PERCEIVED PROSODIC BOUNDARIES AND THEIR PHONETIC CORRELATES

René Collier, Jan Roelof de Pijper and Angelien Sanderman
Institute for Perception Research / IPO
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

ABSTRACT

This paper addresses two main questions: (a) Can listeners assign values of perceived boundary strength to the juncture between any two words? (b) If so, what is the relationship between these values and various (combinations of) suprasegmental features. Three speakers read a set of twenty utterances of varying length and complexity. A panel of nineteen listeners assigned boundary strength values to each of the 175 word boundaries in the material. Then the correlation was established between the variable strength of the perceived boundaries and three prosodic variables: melodic discontinuity, declination reset and pause. The results show that speakers may differ in their strategies of prosodic boundary marking and listeners agree in the perceptual weight they attribute to the prosodic cues.

1. INTRODUCTION

Any two successive words may vary as to their syntactic or semantic cohesiveness. The latter is likely to be stronger if the two words are part of the same linguistic constituent; conversely, the occurrence of a constituent boundary between words decreases their degree of cohesiveness. For example, in the sentence "(the man) (is sitting) (in the chair)", any two words separated by round brackets are structurally farther apart than those within a pair of brackets. Speakers are capable of making the juncture between constituents audible by prosodic means: they may produce appropriate cues in terms of pause, pitch and duration parameters. Listeners, on the other hand, can make use of these cues to segment the incoming flow of speech into word sequences that may be treated as a whole, which facilitates the comprehension process. In certain cases, prosodic demarcation may help in resolving structural ambiguity, for instance in utterances of the type "The girl saw the man with the telescope", in which the prepositional phrase specifies either the verb or its direct object [1, 2]. But in utterances containing no surface syntactic homonymy, too, prosodic boundaries may delineate coherent word groups and lend support to the listener's hypotheses about syntactic-semantic structure as, for instance, in "the beautiful girl / with brown eyes / told her story / to the psychiatrist" [3].

This paper presents results of research that, starting from the observation that listeners do provide prosodic boundary cues, addresses two main questions:

(a) Can listeners assign a value of Perceived Boundary Strength (PBS) to word boundaries?

(b) If so, what is the relationship between PBS and different (combinations of) suprasegmental features?

The answer to these questions may lead to a better model of what prosodic resources a speaker can draw on to highlight the syntactic-semantic structure of an utterance. Such insight may, in turn, contribute to improved prosody in speech synthesis, by making it sound more natural and –more importantly– by making it linguistically more transparent and therefore easier to comprehend. This research may also shed light on how a listener makes use of the demarcative information encoded in prosodic features. In that respect, it has relevance for (knowledge-based) automatic speech recognition, where the inclusion of prosodic information may support the syntactic-semantic parse of the input, especially if the latter contains structural ambiguities.

This line of research is in agreement with the growing interest in the communicative function of prosody, which may contain information not only about utterance-internal phrasing, as already suggested above [4], but also about the topical organization of discourse in monologues and dialogues [5] or about speaker-dependent features such as emotional state [6].

2. EXPERIMENTAL APPROACH

In this section we present part of the results obtained in an experiment that aimed at answering the two questions mentioned in the introduction. To this effect we have collected appropriate speech material, in which we asked listeners to score the PBS of each word boundary. Subsequently, the material was subjected to various phonetic analyses, the results of which were then correlated with the PBS's. Finally, the predictions of an algorithm that assigns prosodic structure to unmarked text were verified against the PBS's.

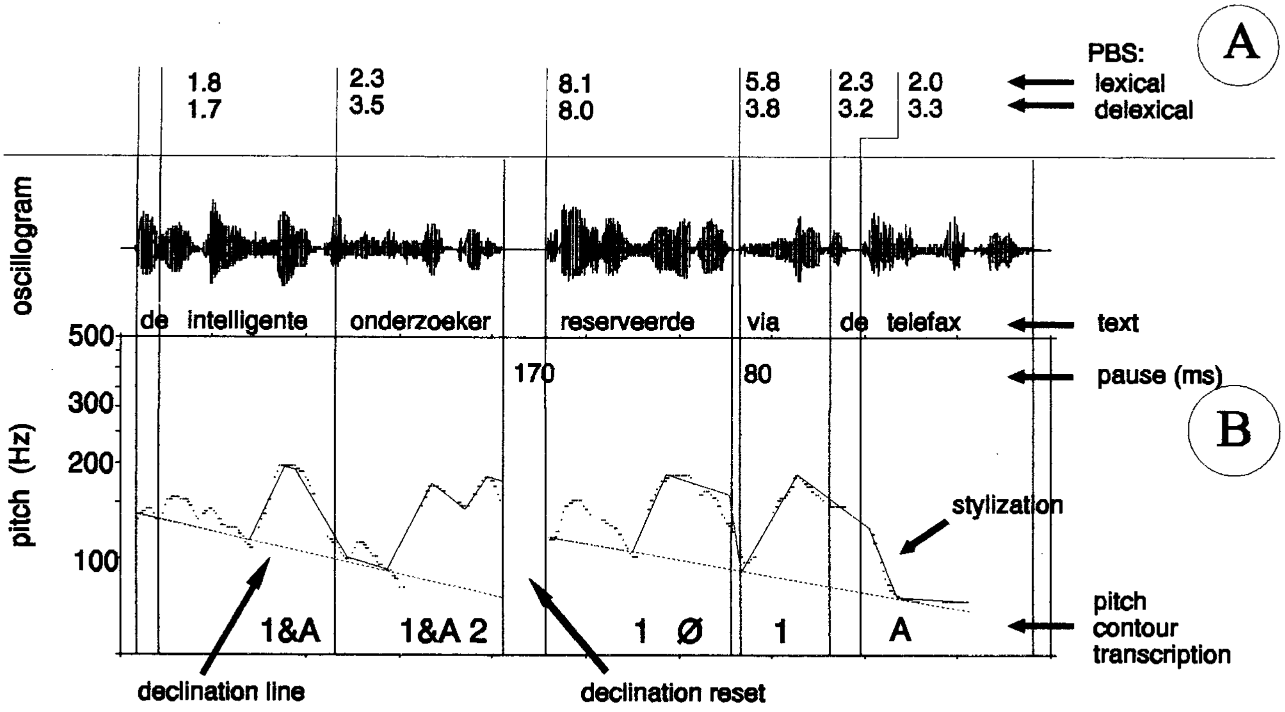

Figure 1: (A) PBS values and (B) results of the phonetic analyses for one of the test utterances of the professional speaker.

2.1. Speech Material

A set of twenty Dutch sentences was constructed, which differed sufficiently in length and complexity to warrant the occurrence of prosodic boundaries of varying strengths. These sentences contained a total of 175 word boundaries. This set was read out by three native speakers: two males, of whom one was a professional speaker, and one female. To evaluate the possible influence of syntactic and semantic information, all 20 utterances spoken by the professional speaker and 3 of the utterances spoken by the other two were processed in such a way that the contents of the utterances was rendered unintelligible, while the prosodic features were kept intact. In this way, a so-called 'delexicalized' version of the test material was created in addition to the 'normal' version.

2.2. PBS Assignment

In a number of successive sessions, nineteen listeners were confronted with the 3 x 20 = 60 utterances in the normal version and the (1 x 20) + (2 x 3) = 26 in the delexicalized version. They were asked to indicate, on a 10-point scale, how strong they felt the juncture at each word boundary to be. The mean of the nineteen scores per word boundary was taken as a measure of the perceptual boundary strength (PBS) of that word boundary. Thus, the PBS was obtained for each word boundary as produced by each of the speakers in each test version.

An example of the PBS values obtained for one of the test utterances is shown in Figure 1a. As can be seen, listeners appear to be quite capable of distinguishing a diversity of PBS values, both in the lexical and delexical conditions.

2.3 Phonetic Analysis

The acoustic / phonetic analysis of the material concentrated on the speakers' use of pauses and intonation to highlight word boundaries. It was determined for each word boundary in the 60 utterances 1) whether there was a pause and, if so, of what length; 2) whether there was melodic discontinuity across the boundary and, if so, of what type; and 3) whether the boundary was associated with a declination reset.

The location and length of pauses were determined by straightforward inspection of the waveforms. Melodic transcriptions of the 60 utterances were obtained by a combination of pitch measurement, pitch stylization and independent perceptual evaluation by experts. Following the typology outlined in 't Hart et al. ([7], p.81), four types of melodic discontinuity were distinguished in the way the speakers marked the word boundaries: '1Ø, 1E, 12, 1A2'.

Figure 1b presents a survey of the results of the phonetic analyses for one of the test utterances.

3. RESULTS

For all three speakers, a high correlation was found between the PBS's obtained in the normal and delexicalized test versions ($r_s = .78$, $p < .01$). This warrants the conclusion that, in this experiment, syntactic and semantic factors did not affect the listeners' judgments. The delexicalized test version is ignored in the rest of this paper.

3.1. Perceptual Boundary Strength And Phonetic Cues

	Prof	Nonprof-1	Nonprof-2
melodic discontinuities	43	37	31
pauses	28	14	11
declination resets	14	2	0

Table 1: Frequency of occurrence of the three phonetic cues across the three speakers.

The three speakers appear to make different use of phonetic cues to mark prosodic boundaries, as shown in Table 1. The table shows that the professional speaker made more extensive use of all three phonetic cues than the other two speakers and was the only one to employ declination resets in a systematic fashion. Not shown in the table is the fact that there were also clear differences between the speakers in preferred type of melodic discontinuity.

Combinations of the three cues can be considered as possible phonetic strategies of the speakers to mark prosodic boundaries. Figure 2 shows the relation of these strategies to PBS. Generally speaking, PBS values are higher as more phonetic cues are associated with a given word boundary. While the speakers differ in their preferences for certain strategies, the impact of strategy on PBS is roughly the same across speakers.

Additional trends not shown in Figure 2 are the following. First, there was a trend for longer pauses to be associated with greater PBS's for all speakers. As for the four types of melodic discontinuity, the main tendency is that melodic discontinuity involving a continuation rise '2' (a steep pitch rise very late in the pre-boundary syllable) is associated with greater PBS's than other types.

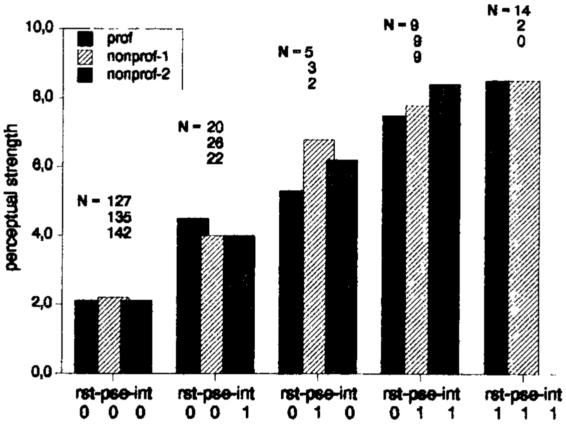

Figure 2: PBS per phonetic cue combination: *rst* = declination reset, *pse* = pause, *int* = melodic cue, *0* and *1* = absence or presence of a cue.

The data show strong interactions between the three phonetic cues. The main observations are that 1) the presence of a declination reset implies the presence of a pause in all cases, 2) the presence of a pause implies the presence of a melodic discontinuity in about 80% of the cases, for all speakers, 3) pauses not accompanied by a melodic cue are usually shorter than 100 ms, and 4) it is quite common for word boundaries to be marked only by a melodic cue.

3.2. Perceptual Boundary Strength And Prosodic Boundaries

The prosodic analysis of the test material consisted of the application of the latest version of the so-called Pros-3 algorithm [8]. This is a program currently under development at IPO to automatically determine accent and prosodic phrase structure of sentences on the basis of syntactic and metrical analysis. In this way, each word boundary was assigned to one of three predicted prosodic boundary categories: no boundary, Phi-boundary or I-boundary.

As can be seen in figure 3, word boundaries that were designated as I-boundaries by the Pros-3 algorithm have greater PBS's than Phi-boundaries, while these in turn are perceived as stronger than unlabelled boundaries. This effect is apparent for all speakers, but is clearest in the professional speaker.

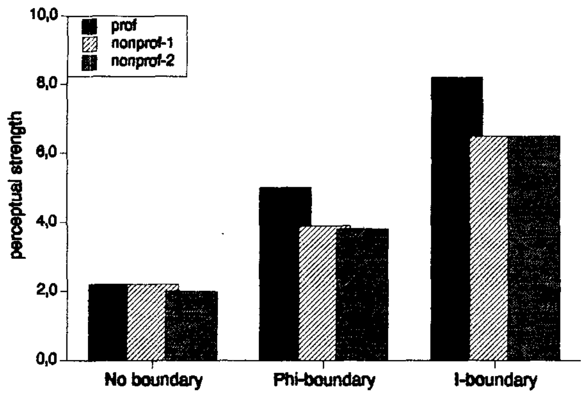

Figure 3: PBS per prosodic boundary.

4. CONCLUSIONS

Our experimental investigation has brought to light that speakers and listeners alike are aware of the role pitch and pause can play in utterance-internal phrasing. These prosodic parameters can effectively highlight how the utterance is to be chunked into coherent word groups. As is often the case with prosody, there is no obligation on the speaker's side to actually use pitch or pause for a communicative purpose, such as boundary marking. In fact, we have seen (in Table 1) that our professional speaker produces more numerous prosodic cues and that all three speakers differ in the relative frequency of use of pitch or pause devices. But, whenever listeners are offered particular (combinations of) prosodic cues, they agree well on how to interpret them in terms of PBS (see Figure 2). Thus, there seems to be prosodic freedom on the speaker's side, while the listener cannot help but pay attention to melodic or temporal cues whenever they are present.

Apparently, a certain amount of gradience is involved in the marking of constituent boundaries: the strength of a prosodic boundary reflects to some extent the depth of the syntactic-semantic juncture at a given point in the utterance. It is not unlikely that the inclusion of additional prosodic parameters, such as local variations in speech rhythm (in particular preboundary lengthening), will add detail to the emerging picture of syntax-to-prosody correspondence. But pitch and pause alone already show sufficiently clear relations to the linguistic structure of the utterance, that their capacity to reveal this structure can be exploited tentatively in text-to-speech conversion and in automatic speech recognition.

Finally, the gradience that can be observed in the PBS values of Figure 2, shows that listeners can do better than merely distinguishing between presence or absence of a boundary. On the basis of our limited set of data, it is not possible to determine exactly how many categories listeners can discriminate reliably. This will be partly determined by the number and the nature of the phonetic cues, and need not be limited to a maximum of three (no boundary, minor boundary, major boundary). Indeed, Figure 2 suggests that it is not unreasonable to assume that listeners can handle five PBS categories. Interestingly, such a five-level distinction is used in the TOBI labelling scheme [9]. However, an important difference between the two approaches is that the TOBI scheme obliges labelers to explicitly assess the nature of the phonetic cues, while PBS values are the result of a purely intuitive judgment.

References

1. Price, P., Ostendorf, M., and Wightman, C. "Prosody and parsing", In: *Proc. of the Second DARPA Workshop on Speech and Natural Language*, 1989, Morgan-Kaufman, San Mateo, CA, pp. 5–11.

2. Price, P., Ostendorf, M., Shattuck-Hufnagel, S. and Fong, C. "The use of prosody in syntactic disambiguation", *J. Acoust. Soc. Amer.*, Vol. 90, 1991, pp. 2956–2970.

3. Terken, J. and Collier, R. "Syntactic influences on prosody", In: *Speech perception, production and linguistic structure*, Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (eds.), Ohmsa Press, Tokyo, 1991, pp. 427–438.

4. Ladd, D.R. "Declination 'reset' and the hierarchical organization of utterances", *J. Acoust. Soc. Amer.*, Vol. 84, 1988, pp. 530–544.

5. Pierrehumbert, J., and Hirschberg, J. "The meaning of intonational contours in the interpretation of discourse", In: *Intentions in communication*, P. Cohen, S. Morgan and M. Pollock (Eds), MIT Press, Cambridge MA, 1990, pp. 271–311.

6. Carlson, R., Granstrom, B., and Nord, L. "Experiments with emotive speech-acted utterances and synthesized replicas", In: *Proceedings Int. Conf. on Spoken Languague Processing, Banff*, 1992, pp. 671-674.

7. 't Hart, J., Collier, R. and Cohen, A. *A perceptual study of intonation*, Cambridge University Press, 1990.

8. Dirksen, A. "Accenting and deaccenting: a declarative approach", In: *Proc. 15th Int. Conf. on Computational Linguis-*

tics, COLING, Association for Computational Linguistics, 1992, pp. 865 -869.

9. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992). "TOBI: A standard for labeling English prosody", In: *Proc. Int. Conf. on Spoken Language Processing,* 1992, pp. 867- 870.