PROSODY, SYNTAX AND PARSING

John Bear
and
Patti Price
SRI International
333 Ravenswood Avenue
Menlo Park, California 94025

Abstract

We describe the modification of a grammar to take advantage of prosodic information provided by a speech recognition system. This initial study is limited to the use of relative duration of phonetic segments in the assignment of syntactic structure, specifically in ruling out alternative parses in otherwise ambiguous sentences. Taking advantage of prosodic information in parsing can make a spoken language system more accurate and more efficient, if prosodic-syntactic mismatches, or unlikely matches, can be pruned. We know of no other work that has succeeded in automatically extracting speech information and using it in a parser to rule out extraneous parses.

1 Introduction

Prosodic information can mark lexical stress, identify phrasing breaks, and provide information useful for semantic interpretation. Each of these aspects of prosody can benefit a spoken language system (SLS). In this paper we describe the modification of a grammar to take advantage of prosodic information provided by a speech component. Though prosody includes a variety of acoustic phenomena used for a variety of linguistic effects, we limit this initial study to the use of relative duration of phonetic segments in the assignment of syntactic structure, specifically in ruling out alternative parses in otherwise ambiguous sentences.

It is rare that prosody alone disambiguates otherwise identical phrases. However, it is also rare that any one source of information is the *sole* feature that separates one phrase from all competitors. Taking advantage of prosodic information in parsing can make a spoken language system more accurate and more efficient, if prosodic-syntactic mismatches, or unlikely matches, can be pruned out. Prosodic structure and syntactic structures are not, of course, completely identical. Rhythmic structures and the necessity of breathing influence the prosodic structure, but not the syntactic structure (Gee and Grosjean 1983, Cooper and Paccia-Cooper 1980). Further, there are aspects of syntactic structure that are not typically marked prosodically. Our goal is to show that at least some prosodic information can be automatically extracted and used to improve syntactic analysis. Other studies have pointed to possibilities for deriving syntax from prosody (see e.g., Gee and Grosjean 1983, Briscoe and Boguraev 1984, and Komatsu, Oohira, and Ichikawa 1989) but none to our knowledge have communicated speech information directly to a parser in a spoken language system.

2 Corpus

For our corpus of sentences we selected a subset of a corpus developed previously (see Price *et al.* 1989) for investigating the perceptual role of prosodic information in disambiguating sentences. A set of 35 phonetically ambiguous sentence pairs of differing syntactic structure was recorded by professional FM radio news announcers. By phonetically ambiguous sentences, we mean sentences that consist of the same string of phones, i.e., that suprasegmental rather than segmental information is the basis for the distinction between members of the pairs. Members of the pairs were read in disambiguating contexts on days separated by a period of several weeks to avoid exaggeration of the contrast. In the earlier study listeners viewed the two contexts while hearing one member of the pair, and were asked to select the appropriate context for the sentence. The results showed that listeners can, in general, reliably separate phonetically and syntactically ambiguous sentences on the basis of prosody. The original study investigated seven types of structural ambiguity. The present study used a subset of the sentence pairs which contained

prepositional phrase attachment ambiguities, or particle/preposition ambiguities (see Appendix).

If naive listeners can reliably separate phonetically and structurally ambiguous pairs, what is the basis for this separation? In related work on the perception of prosodic information, trained phoneticians labeled the same sentences with an integer between zero and five inclusive between every two words. These numbers, 'prosodic break indices,' encode the degree of prosodic decoupling of neighboring words, the larger the number, the more of a gap or break between the words. We found that we could label such break indices with good agreement within and across labelers. In addition, we found that these indices quite often disambiguated the sentence pairs, as illustrated below.

- Marge 0 would 1 never 2 deal 0 in 2 any 0 guys
- Marge 1 would 0 never 0 deal 3 in 0 any 0 guise

The break indices between 'deal' and 'in' provide a clear indication in this case whether the verb is 'deal-in' or just 'deal.' The larger of the two indices, 3, indicates that in that sentence, 'in' is not tightly coupled with 'deal' and hence is not likely to be a particle.

So far we had established that naive listeners and trained listeners appear to be able to separate such ambiguous sentence pairs on the basis of prosodic information. If we could extract such information automatically perhaps we could make it available to a parser. We found a clue in an effort to assess the phonetic ambiguity of the sentence pairs. We used SRI's DECIPHER speech recognition system, constrained to recognize the correct string of words, to automatically label and time-align the sentences used in the earlier referenced study. The DECIPHER system is particularly well suited to this task because it can model and use very bushy pronunciation networks, accounting for much more detail in pronunciation than other systems. This extra detail makes it better able to time-align the sentences and is a stricter test of phonetic ambiguity. We used the DECIPHER system (Weintraub *et al.* 1989) to label and time-align the speech, and verified that the sentences were, by this measure as well as by the earlier perceptual verification, truly ambiguous phonetically. This meant that the information separating the member of the pairs was not in the segmental information, but in the suprasegmental information: duration, pitch and pausing. As a byproduct of the labeling and time alignment, we noticed that the durations of the phones could be used to separate members of the pairs. This was easy to see in phonetically ambiguous sentence pairs: normally the structure of duration patterns is obscured by intrinsic duration of phones and the contextual effects of neighboring phones. In the phonetically ambiguous pairs, there was no need to account for these effects in order to see the striking pattern in duration differences. If a human looking at the duration patterns could reliably separate the members of the pairs, there was hope for creating an algorithm to perform the task automatically. This task could not take advantage of such pairs, but would have to face the problem of intrinsic phone duration.

Word break indices were generated automatically by normalizing phone duration according to estimated mean and variance, and combining the average normalized duration factors of the final syllable coda consonants with a pause factor. Let $\tilde{d}_i = (d_i - \mu_j)/\sigma_j$ be the normalized duration of the ith phoneme in the coda, where μ_j and σ_j are the mean and standard deviation of duration for phone j. d_p is the duration (in ms) of the pause following the word, if any. A set of word break indices are computed for all the words in a sentence as follows:

$$n = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \tilde{d}_i + d_p/70$$

The term $d_p/70$ was actually hard-limited at 4, so as not to give pauses too much weight. The set $\mathcal{A}$ includes all coda consonants, but not the vowel nucleus unless the syllable ends in a vowel. Although the vowel nucleus provides some boundary cues, the lengthening associated with prominence can be confounded with boundary lengthening and the algorithm was slightly more reliable without using vowel nucleus information. These indices n are normalized over the sentence, assuming known sentence boundaries, to range from zero to five (the scale used for the initial perceptual labeling). The correlation coefficient between the hand-labeled break indices and the automatically generated break indices was very good: 0.85.

3 Incorporating Prosody Into A Grammar

Thus far, we have shown that naive and trained listeners can rely on suprasegmental information to separate ambiguous sentences, and we have shown that we can automatically extract information that correlates well with the perceptual labels. It remains to be shown how such information can be used by a parser. In order to do so we modified an already existing, and in fact reasonably large grammar. The parser we

use is the Core Language Engine developed at SRI in Cambridge (Alshawi *et al.* 1988).

Much of the modification of the grammar is done automatically. The first thing is to systematically change all 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. It is common practice to have rules of the form $NP \rightarrow \varepsilon$ and $PP \rightarrow \varepsilon$ in order to handle wh-movement and relative clauses. These rules necessitate the incorporation into the modified grammar of a rule $Link \rightarrow \varepsilon$. Otherwise, a sentence such as a wh-question will not parse because an empty node introduced by the grammar will either not be preceded by a link, or not be followed by one.

The introduction of empty links needs to be constrained so as not to introduce spurious parses. If the only place the empty NP or PP etc. could fit into the sentence is at the end, then the only place the empty Link can go is right before it so there is no extra ambiguity 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 zero through five 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.

sent i.d.	*# parses no prosody*	*# parses with prosody*	*parse time no prosody*	*parse time with prosody*
1a	10	4	5.3	5.3
1b	10	10	5.3	7.7
2a	10	7	3.6	4.3
2b	10	10	3.6	4.0
3a	2	1	2.3	2.7
3b	2	2	2.3	3.7
4a	2	1	3.2	4.7
4b	2	2	3.2	5.5
5a	2	1	1.7	2.5
5b	2	2	1.6	2.9
6a	2	1	2.5	2.8
6b	2	2	2.5	4.1
7a	2	1	0.8	1.3
7b	2	2	0.8	1.5
TOT.	60	46	38.7	53.0

Table 1: The number of parses and parse times (in seconds) with and without the use of prosodic information.

4 Setting Grammar Parameters

Running the grammar through our procedure, to make the changes mentioned above, results in a grammar that gets the same number of parses for a sentence with links as the old grammar would have produced for the corresponding sentence without links.

In order to make use of the prosodic information we still need to make an additional important change to the grammar: how does the grammar use this information? This area is a vast area of research. The present study shows the feasibility of one particular approach. In this initial endeavor, we made the most conservative changes imaginable after examining the break indices on a set of sentences. We changed the rule $N \rightarrow N\ Link\ PP$ so that the value of the link must be between 0 and 2 inclusive (on a scale of 0-5) for the rule to apply. We made essentially the same change to the rule for the construction verb plus particle, $VP \rightarrow V\ Link\ PP$, except that the value of the link must, in this case, be either 0 or 1.

After setting these two parameters we parsed each of the sentences in our corpus of 14 sentences, and compared the number of parses to the number of parses obtained without benefit of prosodic information. For half of the sentences, i.e., for one member of each of the sentence pairs, the number of parses remained the same. For the other members of the pairs, the number of parses was reduced, in many cases from two parses to one.

The actual sentences and labels are in the appendix. The incorporation of prosody resulted in a reduction of about 25% in the number of parses found, as shown in table 1. Parse times increase about 37%.

In the study by Price *et al.*, the sentences with more major breaks were more reliably identified by the listeners. This is exactly what happens when we put these sentences through our parser too. The large prosodic gap between a noun and a following preposition, or between a verb and a following preposition provides exactly the type of information that our grammar can easily make use of to rule out some readings. Conversely, a small prosodic gap does not provide a reliable way to tell which two constituents combine. This coincides with Steedman's (1989) observation that syntactic units do not tend to bridge major prosodic breaks.

We can construe the large break between two words, for example a verb and a preposition/particle, as indicating that the two do not combine to form a new slightly larger constituent in which they are sisters of each other. We cannot say that no two constituents may combine when they are separated by a large gap, only that the two smallest possible constituents, i.e., the two words, may not combine.

To do the converse with small gaps and larger phrases simply does not work. There are cases where there is a small gap between two phrases that are joined together. For example there can be a small gap between the subject NP of a sentence and the main VP, yet we do not want to say that the two words on either side of the juncture must form a constituent, e.g., the head noun and auxiliary verb.

The fact that parse times increase is due to the way in which prosodic information is incorporated into the text. 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. We expect that this overhead will constitute a less significant percentage of the parse time as the input sentences become more complex. We also hope to be able to reduce this overhead with a better understanding of the use of prosodic information and how it interacts with the parsing of spoken language.

5 Corroboration From Other Data

After devising our strategy, changing the grammar and lexicon, running our corpus through the parser, and tabulating our results, we looked at some new data that we had not considered before, to get an idea of how well our methods would carry over. The new corpus we considered is from a recording of a short radio news broadcast. This time the break indices were put into the transcript by hand. There were twenty-two places in the text where our attachment strategy would apply. In eighteen of those, our strategy or a very slight modification of it, would work properly in ruling out some incorrect parses and in not preventing the correct parse from being found. In the remaining four sentences, there seem to be other factors at work that we hope to be able to incorporate into our system in the future. For instance it has been mentioned in other work that the length of a prosodic phrase, as measured by the number of words or syllables it contains, may affect the location of prosodic boundaries. We are encouraged by the fact that our strategy seems to work well in eighteen out of twenty-two cases on the news broadcast corpus.

6 Conclusion

The sample of sentences used for this study is extremely small, and the principal test set used, the phonetically ambiguous sentences, is not independent of the set used to develop our system. We therefore do not want to make any exaggerated claims in interpreting our results. We believe though, that we have found a promising and novel approach for incorporating prosodic information into a natural language processing system. We have shown that some extremely common cases of syntactic ambiguity can be resolved with prosodic information, and that grammars can be modified to take advantage of prosodic information for improved parsing. We plan to test the algorithm for generating prosodic break indices on a larger set of sentences by more talkers. Changing from speech read by professional speakers to spontaneous speech from a variety of speakers will no doubt require modification of our system along several dimensions. The next steps in this research will include:

- Investigating further the relationship between prosody and syntax, including the different roles of phrase breaks and prominences in marking syntactic structure,

- Improving the prosodic labeling algorithm by incorporating intonation and syntactic/semantic information,
- Incorporating the automatically labeled information in the parser of the SRI Spoken Language System (Moore, Pereira and Murveit 1989),
- Modeling the break indices statistically as a function of syntactic structure,
- Speeding up the parser when using the prosodic information; the expectation is that pruning out syntactic hypotheses that are incompatible with the prosodic pattern observed can both improve accuracy and speed up the parser overall.

7 Acknowledgements

This work was supported in part by National Science Foundation under NSF grant number IRI-8905249. The authors are indebted to the co-Principle Investigators on this project, Mari Ostendorf (Boston University) and Stefanie Shattuck-Hufnagel (MIT) for their roles in defining the prosodic infrastructure on the speech side of the speech and natural language integration. We thank Hy Murveit (SRI) and Colin Wightman (Boston University) for help in generating the phone alignments and duration normalizations, and Bob Moore for helpful comments on a draft. We thank Andrea Levitt and Leah Larkey for their help, many years ago, in developing fully voiced structurally ambiguous sentences without knowing what uses we would put them to.

This work was also supported by the Defense Advanced Research Projects Agency 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 Tech Note, Cambridge, England.

[2] E. J. Brisco and B. K. Boguraev (1984) "Control Structures and Theories of Interaction in Speech Understanding Systems," COLING 1984, pp. 259-266, Association for Computational Linguistics, Morristown, New Jersey.

[3] W. Cooper and J. Paccia-Cooper (1980) *Syntax and Speech*, Harvard University Press, Cambridge, Massachusetts.

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

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

[6] A. Komatsu, E. Oohira and A. Ichikawa (1989) "Prosodical Sentence Structure Inference for Natural Conversational Speech Understanding," ICOT Technical Memorandum: TM-0733.

[7] R. Moore, F. Pereira and H. Murveit (1989) "Integrating Speech and Natural-Language Processing," in *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 243-247, February 1989.

[8] P. J. Price, M. Ostendorf and C. W. Wightman (1989) "Prosody and Parsing," *Proceedings of the DARPA Workshop on Speech and Natural Language*, Cape Cod, October, 1989.

[9] M. Steedman (1989) "Intonation and Syntax in Spoken Language Systems," Proceedings of the DARPA Workshop on Speech and Natural Language, Cape Cod, October 1989.

[10] 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, May 1989.

8 Appendix

1a. I 1 read 0 a 0 review 2 of 1 nasality 4 in 0 German.

1b. I 0 read 2 a 1 review 1 of 0 nasality 1 in 0 German.

2a. Why 0 are 0 you 2 grinding 0 in 3 the 0 mud.

2b. Why 1 are 0 you 2 grinding 3 in 0 the 1 mud.

3a. Raoul 2 murdered 1 the 0 man 4 with 0 a 1 gun.

3b. Raoul 1 murdered 3 the 0 man 1 with 0 a 0 gun.

4a. The 0 men 1 won 3 over 0 their 0 enemies.

4b. The 0 men 2 won 0 over 1 their 0 enemies.

5a. Marge 1 would 0 never 0 deal 3 in 0 any 0 guise.

5b. Marge 0 would 1 never 2 deal 0 in 2 any 0 guys.

6a. Andrea 1 moved 1 the 0 bottle 3 under 0 the 0 bridge.

6b. Andrea 1 moved 3 the 0 bottle 1 under 0 the 0 bridge.

7a. They 0 may 0 wear 4 down 0 the 0 road.

7b. They 0 may 1 wear 0 down 2 the 0 road.