PROSODY AND PARSING

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

†SRI International ‡Boston University

ABSTRACT

We address the role of prosody as a potential information source for the assignment of syntactic structure. We consider the perceptual role of prosody in marking syntactic breaks of various kinds for human listeners, the automatic extraction of prosodic information, and its correlation with perceptual data.

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). We concentrate in this report on the use of prosody in parsing, through perceptual studies and through modeling the correlation of duration patterns with stress patterns and with syntactic structures.

It is rare that prosody alone disambiguates otherwise phonetically identical phrases. However, it is also rare that any one source of information (spectral or temporal) is the *sole* feature that separates one phrase from all competitors. We argue here that prosody can provide information useful to a parser. Taking advantage of this 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. There is a vast literature on the perception and production of prosodic information. Our goal is to show that at least some of this information can be automatically extracted and used to improve speech recognition and understanding.

Figure represents a strategy for using prosody to improve speech understanding. Prosodic features extracted from speech will be analyzed by the prosody module which will communicate both with the speech recognition component and with the language processing component. For example, phone durations from the speech recognition module can be used by the prosodic module to hypothesize stress patterns and prosodic structure, which in turn can be checked in the speech recognition component for consistency with the dictionary's lexical stress patterns and with the application of phonological rules in the hypothesized word strings. Similarly, the consistency of syntactic, semantic and discourse structures from the language processing component can be verified against the prosodic structures hypothesized.

The complete integration of the prosodic module with speech and language components is an ambitious goal. Our strategy for attaining this goal is to 1) assess the potential of various types of prosodic information through perceptual experiments and through analysis of recognition and parsing errors, 2) for those aspects of prosody that appear to have the greatest potential, develop computational models using linguistic theory to help determine the units and structure of the models, and using statistical models to help account for variability and to allow for automatic training, and 3) evaluate the models and algorithms by comparing the model's output with hand labels, by comparing human discrimination with the system's performance, and by assessing the accuracy and speed of an SLS with and without the prosodic module. We report here on our initial experiments involving a) the perception of various types of syntactic structures marked by prosody, b) the coding of prosodic information for parsing, and c) the automatic labeling of prosodic structures.

PERCEPTUAL EXPERIMENTS

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

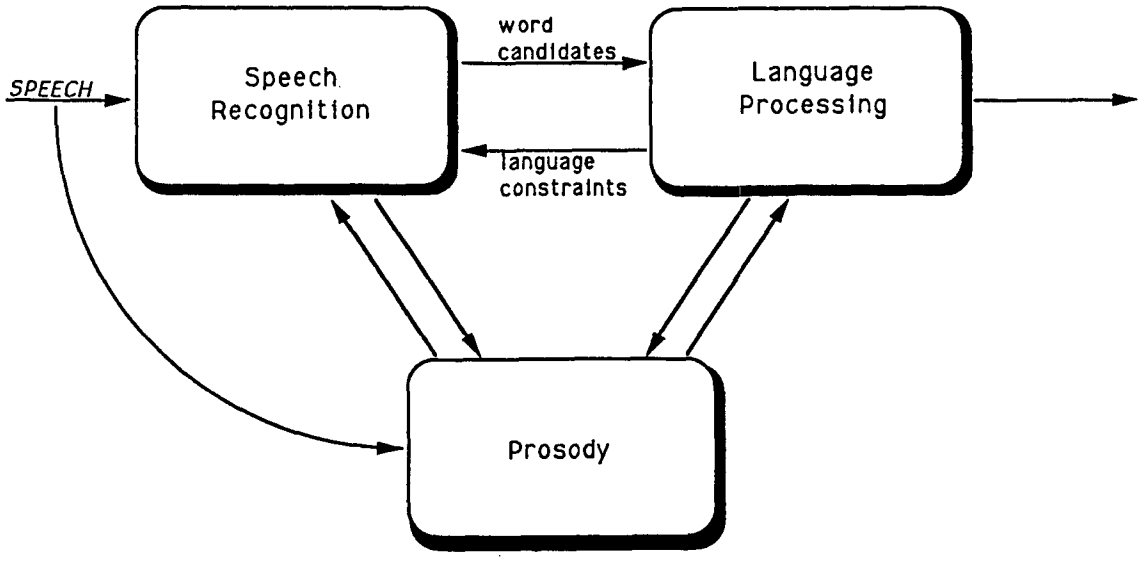

Figure 1: Prosody in Speech Understanding

1983 , Cooper and Paccia-Cooper 1980). Further, there are likely some types of syntactic structures that are not typically marked prosodically. In order to help focus our study on syntactic structures that correlate well with prosody, we designed a perceptual experiment involving phonetically-ambiguous, structurally-different pairs of sentences.

The sentence pairs were read by a professional radio announcer in disambiguating contexts. In order to discourage unnatural exaggerations of any differences between sentences, the materials were recorded in different sessions with several days in between. In each session only one sentence of each pair occurred. Sixteen listeners heard the sentences without the disambiguating contexts and were asked to select which context on the answer sheet was more likely for the sentence heard. Listeners were also asked to put a check next to contexts for which they were particularly confident of their judgments. The listeners were all native speakers of English, naive with respect to the purpose of the study. Again, two different listening sessions with several weeks in between were used, and only one member of each sentence pair occurred in the same session.

Seven types of structural ambiguity were investigated: 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. We list a sample from each class with their disambiguating contexts in the Appendix. In each case, the "a" member of the sentence-pairs has at least one location with a more major break than the corresponding location in the "b" version.

The results in Table 1 show that, on the average, subjects did well above chance (86% correct) in assigning the sentences to their appropriate contexts, although subjects were confident of their judgments only about 58% of the time. In general, the sentences with the more major breaks ("a" versions) were more reliably identified (90% accurate) compared to the "b" versions (83% accurate), though this difference did not correspond to a difference in the subjects' confidence. This suggests that breaks that are strongly marked prosodically may be relatively unambiguous, whereas locations where no break information is observed should not rule out the possibility that a prosodically unmarked syntactic break may exist.

The subjects were least able to identify the main-main vs. main-subordinate pairs (77% accurate, 41%

Ambiguity	a		b		total	
	% correct	% confident	% correct	% confident	% correct	% confident
1. Parenthetical	71	41	93	73	83	59
2. Apposition	98	66	91	72	94	75
3. M-M vs. M-S	92	50	63	33	77	41
4. Tags	98	77	78	51	88	65
5. Near/far attach.	82	46	75	30	79	38
6. Left/right attach.	95	70	93	77	94	74
7. Particle/Prep.	91	55	84	68	87	61
Average	90	58	83	58	86	58

Table 1: Perceptual experiment results for ambiguous sentence interpretation.

confident) and the near vs. far attachment pairs (79% accurate, 38% confident), and were best at the apposition pairs (94% accurate, 75% confident) and the left vs. right attachment pairs (94% accurate, 74% confident). The table also indicates differences related to the "a" and "b" versions by structural category. Because of significant differences across sentences even within a structural category, more perceptual judgments and productions of the same sentences by different speakers are required to assess these differences.

Of particular interest for database query tasks are categories 3, 5, 6 and 7 (parenthetical expressions, apposition and tags may be rare in database queries). These categories cover the range of observed accuracies from worst to best. We plan to concentrate in the future on prepositional phrases. Prepositions are very frequent (about 75% of the Resource Management sentences (Price *et al.* 1988) have one or more prepositional phrases), and nearly all sentences with prepositional phrases are syntactically ambiguous. The perceptual results indicate that it may be possible to help resolve syntactic ambiguity on the basis of prosodic information. One can expect structural ambiguity to pose even more of a problem for a parser when the input is speech rather than text, because of the additional ambiguities of word identity and word boundaries.

CODING PROSODIC INFORMATION FOR PARSING

A formalism has been devised for coding, or labeling, prosodic information in a form useful to a parser. Phrase boundary information is indicated by a "break" index, a number which indicates by its relative size the degree of prosodic decoupling of neighboring words. The example below illustrates how phrase boundary information can be used to distinguish particles and prepositions.

7a Marge 0 would 1 never 2 deal 0 in 2 any 0 guys 5
7b Marge 1 would 0 never 0 deal 3 in 0 any 0 guise 5

In addition, words with relatively high prominence can be marked with * (prominence) and ** (high prominence) following that word. Another example illustrates that prominence information may provide cues which disambiguate the sentences, even when phrase boundary information does not disambiguate the sentence.

6a They 0 rose 2 early** 1 in 0 May 5
6b They 0 rose 2 early 1 in 0 May** 5

We found that we could reliably label these sentences with prominences and with break indices, with good agreement within and across labelers. In ten pairs of sentences examined from categories 5, 6, and 7, the syntactic structures were all clearly disambiguated by the perceived (hand-labeled) phrase break indices in

all but one case. The prominences usually provided additional support for the syntactic structure, and in one case ("they rose early in May") provided the sole support for disambiguation. Because these initial results were encouraging, we began to search for acoustic correlates of the perceptual results.

AUTOMATIC DETECTION OF PROSODIC CUES

In order to automatically label phrase breaks, several sources of information will no doubt be useful, including relative duration, pausing phenomena, boundary tones and syntactic structure. In our initial work, we have investigated independent modeling and detection of these cues. Later, the different algorithms will be combined in a statistical framework, for integration with the SRI spoken language system. Algorithms and experimental results for speaker-dependent detection of breaks, lexical stress and boundary tones are described here. The results for the different algorithms are based on three databases – the ambiguous sentence database, a radio news story database, and the Resource Management database – according to the nature of the information being detected. Other work in prosodic phrase "language modeling", based on a deterministic parser coupled with a Markov model, is described elsewhere (Veilleux *et al.*).

Break Labeling

Our main efforts have involved automatically generating break indices using phoneme duration, a very powerful cue. Phoneme durations were obtained from the SRI speaker-independent word recognition system (Weintraub *et al.* 1989) by constraining the the recognizer so that the correct word sequence is recognized. The SRI recognition system is especially useful for this task, since the phonological rules allow for quite bushy word pronunciation networks. This means that the alignments have a better chance of being highly accurate, and that the phonetic labels thus obtained can be used to independently assess the phonetic ambiguity of the sentences investigated.

Word break indices were generated by normalizing phoneme 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.

As an initial step in assessing the use of prosody in parsing, we have examined the differences between ten of the phonetically ambiguous, minimal pair sentences described in the perceptual studies. The ten pairs included the particle-preposition and attachment ambiguities. These were chosen because these types of ambiguities seem to be relatively frequent in database queries.

Word break indices were automatically generated using the algorithms described above, and the results were compared to perceptually labeled data. In 19 of the 20 sentences, the largest automatically labeled break within a sentence occured at the location of the largest perceived break, which disambiguated the sentence. In the other sentence, there was a large break (3 or 4) that correctly disambiguated the sentence, though not the largest break. Hence, the duration model disambiguates sentences with 100% accuracy for this small set. The correlation coefficient between the hand-labeled break indices and the automatically generated break indices is 0.85.

Lexical Stress Assignment

Duration was also investigated as a cue for predicting lexical stress. Assuming known word segmentations, the lexical stress pattern for multi-syllabic words was estimated based on vowel durations normalized by a window of three successive, neighboring vowels. On a 25-sentence test set from the Resource Management database, this algorithm yielded 90% correct lexical stress pattern prediction.

Breath Detection

For speech that involves more than one sentence or long sentences, speakers typically take breaths. In addition, breaths are highly correlated with major prosodic breaks. We have studied the use of breaths in several minutes of radio news speech. 85% of sentence boundaries (break index 5) are marked by breaths and 53% of major phrase boundaries (break index 4) are also marked by breaths. In addition, the acoustic realization of a breath is quite distinctive, indicating that a breath detection might provide a very reliable phrase break cue.

A breath detection algorithm was developed based on a Gaussian classifier. A fourteen-dimensional full covariance Gaussian distribution was estimated for cepstral features. To compensate for session dependent effects, the cepstral features were normalized by session mean and variance. The classifier labeled successive frames of speech according to a threshold chosen to constrain false detection rate. A segment of speech was labeled as a breath if 23 out of 25 sequential frames were labeled as "breath".

The Gaussian distribution was estimated using data from three different 2-3 minute news stories (three sessions). Separate data from the same stories was used to evaluate the algorithm. The algorithm detected 78 out of 83 breaths (93% correct detection) and inserted one breath. The breaths labeled by the algorithm are within 50 ms of those labeled by hand 95% of the time. In addition, the algorithm detected one breath that was not detected perceptually by four listeners.

Boundary Tone Classification

Boundary tones are another important cue for phrase boundary detection and for semantic interpretation. The boundary tone study is based on a speaker-dependent radio news story database.

The three news stories described above were perceptually labeled for two types of boundary tones at every major phrase break: final fall and continuation rise. A classifier for boundary tones in known final phrase syllable locations was designed using hidden Markov models of intonation. The classifier was based on earlier work with hidden Markov models of isolated intonation patterns (Butzberger, Ostendorf and Price). The classifier used as input only quantized F0 estimates. Using a resubstitution error estimate, 76% of the boundary tones were correctly classified. These results are encouraging because many of the boundary tones were impossible to visually classify because of pitch tracking errors.

CONCLUSIONS

We have shown through perceptual experiments that several types of syntactic ambiguities can be resolved with prosodic information, we have developed a prosodic information coding system suitable for a parser, and we have developed automatic algorithms for extraction of information that correlates well with perceived prosodic phenomena.

These initial results are very encouraging; we plan to test the algorithm on a larger set of sentences by more talkers. Changing from speech read by professional speakers to spontaneous speech from random speakers will no doubt require modification of the algorithms. The next steps in this research will include 1) further investigation of the relationship between prosody and syntax, including the different roles of phrase breaks and prominences in marking syntactic structure; 2) improvement of the prosodic labeling algorithm by incorporating intonation and syntactic/semantic information; and 3) incorporating the automatically

labeled information in the parser of the SRI Spoken Language System (Moore, Pereira and Murveit 1989). We expect the prosodic information to resolve structural ambiguities, and also to increase the speed of the parser by eliminating prosodically inconsistent hypotheses. The tighter the integration scheme between the acoustic information and the syntactic information, the more potential gain we can expect from prosody.

Acknowledgements

This work was supported jointly by NSF and DARPA under NSF grant number IRI-8905249. There were several other researchers involved in this effort. The authors wish to thank Stefanie Shattuck-Hufnagel (MIT) for help with data collection and numerous conversations on prosody, John Butzberger (BU) for help with data collection and boundary tone recognition, Cynthia Fong (BU) for help with the perceptual experiments, Hy Murveit (SRI) for help in generating the phoneme alignments, and John Bear (SRI) for work and insights on integration of prosodic information with a parser.

References

[1] J. Butzberger, M. Ostendorf, P. Price, "Isolated Intonation Recognition Using Hidden Markov Models," submitted to the 1990 Inter. Conf. on Acoustics, Speech and Signal Processing.

[2] Cooper, W. and J. Paccia-Cooper, *Syntax and Speech*, Harvard University Press, Cambridge, MA, 1980.

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

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

[5] P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 651–654, New York, NY, April 1988.

[6] N. Veilleux, M. Ostendorf, P. Price, S. Shattuck-Hufnagel, "Markov Modeling of Prosodic Phrase Structure," submitted to the 1990 Inter. Conf. on Acoustics, Speech and Signal Processing.

[7] M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin and D. Bell, "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.

EXAMPLES OF AMBIGUOUS SENTENCE PAIRS

1. Parentheticals:
   (a) Mary leaves on Tuesday. She will have no problem Europe. *Mary knows many languages, you know.*
   (b) Mary and you have similar backgrounds and have both learned many languages. *Mary knows many languages you know.*
2. Apposition:
   (a) The Smiths didn't know what to do with their time while their television was broken. *The neighbors who usually read, the Daleys, were amused.*
   (b) There was a funny Doonesbury today in all the local papers. *The neighbors who usually read the dailies were amused.*

3. Main-main vs. main-subordinate clauses:

   (a) His mother and father did not have the same reaction when he announced he was going to become a hairdresser. ***Mary was amazed and Dewey was angry.***

   (b) Mary couldn't believe anyone would object to such a harmless prank. ***Mary was amazed Ann Dewey was angry.***

4. Tags:

   (a) Dave is always very angry, but it's futile to ask him why. ***Dave will never know why he's enraged, will he?***

   (b) Dave can be obnoxious without realising it. He just insulted Willy and is puzzled by his anger. ***Dave will never know why he's enraged Willy.***

5. Attachment of final phrase:

   (a) You'll never believe what she had on when she eloped. ***Laura ran away with the man wearing a green robe.***

   (b) Which man did Laura run away with? ***Laura ran away with the man wearing a green robe.***

6. Attachment of middle phrase:

   (a) In spring there was always more work to do on the farm. May was the hardest month. ***They rose early in May.***

   (b) Bears sleep all winter long, usually coming out of hibernation in late April, but this year they were a little slow. ***They rose early in May.***

7. Particles and prepositions:

   (a) Marge loves cards but she refuses to deal. We would often try to trick her into doing it, but it never worked. ***Marge would never deal in any guise.***

   (b) Marge is a real card shark and adores dealing poker, but she will only play with women. We would sometimes try to get her to let one of our male friends int the game, but she always refused. ***Marge would never deal in any guys.***