Umeå University, Department of Philosophy and Linguistics
PHONUM 9 (2003), *117-120*
Available online at http://www.ling.umu.se/fonetik2003/

The acoustic and morpho-syntactic context of prosodic boundaries in dialogs

Mattias Heldner and Beáta Megyesi

Department of Speech, Music and Hearing, KTH

This study investigates the structuring of speech in terms of prosodic boundaries. In particular, the relation between boundaries as perceived by listeners, and their acoustic and linguistic realizations as uttered by speakers is examined.

1. Introduction

The structuring of speech in terms of prosodic boundaries is fundamental for spoken communication. By reflecting the speakers' internal organization of the information, prosodic boundaries facilitate the listeners' processing of the message. This study is viewed as a step towards a general model of the structuring of speech with applications in speech technology; e.g. to predict prosodic boundaries from input texts for speech synthesis, to produce natural sounding boundaries in synthetic speech, and to predict boundaries from input speech for automatic speech recognition and understanding. To arrive at such a model, however, several kinds of information have to be taken into account. Perceptual classifications of prosodic boundaries in a speech material and detailed acoustic and linguistic descriptions of these boundaries and their context will be required.

Each type of information has its own problems and limitations. For example, although most researchers agree that several boundary strengths must be assumed, there is no general agreement on issues such as the number and types of boundaries that need to be distinguished. This is perhaps reflected in the multitude of prosodic transcription systems available; several different systems have been proposed for Swedish (e.g. Bruce, 1995; Horne, Strangert & Heldner, 1995). Moreover, an extensive literature has shown that phenomena such as silent pauses, final lengthening and F0 resets are involved in the acoustic signaling of prosodic boundaries in Swedish (Bruce, Granström, Gustafson & House, 1993; Fant & Kruckenberg, 2002). Capturing such phenomena automatically in real-world speech, however, is a non-trivial task. Furthermore, it is also known that prosodic and linguistic structures are related (e.g. Strangert, 1990; Gustafson-Capkova & Megyesi, 2002), and prosodic boundaries in TTS systems are often predicted on the basis of content/function words, part-of-speech (PoS), or phrase structure (Ostendorf & Veilleux, 1994; Taylor & Black, 1998). Yet, we need to further explore what kind of linguistic features and the detail of analysis needed for making correct predictions about prosody.

In this paper, we investigate weak and strong perceived boundaries and their acoustic and linguistic context in spontaneous dialogs in Swedish. At present, the acoustic features reflect silent pauses and final lengthening. The linguistic features include information about content and function words, and parts-of-speech with and without subcategorization features.

2. Data and Method

The speech material consists of a radio interview of about 25 minutes (approx. 4100 words). The format is one interviewee and two interviewers. The interview contains examples of interactive dialog, as well as longer stretches of uninterrupted or monologue-like speech.

The speech material was manually annotated for perceived boundaries by three experienced transcribers. Each word was marked as being followed either by a weak or a strong boundary, or as not followed by a boundary. The inter-rater reliability of this task was fairly high; the pair-wise agreement in the three-way classification was 91% and the corresponding Kappa value 0.68. The agreement and Kappa values on presence vs. absence of a boundary were 94% and 0.77, respectively. However, to further increase the quality of the annotations, the perceived boundaries were determined by the majority votes of the three transcribers which resulted in 211 strong boundaries, 407 weak boundaries, 3459 no boundaries, and 25 cases of total disagreement among the annotators.

The segmentation of the speech material into words and phonemes was achieved by means of an automatic alignment algorithm developed at our department (Sjölander, 2003). The input to the aligner was a speech file and a verbatim transcription of the speech (including various disfluencies), supplemented by anchor points at approximately one-minute intervals. The output consists of two tiers marking words in standard orthography, and phonemes, respectively. The phoneme tier is supplemented with lexical prosodic features such as primary and secondary stress, and word accent type (i.e. accent I or II). The grapheme-to-phoneme conversion, as well as the lexical prosodic markup was accomplished with the KTH text-to-speech system.

A number of duration and pause features intended to capture final (or pre-boundary) lengthening and silent pause durations were extracted. These features included the (absolute) durations of the word, the word final rhyme, and any silent pauses after and before the word. In addition, four different normalized measures of duration were calculated for each constituent as the average z-score normalized segment durations across the constituents. The first two used standard z-score normalization with respect to inherent duration (Wightman, Shattuck-Hufnagel, Ostendorf & Price, 1992), with means and standard deviations either from all speakers or per speaker. The following two were variants including a compensation for speaking rate based on a moving window of ±15 segments.

The linguistic description of the transcribed speech materials included features that have been shown to be (partly) relevant for prosodic structuring. The linguistic features used in this study were: content (adjective, noun and verb) or function (others) word; and PoS, such as adjective (A), adverb (R), conjunction (C), determiner (D), disfluency (F), interjection (I), noun (N), numeral (M), particle (Q), preposition (S), pronoun (P), and verb (V). The words in each utterance were automatically annotated with part-of-speech, and manually post-edited where necessary.

3. Results and Discussion

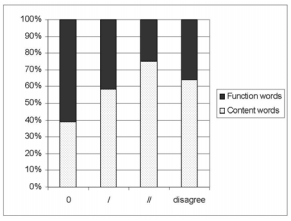
The three types of analysis of the material (the perceived boundaries, the acoustic and the linguistic features) were combined in various ways and analyzed. The mapping of the acoustic features onto the perceived boundaries (see Table 1) revealed, as expected that, (i) words and word-final rhymes before boundaries were longer than those not followed by a boundary, (ii) boundaries were characterized by longer silent pauses after the word than non-boundaries, and (iii) strong boundaries were characterized by longer silent pauses than weak boundaries. In addition, silent pauses before the word were slightly longer before no

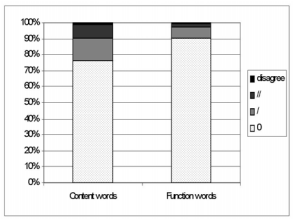
boundary words than before weak or strong boundary words. Somewhat to our surprise, however, the analyses also showed that words and word-final rhymes before weak boundaries were considerably longer than before strong boundaries, thus indicating relatively more final lengthening, and in a sense a stronger signaling before weak boundaries. These tendencies were all present in the absolute durations, but were generally more pronounced in the normalized duration measures, cf. Table 1. Due to lack of space, only the z-score measure without speaking rate compensation and using means and standard deviations from all speakers is presented. However, the other measures generally gave a poorer separation between boundary categories.

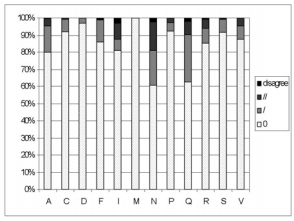
Table 1. Means (and standard deviations) of absolute duration (in ms) and average z-score normalized duration (in std devs) for the duration features. The z-scores were calculated without rate compensation, and using means and standard deviations from all speakers.

	No boundary		**Weak boundary**		**Strong boundary**	
	Duration	**z-score**	**Duration**	**z-score**	**Duration**	**z-score**
Word	278 (210)	-.13 (.70)	560 (309)	.50 (.90)	523 (276)	.17 (.65)
Word-final rhyme	122 (89)	-.15 (.84)	231 (147)	.96 (1.45)	199 (112)	.48 (.97)
Silence after	12 (45)	-.09 (.24)	188 (183)	.27 (.59)	362 (402)	.84 (1.65)
Silence before	51 (153)	-.01 (.55)	36 (107)	.02 (.31)	27 (102)	.02 (.30)

The distribution of content/function words, as well as of parts-of-speech for the different boundary types is shown in Figures 1-3. Figure 1 clearly shows that there is a relationship between boundary types and content/function words; the stronger the boundary, the more probable that the word before the boundary is a content word. However, we found only 21% of the content words before a boundary. Among function words, on the other hand, only 9% was found in a boundary position, as could be expected, cf. Figure 2. Our results indicate that various boundary types cannot be predicted by distinguishing between content and function words in spontaneous dialogs. However, we could make predictions about whether the word is a content or function word if we have knowledge about a boundary type that follows the word, which might be useful in speech recognition. Figure 3 shows that nouns and particles are the categories that most frequently occur before boundaries. Adjectives, interjections, disfluencies, adverbs, and verbs were followed by a boundary between 10% and 20% of the cases, while prepositions, pronouns, and determiners were only occasionally found in connection to a boundary. Numerals belong to the only category that never co-occurred with a boundary.







Figures 1-3. The distribution of content/function words, as well as of parts-of-speech (see Data and Method for abbreviations) for the different boundary types.

Lastly, to investigate the extent to which the acoustic and linguistic features model perceived boundaries, prediction experiments using discriminant analysis were conducted. The results are presented in Table 2. The best combination of acoustic features, the absolute duration of the silence after the word and the average z-score normalized duration of the word-final syllable rhyme, yielded 86.2% correctly classified cases. Similarly, the best

combination of linguistic features, the content/function word distinction, the PoS, and the following PoS, gave 66.3% correct results. Combining the best acoustic and linguistic features yielded slightly lower prediction accuracy (85.3%).

Table 2. Correctly classified cases (%) for no, weak and strong boundaries using acoustic and linguistic features, and a combination of both.

Boundary	Acoustic	Linguistic	Combination
No	93.6	72.6	92.5
Weak	42.5	14.0	42.3
Strong	48.8	63.0	49.3
Total	*86.2*	*66.3*	*85.3*

4. Conclusions

The study provides insight into the factors that govern the structuring of speech. It is viewed as a first step towards a more general model with applications in speech technology. In the near future, we plan to extend the analyses with additional acoustic and linguistic features, and our speech material with other speaking styles. We intend to add F0 movements to the acoustic features, and phrase and clause boundaries to the syntactic features. Finally, in order to improve the prediction of prosodic boundaries, we will explore other machine learning techniques better suited to combine continuous and discrete variables.

5. References

Bruce, G., Granström, B., Gustafson, K. & House, D. (1993) Interaction of F0 and duration in the perception of prosodic phrasing in Swedish. In *Nordic Prosody VI*, pp. 7-22. Stockholm.

Bruce, G. (1995) Modelling Swedish intonation for read and spontaneous speech. In *Proceedings ICPhS 95*, pp. 28-35. Stockholm.

Fant, G. & Kruckenberg, A. (2002) A new approach to intonation analysis and synthesis of Swedish. In *Speech Prosody 2002*, pp. 283-286. Aix-en-Provence.

Gustafson-Capkova, S. & Megyesi, B. (2002) Silence and discourse context in read speech and dialogues in Swedish. In *Speech Prosody 2002*. Aix-en-Provence.

Horne, M., Strangert, E. & Heldner, M. (1995) Prosodic boundary strength in Swedish: Final lengthening and silent interval duration. In *Proceedings ICPhS 95*, pp. 170-173. Stockholm.

Ostendorf, M. & Veilleux, N. (1994) A hierarchical stochastic model for automatic prediction of prosodic boundary location, *Computational Linguistics,* 20(1), 27-54.

Sjölander, K. (2003) An HMM-based system for automatic segmentation and alignment of speech. In *Proceedings of Fonetik 2003*. Umeå.

Strangert, E. (1990) Pauses, syntax, and prosody. In *Nordic Prosody V*, pp. 294-305. Turku.

Taylor, P. & Black, A. W. (1998) Assigning phrase breaks from part-of-speech sequences, *Computer Speech and Language*, 12, 99-117.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. & Price, P. J. (1992) Segmental durations in the vicinity of prosodic phrase boundaries, *Journal of the Acoustical Society of America,* 91(3), 1707-1717.