SESSION 13: PROSODY

Patti Price[1] and Julia Hirschberg[2]

[1]SRI International, 333 Ravenswood Avenue, EJ 133, Menlo Park, CA 94306.
[2]AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974

ABSTRACT

The aim of this introductory section is to set the context for Session 13: Prosody. It will do so by defining some basic terms, by considering the status of current research on prosody, and by outlining the papers in the session and how they contribute to and complement previous work in the area.

1. What is Prosody?

Prosody, perceptually, can be thought of as the relative temporal groupings of words and the relative prominence of certain syllables within these groupings. Acoustic correlates of prosody include patterns of relative duration of segments and of silences, fundamental frequency, amplitude, and "vowel color." Phonological variation is related in part to prosodic structure and is sometimes also considered part of prosody. The bulk of the research on prosody, in the literature as well as in this session, is focused on the primary acoustic correlates of prosody, namely patterns of fundamental frequency and duration.

It is appropriate to end this workshop on this topic, since prosody, perhaps more than any other area in spoken language systems, requires the involvement of both speech and natural language. Prosody can provide for natural language a source of acoustic information bearing on higher linguistic levels. Further, this information is largely unrepresented in text. One of the questions addressed in this session is: How can the acoustic attributes of prosody be transmitted not just to a speech recognizer (which is used to interpreting acoustic information), but also to natural language understanding components? The temporal (grouping) aspect of prosody appears to be related to the syntactic structure of an utterance, and one could imagine a component that would pass temporal information to a parser. The prominence aspect of prosody appears to be related to the semantic and discourse/pragmatic structure, and one could imagine a component that would pass prominence information to these levels. However, both grouping and prominence relationships are involved to some extent in all linguistic levels, and a more complex architecture is required.

Prosody is an area ripe for further research since it requires the integration of information from all levels, from the acoustics through morphology, syntax, semantics and pragmatics. Few, if any, researchers are comfortable in all these areas, and the work requires collaboration across traditional divisions among disciplines, as these papers show.

2. HOW WELL DOES THIS SESSION REPRESENT THE RESEARCH ON PROSODY?

In this session, in contrast to historical trends in prosody research, there is a focus on using statistical and corpus-based techniques. Further, in this session, these techniques are used not only to model prosody, but also to acquire information about prosody and its role in language. The traditional literature on prosody, while lacking this statistical aspect, does include much that is not covered in this session. Much previous prosody research has been from two rather different traditions, neither of which overlaps greatly with the content of this session. The two traditions are (1) speech science, which has focused on the search for acoustic correlates of linguistic entities (such as stress and accent) in laboratory conditions, and (2) linguistics, which has produced a volume of intuitive, anecdotal attributions of prosody's role in higher linguistic levels, such as pragmatics and discourse.

Growing interest in prosody research, due in part to the demands of the recent merging of speech and natural language, has been somewhat limited by the lack of an agreed upon convention for prosodic notation. However, as mentioned in the paper in this session by Silverman, this issue is currently being addressed, with quite encouraging results. A draft notation was developed last fall by a group of prosody researchers from a number of academic and industrial research laboratories. The draft notation is currently being evaluated, and a second meeting of the group will occur in early April to refine the system and to plan for future needs.

In brief, the prosody notation project consists of researchers meeting periodically and exchanging volumes of email. The goal of the group is to define a "core" notational system with which spoken English can be transcribed quickly and with good agreement across labelers. Agreeing on a standard with these characteristics means that large corpora could be labeled with prosodic and other linguistic notations (e.g., those of the Penn Treebank). It also means that what would otherwise have been independent corpora can now be shared, and that researchers can much more easily test hypotheses on new and larger corpora. The existence of large corpora will also encourage the further development of statistical methods for modeling and for acquiring knowledge. The notion of a "core" system is meant to imply that different researchers can add different details to the "core," and thus reap the benefits of the shared portion while not limiting research to those aspects that all can agree on now.

It is hoped that the existence of large labeled corpora can help bridge the traditional conflict between using data in which known sources of variability are strictly controlled (e.g., readings of isolated utterances in a laboratory environment) versus using naturally occurring data (which may form a sample too small and too variable to be used for anything other than impressionistic analysis). Such corpora should also encourage work in evaluation, since experimental hypotheses reported by one researcher can be tested by another researcher on the same corpus or on a new corpus labeled according to the same conventions.

In sum, this session does not represent the field of prosody research as a whole. Rather, it represents that part of prosody research which hardly existed previously, but which is currently growing rapidly. In particular, this session includes statistical methods, corpora-based approaches, analyses of field data and spontaneous speech phenomena. These are areas that have previously played a small or non-existent role, and in which there are likely to be some important new results.

3. OVERVIEW OF THE PAPERS IN THIS SESSION

The first paper in this session, by Shriberg et al., focuses on the prosody of speech and language repairs in spontaneous speech using prosody as one knowledge source. The paper introduces a notational system for categorizing and analyzing repairs, discusses the distribution of these repairs and proposes an initial algorithm for "repairing" the repairs. The algorithm is based on pattern matching in combination with other knowledge sources including syntax, semantics and acoustics.

Abney's paper concerns the role of prosody and syntax. This paper points out, as others have noted, that phrase structure and prosodic structure are not identical, but neither are they entirely unrelated. The paper describes a particular approach to syntactic analysis that appears to provide interesting correlations between syntactic and prosodic structure. The study evaluates the hypothesized structures on two data sets, including one from the ATIS domain.

While it is true that we need an increased understanding of the relationship between prosody and syntax (both for synthesis and for understanding), we also need to address the fact that given a syntactic structure, there is some choice about how to assign a prosodic structure. The paper by Ostendorf and Veilleux describes a strategy that models the choice, or variability, in prosodic structure. The proposed strategy, an extension of work done by Wang and Hirschberg, is used in synthesis as well as analysis (using an analysis-by-synthesis approach). The new contribution is a richer set of segmentation (grouping) levels to be modeled.

The paper by Silverman and others from Nynex focuses on locating discourse-relevant information-bearing words within an utterance. This concept is similar, but not identical, to what may be called discourse "focus." The study is a field analysis using actual directory assistance calls, and compares the use of prosody in read and in spontaneous speech, using the draft prosody notation mentioned earlier. This study, like others at the workshop, point out potential differences in these two speech styles, differences that need to be better understood if we are to adequately model human speech.

At the workshop, Mark Liberman reported a related study of calls left on answering machines compared to those made to people. The calls left on the machine were similar prosodically to the read style in the Nynex study. This appears to be evidence that the change in prosody may be related more to the interactivity of the situation than to the read vs. spontaneous contrast. Although most read speech is non-interactive, spontaneous speech can be either interactive (as in the people talking to the Nynex simulated understanding machine) or non-interactive (as in talking to an answering machine).

The last paper in the session, by Hirschberg and Grosz, takes the position that similar models of prosody should be used for synthesis and understanding. The study uses Grosz and Sidner's discourse model for segmentation and attempts to correlate the resulting structures with various prosodic attributes. Seven labelers are used, some of whom segmented based on the text only, and others who segmented on the basis of both text and through listening. The inter-labeler reliability and the correlations between acoustic and discourse attributes of prosody are results important for those who want to develop spoken language systems as well as those who want to better understand human communication processes.

The use of decision trees, evident in this session as well as in the statistical language modeling session of this workshop, is a growing research area. Decision trees have the interesting property of providing a straightforward mecha-

nism for combining the virtues of probabilities with knowledge from rule-based approaches. Because the structure of the knowledge from the rules is preserved, these methods can be used not only to model the phenomena of interest, but also to test hypotheses and to gain knowledge about which information sources are providing the most gain in the models.

During the discussion period for this session, Bill Fisher pointed out that one of the utterances in the test set ("Does this flight leave on Friday or Saturday") had to be marked class X (not evaluable) because the lexical SNOR was ambiguous. The sentence could, however, be disambiguated on the basis of its prosody (either by listening or by observing the .sro transcription). Throwing out sentences that prosody could disambiguate discourages work on integrating prosody in our systems. Bob Moore pointed out a related issue: the impoverished representation of disfluencies in the lexical SNOR representation. This is important since disfluencies form a significant portion of the errorful sentences in our systems. Our representations and our evaluations are not yet in phase with our research goals.

We could address these issues by: (1) having the annotators listen to sentences before classification (which is costly if done for every sentence), (2) developing a new transcription level that includes more than the lexical SNOR but less than the .sro transcriptions, or (3) providing two .cat files for some utterances (one categorization on the basis of lexical SNOR and the other on the basis of the .sro file). These issues should be addressed by the MADCOW committee in cooperation with the Principles of Interpretation committee. As pointed out by Hirschberg and Grosz, it is of interest to know what different interpretations may be derived on the basis of text alone versus text and speech. We propose that a sample of the sentences be annotated from speech, so that at least we know what is being missed by relying on the text alone.

4. HOW DO THESE PAPERS FIT INTO THE REST OF PROSODY RESEARCH?

These papers differ from traditional prosody research in their focus on statistical and corpus-based approaches, but they also differ in other ways. Table 1 outlines the areas covered by most previous work in prosody. The traditional linguistic divisions are labeled down the left-most column, and the other column headings address the traditional division between perception (analysis, understanding) on the one hand and production (synthesis, generation) on the other. The perception area consists of two columns, for read and spontaneous styles. The production column is not so subdivided because we have not yet addressed the issue of generating spontaneous speech, though a better understanding of these mechanisms could eventually lead to more natural synthesis. The read speech column is located closer to the production column since this style is closer to what is currently used in synthesis.

Table 1: How do these papers fit into the rest of prosody research?

Source of information	Production	Perception (read speech)	Perception (spontaneous speech)
pragmatic	5	4	4
discourse	5		
semantic			1
syntactic	2 3		1
morphological			
lexical			
phonetic			
acoustic	2 5 3	4	1 4

The group of researchers dealing with speech and language is not culturally homogeneous. Even among those with degrees in the same subject area (e.g., linguistics), specialists from different sub-disciplines manifest significant cultural differences in background, strategies, beliefs and goals. The shaded areas of Table 1 indicate where traditional research on prosody has focused its efforts. Please forgive any biases in this interpretation which is meant to be suggestive only.

There is much work in the pragmatics literature, which draws greatly from semantics and which touches on syntax, but seldom has involved acoustic analyses. On the other hand, the tradition of speech science (which includes linguists of the more phonetic ilk) has focused a great deal of attention on the production and perception of read speech styles. This work occasionally touches on syntactic issues, but largely ignores the higher linguistic levels as well as issues related to spontaneous speech. Largely, the past work in prosody has been performed by two different communities with goals and results somewhat offset from one another.

The papers in this session start to fill in some of the gaps in the field of prosody. For example, the Shriberg et al. paper (labeled '1' in the table) begins to integrate aspects of acoustics, syntax and semantics for spontaneous speech. Abney's paper (labeled '2' in the table) considers syntax and some aspects of acoustics as related principally to production. The Ostendorf and Veilleux paper (labeled '3' in the table) integrates aspects of acoustics and syntax in a representation neutral with respect to perception and production. The paper by Silverman et al. (labeled '4' in the table) integrates aspects of acoustics and pragmatics in read and spontaneous speech. The Hirschberg and Grosz paper, (labeled '5' in the table) strives for a representation neutral with respect to production and perception and integrates aspects of pragmatics and discourse with acoustics.

As illustrated in Table 1, the papers in this session represent rather new research areas and begin to fill out the field of prosody. We are very hopeful that this area will provide results that will improve our understanding of human communication, and that will be useful in the development of spoken language systems. However, though these papers begin to fill some gaps in our understanding of prosody and its relationship to other areas of speech and language, it is clear that far more research is needed. To fully understand the nature of prosody, and to be able to use it effectively, we still have a good deal more integration work to achieve.