start-up for faster initial adaptation, σ is the standard deviation of the phone's duration, unadapted, and μ_α represents the mean duration for phone α.

A fully connected seven-state HMM is used to recognize break indices, given the raw break feature. Each HMM state corresponds to a break index (state number = break index) and the output distribution in each state describes the raw indices observed while in that state. In this work, we investigated the use of Gaussian output distributions of the scalar break feature, but joint use of several features in multivariate output distributions will best utilize the power of the HMM approach. Viterbi decoding was used to obtain the state sequence for an utterance, corresponding to the break index sequence.

The parameters of the break HMM were estimated in two different ways, involving either supervised or unsupervised training. By supervised training, we mean that the hand-labeled break indices are given, so the state sequence is fully observable and simple maximum likelihood estimation (as opposed to the Estimate-Maximize, or forward-backward, algorithm) is used. In unsupervised training, no hand-labeled data is used. Mean output distributions of the states are initialized to values on a scale that increases with the corresponding break index, and the transition probabilities were initialized to be essentially uniform. The forward-backward algorithm was then run, effectively clustering the states, to estimate the final output distribution parameters. A surprising and very encouraging result was that the unsupervised HMM correlated as well with the hand-labeled data as did the HMM with supervised parameter estimates.

Integration With a Parser

The question of how best to incorporate prosodic information into a grammar/parser is a vast area of research. The methodology used here is a novel approach, involving automatic modification of the grammar rules to incorporate the break indices as a new grammatical category. We modified an existing, and reasonably large grammar, the grammar used in SRI's spoken language system. The parser used is the Core Language Engine developed at SRI in Cambridge.

Several steps are involved in the grammar modification. The first step is to systematically change all of the rules of the form $A \rightarrow B\ C$ to the form $A \rightarrow B\ Link\ C$, where $Link$ is a new grammatical category, that of the prosodic break indices. Similarly all rules with more than two right hand side elements need to have $Link$ nodes interleaved at every juncture, e.g., a rule $A \rightarrow B\ C\ D$ is changed into $A \rightarrow B\ Link_1\ C\ Link_2\ D$.

Next, allowance must be made for empty nodes, denoted ϵ. It is common practice to have rules of the form $NP \rightarrow \epsilon$ and $PP \rightarrow \epsilon$ in order to handle wh-movement and relative clauses. These rules necessitate the incorporation into the modified grammar of a rule $Link \rightarrow \epsilon$; otherwise, the sentence will not parse, because an empty node introduced by the grammar will either not be preceded by a link, or not followed by one.

The introduction of empty links needs to be constrained to avoid the introduction of spurious parses. If the only place the empty NP or PP could go is at the end of the sentence, then the only place the empty $Link$ could go is right before it and no extra ambiguity is introduced. However, if an empty wh-phrase could be posited at a place somewhere other than the end of the sentence, then there is ambiguity as to whether it is preceded or followed by the empty link.

For instance, for the sentence, "What did you see on Saturday?" the parser would find both of the following possibilities:

- What L did L you L see L empty-NP empty-L on L Saturday?

- What L did L you L see empty-L empty-NP L on L Saturday?

Hence the grammar must be made to automatically rule out half of these possibilities. This can be done by constraining every empty link to be followed immediately by an empty wh-phrase, or a constituent containing an empty wh-phrase on its left branch. It is fairly straightforward to incorporate this into the routine that automatically modifies the grammar. The rule that introduces empty links gives them a feature-value pair: *empty-link* $=$ y. The rules that introduce other empty constituents are modified to add to the constituent the feature-value pair: *trace-on-left-branch* $=$ y. The links 0 through 6 are given the feature-value pair *empty-link* $=$ n. The default value for *trace-on-left-branch* is set to n so that all words in the lexicon have that value. Rules of the form $A_0 \rightarrow A_1 \, Link_1 \ldots A_n$ are modified to insure that A_0 and A_1 have the same value for the feature *trace-on-left-branch*. Additionally, if $Link_i$ has *empty-link* $=$ y then A_{i+1} must have *trace-on-left-branch* $=$ y. These modifications, incorporated into the grammar modifying routine, suffice to eliminate the spurious ambiguity.

Additional changes to the grammar were necessary to actually make use of the prosodic break indices. In this initial endeavor, a very conservative change was made after examining the break indices on a set of sentences with preposition ambiguities. The rule $N \rightarrow N \, Link \, PP$ was changed to require the value of the link to be between 0 and 2 inclusive for the rule to apply. A similar change was made to the rule $VP \rightarrow V \, Link \, PP$, except that the link was required to have the value of either 0 or 1.

Experimental Results

We have achieved encouraging results both in detection of break indices and in their use in parsing. The automatic detection algorithm yields break labels having a high correlation with hand-labeled data for the various algorithms described. In addition, when we chose a subset (14) of these sentences exhibiting prepositional phrase attachment ambiguities or preposition/particle ambiguities, we found that the incorporation of the prosodic information in the SRI grammar resulted in a reduction of about 25% in the number of parses, without ruling out any correct parses. For sentences to which the prosodic constraints on the rules actually applied, the decrease in number of parses was about 50%. In many cases the use of prosodic information allowed the parser to correctly identify a unique parse. Below we describe the results in more detail.

Corpus

The first corpus we examined consisted of a collection of phonetically ambiguous, structurally different pairs of sentences. The sentence pairs were read by three female professional radio announcers in disambiguating contexts. In order to discourage unnatural exaggerations

	Correlation with Hand Labels		
Speaker	SD super.	SD unsup.	SI unsup.
F1A	0.89	0.88	0.89
F2B	0.86	0.87	0.85

Table 1: Average correlation between automatically labeled break indices and hand-labeled break indices, using different methods of training. SD,SI = speaker-(in)dependent; super. = supervised training with hand-labeled data; unsup. = unsupervised training.

of any differences between the sentences, the materials were recorded in different sessions with several days in between. In each session only one sentence of each pair occurred. Seven types of structural ambiguity were investigated: parentheticals, apposition, main-main versus main-subordinate clauses, tags, near versus far attachment, left versus right attachment, and particles versus prepositions. Each type of ambiguity was represented by five pairs of sentences.

Detection Algorithm

In finding break indices for the ambiguous sentence pairs, the seventy sentences were concatenated together as though the speaker read them as a paragraph. Concatenation allowed the algorithm to avoid initialization for every sentence, but since the speaking rate is then tracked across several sentences that were not actually read in connection, there was probably some error associated with estimating the speaking rate factor. The HMM was used to generate break indices, and the results were evaluated according to how highly correlated the automatically generated labels were with the hand-labeled data. The correlation reported here is the average of the sample correlation for each sentence. The experiments yielded good accuracy on the detected break labels, but also some important results on unsupervised training and speaker-independence.

In comparing the supervised and unsupervised parameter estimation approaches for the HMM, we found that both yielded break indices with similar correlation to the hand labeled indices (Table 1). In addition, the indices obtained using the two training approaches were very highly correlated with each other (> 0.92). This is a very important result, because it suggests that we may be able to automatically estimate models without requiring hand-labeled data. Results for speaker-dependent training on two speakers are summarized in Table 1.

For the moment, we are mainly interested in detecting major phrase breaks (4-6) and not the confusions between these levels, since the parser uses major breaks as constraints on grammar rules. Using supervised parameter estimation, the false rejection/false acceptance rates are 14%/3% for speaker F1A and 21%/6% for speaker F2B. The unsupervised parameter estimation algorithm has a bias towards more false rejections and

fewer false acceptances. The most important confusions were between minor phrase breaks (2,3) and intonational phrases (4). Since a boundary tone is an important cue to an intonational phrase, we expect performance to improve significantly when intonation is included as a feature.

In the experiments comparing supervised to unsupervised training, speaker-dependent phone means and variances were estimated from the same data used to train the HMM as well as to evaluate the correlation, because of the limited amount of speaker-dependent data available. Though the speaker-dependent experiments were optimistic in that they involved testing on the training data, the results are meaningful in the sense that other speaker-independent experiments showed the parameters were robust with respect to a change in speakers. Using unsupervised training with two speakers to estimate both HMM parameters and duration means and variances for normalization for a different speaker, the correlation of the resulting automatically detected break indices with the hand-labeled indices was close to the speaker-dependent case (Table 1). Also, the speaker-dependent predictions and speaker-independent predictions were highly correlated with each other (0.96). We conclude that, at least for these radio news announcers, the algorithm seems to be somewhat robust with respect to different speakers. Of course, the news announcers had similar reading styles, and the hand-labeled data for two speakers had a correlation of 0.94.

Overall, the HMM provided improvement over the previously reported algorithm, with a correlation of 0.90 compared to 0.86 for six levels. On the other hand, there was a small reduction in the correlation when using seven levels of breaks (0.87) compared to six levels (0.90).

Use in Parsing

A subset of 14 sentences with preposition ambiguities was chosen for evaluating the integration of the break indices in the parser. We evaluated the results by comparing the number of parses obtained with and without the prosodic constraints on the grammar rules, and noted the differences in parse times. On average, the incorporation of prosody resulted in a reduction of about 25% in the number of parses found, with an average increase in parse times of 37%. The fact that parse times increase is due to the way in which prosodic information is incorporated. The parser does a certain amount of work for each word, and the effect of adding break indices to the sentence is essentially to double the number of words that the parser must process. It may be possible to optimize the parser to significantly reduce this overhead.

The sentences were divided into those to which the additional constraints would apply (type 'a') and those about which the constraints had nothing to say (type 'b'). Essentially the constraints block attachment if there is too large a break index between a noun and a following prepositional phrase or between a verb and

| | | Number of Parses | | Parse Time | |
Sentence	Humans % correct	No Pros.	With Pros.	No Pros.	With Pros.
1a	81	10	4	5.3	5.3
2a	94	10	7	3.6	4.3
3a	94	2	1	2.3	2.7
4a	87	2	1	3.2	4.7
5a	100	2	1	1.7	2.5
6a	56	2	1	2.5	2.8
7a	100	2	1	0.8	1.3
TOTAL	87	30	16	19.4	23.6

Table 2: Sample of sentences to which the added constraints applied. Parse times are in seconds

a following particle. Thus the 'a' sentences had more major prosodic breaks at the sites in question than did the 'b' sentences.

The results, shown in Tables 2 and 3, indicate that for the 'a' sentences the number of parses was reduced, in many cases to a unique parse. The 'b' sentences, as expected, showed no change in the number of parses. No correct parses were eliminated through the incorporation of prosodic information.

This corpus was also used in perceptual experiments to determine which types of syntactic structures humans could disambiguate using prosody. It is interesting to note that in many cases, sentences which were automatically disambiguated using the added constraints were also reliably disambiguated by humans. The fact that the perceptual results and parsing results are not more correlated than they are may be due to the fact that humans use other prosodic cues such as prominence, in addition to duration, for disambiguation.

Discussion

We are encouraged by these initial results and believe that we have found a promising and novel approach for

| | | Number of parses | | Parse Time | |
Sentence	Humans % correct	No Pros.	With Pros.	No Pros.	With Pros.
1b	61	10	10	5.3	7.7
2b	81	10	10	3.6	4.0
3b	75	2	2	2.3	3.7
4b	94	2	2	3.2	5.5
5b	100	2	2	1.6	2.9
6b	78	2	2	2.5	4.1
7b	100	2	2	0.8	1.5
TOTAL	84	30	30	19.3	29.4

Table 3: Sample of sentences to which the added constraints did not apply. Parse times are in seconds.

incorporating prosodic information into a natural language processing system. The break index representation of prosodic phrase levels is a useful formalism which can be fairly reliably detected and can be incorporated into a parser to rule out prosodically inconsistent syntactic hypotheses.

The results reported here represent only a small study of integrating prosody and parsing, and there are many directions in which we hope to extend the work. In detection, integrating duration and intonation cues offers the potential for a significant decrease in the false rejection rate of major phrase boundaries, and previous work by Butzberger on boundary tone detection [4] provides a mechanism for incorporating intonation. As for integration with the parser, investigation of other types of structural ambiguity should lead to similar improvements in the reduction of the number of parses. Finally, we hope to verify and extend these results by considering a larger database of speech and as well as the prosody of nonprofessional speakers. We are already evaluating the techniques on the ATIS database.

Acknowledgements

References

[1] H. Alshawi, D. M. Carter, J. van Eijck, R. C. Moore, D. B. Moran, F. C. N. Pereira, S. G. Pulman and A. G. Smith (1988) *Research Programme in Natural Language Processing: July 1988 Annual Report,* SRI International Technical Note, Cambridge, England.

[2] J. Bear and P. J. Price (1990) "Prosody, Syntax and Parsing," *Proceedings of the ACL Conference.*

[3] M. Beckman and J. Pierrehumbert (1986) "Intonational Structure in Japanese and English," *Phonology Yearbook 3,* ed. J. Ohala, pp. 255-309.

[4] J. Butzberger (1990) *Statistical Methods for Analysis and Recognition of Intonation Patterns in Speech,* M.S. Thesis, Boston University.

[5] J. Gaitenby (1965) "The Elastic Word," *Status Report on Speech Research SR-2,* Haskins Laboratories, New Haven, CT, 1-12.

[6] J. P. Gee and F. Grosjean (1983) "Performance Structures: A Psycholinguistic and Linguistic Appraisal," *Cognitive Psychology,* Vol. 15, pp. 411-458.

[7] J. Harrington and A. Johnstone (1987) "The Effects of Word Boundary Ambiguity in Continuous Speech Recognition," *Proc. of XI Int. Congr. Phonetic Sciences,* Tallin, Estonia, Se 45.5.1-4.

[8] D. Klatt (1975) "Vowel Lengthening is Syntactically Determined in a Connected Discourse," *J. Phonetics 3,* 129-140.

[9] M. Y. Liberman and A. S. Prince (1977) "On Stress and Linguistic Rhythm," *Linguistic Inquiry 8,* 249-336.

[10] D. R. Ladd, (1986) "Intonational Phrasing: the Case for Recursive Prosodic Structure," *Phonology Yearbook,* 3:311-340.

[11] P. Price, M. Ostendorf and C. Wightman (1989) "Prosody and Parsing," in *Proc. Second DARPA Workshop on Speech and Natural Language,* October 1989, pp.5-11.

[12] E. Selkirk (1980) "The Role of Prosodic Categories in English Word Stress," *Linguistic Inquiry,* Vol. 11, pp. 563-605.

[13] E. Selkirk (1984) *Phonology and Syntax: The Relation between Sound and Structure,* Cambridge, MA, MIT Press.

[14] M. Steedman (1989) "Intonation and Syntax in Spoken Language Systems," presented at the BBN Natural Language Symposium.

[15] L. Streeter, "Acoustic Determinants of Phrase Boundary Perception," *Journal of the Acoustical Society of America,* Vol. 64, No. 6, pp. 1582-1592 (1978).

[16] J. Vaissiere (1983) "Language-Independent Prosodic Features," in *Prosody: Models and Measurements,* ed. A. Cutler and D. R. Ladd, pp. 53-66, Springer-Verlag.

[17] M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin and D. Bell (1989) "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* pages 699–702, Glasgow, Scotland.

[18] A. Zwicky (1970) "Auxiliary Reduction in English," *Linguistic Inquiry,* Vol. 1, 323-336.