SESSION 11: PROSODY

M. Ostendorf

Electrical, Computer and Systems Engineering
Boston University, Boston, MA 02215

ABSTRACT

This paper provides a brief introduction to prosody research in the context of human-computer communication and an overview of the contributions of the papers in the session.

1. WHAT IS PROSODY?

In large part, prosody is "the relative temporal groupings of words and the relative prominence of certain syllables within these groupings" (Price and Hirschberg [1]). This organization of the words, as Silverman points out [2], "annotates the information structure and discourse role of the text, and indicates to the listener how the speaker believes the content relates to the ...prior knowledge within the discourse context." For example, the relative groupings of words can provide cues to syntactic structure as well as discourse segmentation, and the relative prominence of words can provide cues to semantically important or focused items. Segmentation and focus represent two of the major uses of prosody, but other information may also be cued by intonation patterns, e.g. indication of continuation, finality or a yes-no question with phrase final "boundary tones".

Prosody is typically also defined with a reference to its suprasegmental nature: "Prosody comprises all the sound attributes of a spoken utterance that are not a property of the individual phones" (Collier) [2]. In addition, prosody can operate at multiple levels (e.g., word, phrase, sentence, paragraph), making computational modeling of prosody particularly challenging. The acoustic correlates of prosody, which include duration of segments and pauses, fundamental frequency (F0), amplitude and vowel quality, may be influenced by prosodic patterns at more than one level, as well as inherent segmental properties. Modeling the interactions among the different factors is an important and difficult problem.

Most current linguistic theories of prosody include an abstract or phonological representation of prosody to characterize aspects of phrasing, prominence, and intonation or melody. However, here we also see that abstract representations are of interest for computational modeling. Since it is generally agreed that prosody is not directly related to standard representations of syntactic structure, it is useful to have an intermediate representation to facilitate automatic learning and to simplify model structure. Thus, the form of an abstract representation is an important issue. Ideally, it should include all three main aspects of prosody, and address the needs of theory and computational models. Many different schemes have been proposed, and variations of two different prosodic transcription systems are used in the papers presented in this session. The TOBI (Tones and Break Indices) system for American English [3] is a prosodic transcription system that has evolved from a series of workshops where researchers met with the goal of defining a common core of transcription labels. The TOBI system is used to varying degrees in the papers by Silverman, Veilleux and Ostendorf, and Nakatani and Hirschberg. The IPO taxonomy of intonation for Dutch [4], which is used in the work of Collier, de Pijper and Sanderman, was developed from a long tradition of research in intonation that has recently been applied to several languages.

2. PROSODY & HUMAN-COMPUTER COMMUNICATION

The theme of this workshop is on technology for automated language processing, and thus the emphasis in this overview is on representations and computational models of prosody for spoken language processing applications. There are two classes of problems in speech processing for human-computer interactions: speech synthesis and speech understanding. Prosody plays a role in both problems, as is clearly seen in the different papers covered in this session. Prosodic patterns are determined by the information structure of language and realized in the speech waveform in terms of F0, duration and energy patterns. As illustrated in Figure 1, the overall problem in computational modeling of prosody is to move from one domain to the other, optionally via an intermediate abstract representation.

Until recently, almost all research in computational modeling of prosody has been in speech synthesis applications, where it has been claimed that good prosody models are among the most important advances needed for

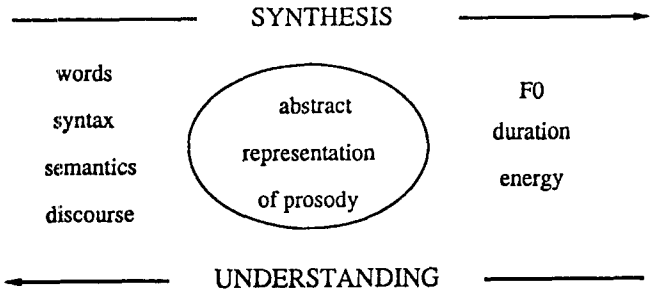

Figure 1: Problems in computational modeling of prosody for human-machine communication.

high quality synthesis. The papers by Silverman, van Santen, and Collier *et al.* each address different problems related to prosody synthesis. Silverman attacks the problem of predicting abstract prosodic labels, while van Santen presents a model for predicting duration from text (and optionally abstract labels). Collier *et al.*, on the other hand, analyzes the relation between automatically predicted boundary levels and perceived level in natural speech. Both Silverman and van Santen make the point that good prosody models can improve naturalness, but Silverman also shows that *intelligibility* can be improved.

Speech understanding is a relatively recent area of research for prosody, although researchers have long cited anecdotal evidence for its usefulness. Within the speech understanding domain, the papers in this session are directed mainly at contributions of prosody to natural language processing. An example is the use of prosody in combination with other "knowledge sources" to choose among the different possible interpretations of an utterance, investigated by Veilleux and Ostendorf. Some utterances from the ATIS domain that illustrate the potential role of prosody in interpretation include:

Does flight US six oh four leave San Francisco on Friday or Thursday?

where both intonation and phrase structure can be used to distinguish between the yes-no question and the "Thursday vs. Friday" alternative, and

Show me the itineraries for Delta flights eighty two one three one two seven five and one seven nine.

where knowledge of phrasing can help determine the specific flights referred to. Prosody can also serve speech understanding systems in an entirely different way, as discussed in the paper by Nakatani and Hirschberg, which is to cue the presence of a disfluency and the interval of replacement. As an example, consider another sentence from the ATIS domain, where prosody would be useful in automatically distinguishing a disfluency from a speech recognition error:

What is the <light> latest flight on Wednesday going from Atlanta to Washington DC?

Of course, the presence of disfluencies complicates the design of prosodic models, e.g. since fluent and disfluent pauses may cue different types of syntactic constituents.

An important question in current approaches to computational modeling of prosody is the specification of (or even use of) an intermediate phonological representation. Although all papers use some sort of discrete prosody labels, the paper by Collier *et al.* specifically investigates the perceptual relevance of one type of prosodic label – an integer representation of relative phrase breaks – and its acoustic correlates.

3. IMPORTANT THEMES

Several important and common themes, indicative of recent research trends, cut across subsets of these papers. First, it is significant that both synthesis and understanding applications of prosody are represented in this session, and useful since the developments in one field can benefit the other. Second, we see corpus-based analysis and automatic training methods being introduced into many aspects of prosody modeling. Third, Silverman's results argue the case for developing models in constrained domains, but this approach is also supported by the development of automatic training methods and probably used to advantage in the papers focussed on the ATIS domain. Fourth, all of the papers use an intermediate prosodic representation at some level, which raises the issue of representation as an important research question in its own right. Perhaps the most important contribution of this session is the collection of experimental results demonstrating the benefits of prosody in actual synthesis and understanding applications, providing concrete and not just anecdotal evidence that prosody is a useful component of a spoken language system. Since these themes represent relatively new directions in computational modeling of prosody, the applications and modeling possibilities are only beginning to open up and we can expect many more gains in the future.

References

1. P. Price and J. Hirschberg, "Session 13: Prosody," *Proc. of the DARPA Workshop on Speech and Natural Language*, pp. 415–418, 1992.
2. Prosody definitions, personal communication.
3. K. Silverman *et al.*, "TOBI: A Standard Scheme for Labeling Prosody," *Proc. of the Inter. Conf. on Spoken Language Processing*, pp. 867–870, 1992.
4. J. 't Hart, R. Collier and A. Cohen, *A Perceptual Study of Intonation*, Cambridge University Press, 1990.