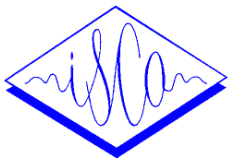

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

The Use of Relative Duration in Syntactic Disambiguation

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

† 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. Automatically detected phrase break indices had a correlation of greater than 0.8 with hand-labeled data for speaker-dependent and independent models; and in a subset of sentences with preposition ambiguities, the number of parses was reduced by 25% with a simple grammar modification.

1. INTRODUCTION

"Prosody", the suprasegmental information in speech, 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 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.

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, duration of each phone 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.

2. 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, e.g., [3,6,4]. Although not all levels of these hierarchies are universally accepted, our data appear to provide evidence for: prosodic words (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 (sentence, intonational phrase).

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. 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 break indices, while influenced by several cues, are closely related to the prepausal lengthening observed in each word. Prepausal lengthening describes the tendancy of speakers to increase the duration of the phones occuring just prior to a major boundary, and is observed in English and many, but not all, other languages (see Vaissiere [8] for a summary). Our initial data indicate that duration lengthening is a fairly reliable cue to phrase boundaries.

The typical representation of syntactic structure is not identical to prosodic phrase structure.[1] 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% [5]. 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 connection to the following object of the preposition than to the preceding verb.

3. 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 focussed on demonstrating the feasibility of this approach, and 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. Below 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.
- 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 sentences were read by professional FM radio announcers who have a clear and consistant speaking style.

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), and (2) sentences with prepositions are usually syntactically ambiguous; and (3) our perceptual experiments suggested that prosody could be used effectively in many sentences with preposition ambiguities.

[1]Steedman [7] has proposed a syntactic structure that directly follows the prosodic structure of an utterance. However, most existing parsers are not based on this approach, so we have not yet considered the use of our algorithms in this formalism.

3.1 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 [9], which has the advantage that the use of phonological rules can generate bushy pronunciation networks that provide a more accurate phonetic transcription and alignment.

A new algorithm, using a hidden Markov model, was investigated for computing the break indices from "Raw" break features generated by averaging the normalized phone durations over the rhyme (final syllable vowel nucleus and coda) of each word and adding a pause factor. 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 50 phones, excluding pauses. 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 ($N = 100$ at steady state, but varies at start-up for faster initial adaptation), σ is the standard deviation of the phone's duration (not adapted), 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 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. The forward-backward algorithm was then run, effectively "clustering" the states, to estimate the final output distribution paramaters. For both algorithms, the transition probabilities were initialized to be essentially uniform.

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.

3.2 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 already existing, and in fact reasonably large grammar: the Core Language Engine developed at SRI in Cambridge [1].

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 be of 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 so as not to introduce spurious parses. 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.

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 be either 0 or 1.

4. EXPERIMENTAL RESULTS

We have achieved encouraging results in both detection of break-indices and in their use in parsing. The automatic detection algorithm yields break labels which have 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 found, without ruling out any correct parses.

4.1 Corpus

The corpus we examined consisted of a collection 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 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 are included: 1) parentheticals, 2) apposition, 3) main-main vs. main-subordinate clauses, 4) tags, 5) near vs. far attachment, 6) left vs. right attachment, and 7) particles vs. prepositions. Each type of ambiguity was represented by five pairs of sentences.

4.2 Detection Algorithm

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: 0.87 for supervised estimation and 0.88 for unsupervised. The correlation between the two results was 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.

For the moment, we are mainly interested in detecting major phrase breaks (4-6) and not the confusions between these levels. Using supervised MLE parameter estimation, the false rejection rate is 14% and the false detection rate is 3%. 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 with supervised vs. unsupervised training, speaker-dependent phone means and variances were estimated from the same data that was used to train the HMM as well as evaluate the correlation (because of the limited amount of speaker-dependent data available). Though this experiment was unfair in that it involved testing on the training data, the results are meaningful in the sense that other experiments showed the parameters were robust to a change in speakers. Using two speakers to train 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 (0.85 - 0.88). Also, the speaker-dependent predictions and speaker-independent predictions were highly correlated with each other (0.96). We conclude that, at least for ra-

dio news announcers, the algorithm seems to be somewhat robust to different speakers.

4.3 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 grammar rules for break indices, and noted the difference in parse time associated with the added rules. On average, the incorporation of prosody resulted in a reduction of about 25% in the number of parses found.

	number of parses			
	type A		*type B*	
sent. i.d.	*no pros.*	*with pros.*	*no pros.*	*with pros.*
1	10	4	10	10
2	10	7	10	10
3	2	1	2	2
4	2	1	2	2
5	2	1	2	2
6	2	1	2	2
7	2	1	2	2
TOTAL	30	16	30	30

Table 1: Number of parses found with and without using prosody.

The sentences were divided into those to which the rules added to the grammar would apply (type 'a') and those about which the rules had nothing to say (type 'b'). Essentially the rules block attachment if there is too large a break index between a noun and a following prepositional phrase or between a verb and a following particle. Thus the 'a' sentences had more major prosodic breaks at the site in question than did the 'b' sentences.

The results, shown in Table 1, indicate that for the 'a' sentences the number of parses was reduced, in many cases from 2 to 1. The 'b' sentences, as expected, showed no change in the number of parses. No correct parses were eliminated through the incorporation of prosodic information.

V. DISCUSSION

We are encouraged by these initial results and believe that we have found a promising and novel approach for 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 [2] provides a mechanism for incorporating intonation. In integration with the parser, investigating more syntactic categories which would require additional rules in the grammar can expand the scope of these results. Finally, we hope to verify and extend these results by considering a larger database of speech and as well as the prosody of non-professional speakers.

VI. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the help and advice from Stefanie Shattuck-Hufnagel, for her role in defining the prosodic break representation, and Hy Murveit from SRI, for his help in generating the phone alignments. This research was jointly funded by NSF and DARPA under NSF grant number IRI-8905249, and in part be DARPA under the Office of Naval Research contract N00014-85-C-0013.

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

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

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

[5] 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.

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

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

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

[9] 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.