

Exploring the Prosody-Syntax Interface in Conversations

Mattias Heldner and Beáta Megyesi

Centre for Speech Technology, Dept. of Speech, Music and Hearing, KTH, Sweden

[*mattias|bea*]@speech.kth.se

ABSTRACT

The goal of this study is to investigate the structuring of speech in terms of prosodic boundaries in spontaneous dialogues in Swedish. 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.

Several kinds of information need to be taken into account to model the structuring of speech. First, we need a classification of prosodic boundaries. 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 reflected in the multitude of prosodic transcription systems available; ToBI [1] is perhaps the best known, but there are several alternative systems.

Second, we need an acoustic modeling of prosodic boundaries. Regarding the acoustic information, it is widely known that perceived boundaries are signaled by silent pauses, F0 resets, final lengthening, speaking rate variation, among others.

Furthermore, it is also known that prosodic and linguistic structures are related to each other, although the exact nature of this relationship is not clearly understood. For example, in text-to-speech systems, phrase breaks are often predicted on the basis of content/function words, part-of-speech (PoS) categories, or detailed but incomplete syntactic constituent structure (e.g. [2], [3], and [4]) indicating that there is some relation between the morpho-syntactic and prosodic structure. We need to explore in more detail what kinds of linguistic features and what detail of linguistic analysis that are needed for making correct predictions about prosody.

The various kinds of information needed to model the structuring of speech have been previously investigated. Several classification systems for prosodic boundaries based on auditive analysis have been proposed (see [5] and [6]). Studies concerning prosodic characteristics in the vicinity of boundaries include e.g. [7] and [8]. An intonation model described in [9] takes duration and pausing into account, as well as clause structure and part-of-speech (PoS). The relationship between syntactic structure and prosody has been investigated in [10] and [11], and the relationship between prosodic structure and prosody in [6].

In this paper, we investigate weak and strong perceived boundaries and their acoustic and linguistic context in spontaneous dialogues 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. This study is a first step towards an integrative model of the structuring of speech.

2 METHOD

2.1 SPEECH MATERIAL

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

2.2 PERCEIVED BOUNDARIES

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 (see [12] for agreement and Kappa methods).

However, to further increase the quality of the anno-

tations, the perceived boundaries referred to and used in the remainder of this paper were determined by the majority votes of the three transcribers. The majority votes resulted in 211 strong boundaries, 407 weak boundaries, 3459 no boundaries, and 25 cases of total disagreement among the labelers.

2.3 AUTOALIGNER

The segmentation of the speech material into words and phonemes was achieved by means of an automatic alignment algorithm developed at our department [13]. The input to the auto aligner was a speech file and a verbatim transcription of the speech (presently 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.

2.4 ACOUSTIC FEATURES

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 (e.g. [14]), with means and standard deviations either from all speakers or per speaker. The following two are variants including a compensation for speaking rate (e.g. [15]), based on a moving window of 15 segments.

2.5 LINGUISTIC FEATURES

The linguistic description of the transcribed speech materials includes features that have been shown to be (partly) relevant for prosodic structuring. The linguistic features used in this study are listed below.

- (i) the words
- (ii) content (adjective, noun and verb) or function (others) word
- (iii) part-of-speech (PoS): adjective (A), adverb (R), conjunction (C), determiner (D), disfluency (F), interjection (I), noun (N), numeral (M), particle (Q), preposition (S), pronoun (P), or verb (V)

- (iv) the same part-of-speech as listed in (iii) but some with subcategorization: adjective (AQ), participle (AF), relative/interrogative adverb (RH), other adverb (RG), conjunction (CC), infinitive “att” (to) (CI), subjunction (CS), relative determiners (DTR), other determiner (DT), common noun (NC), proper noun (NP), personal pronoun (PF), relative/interrogative pronoun (PH), or possessive pronoun (PS)

- (v) part-of-speech including morphological features such as number, person, case, etc.

The words in each utterance were automatically annotated with part-of-speech including morphological information, and manually post-edited where necessary. Then, the detailed description was mapped automatically into less detailed subcategories as described above in (ii–iv).

3 RESULTS AND DISCUSSION

The three types of analysis of the material (the perceived boundaries, the acoustic and the linguistic features) are combined in various ways. The relationship between the feature types will be presented next.

3.1 PERCEIVED BOUNDARIES AND ACOUSTIC FEATURES

The mapping of the acoustic features onto the perceived boundaries (see Table 1 for results) revealed, as expected, (i) that words and word-final rhymes before boundaries were longer than those not followed by a boundary, (ii) that boundaries were characterized by longer silent pauses after the word than non-boundaries, and (iii) that 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 the boundary categories.

	BOUNDARY TYPE		
	NO	WEAK	STRONG
ABSOLUTE DUR.			
Word	278 (210)	560 (309)	523 (276)
Word-final rhyme	122 (89)	231 (147)	199 (112)
Silence after	12 (45)	188 (183)	362 (402)
Silence before	51 (153)	36 (107)	27 (102)
AVG. Z-SCORE			
Word	-.13 (.70)	.50 (.90)	.17 (.65)
Word-final rhyme	-.15 (.84)	.96 (1.45)	.48 (.97)
Silence after	-.09 (.24)	.27 (.59)	.84 (1.65)
Silence before	-.01 (.55)	.02 (.31)	.02 (.30)

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

3.2 PERCEIVED BOUNDARIES AND LINGUISTIC FEATURES

The relationship between various boundary types and the distribution of content and function words are shown in Figures 1 and 2. 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.

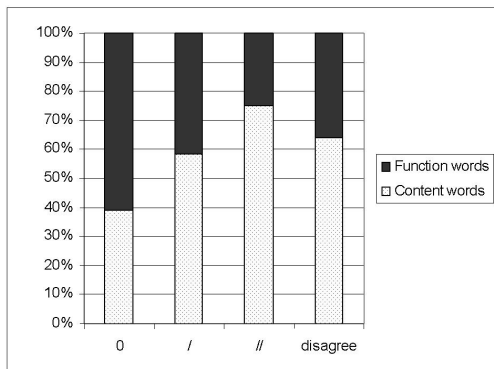


Figure 1: The distribution between content and function words for various boundary types.

However, we found only 21% of the content words occurring before a boundary: 9% before strong and 14% before weak boundary positions. Among function words, on the other hand, only 9% was found in a boundary position, as could be expected, of which 7% was located before weak and 2% before strong boundaries, cf. Figure 2. Our results indicate that various boundary types cannot be predicted on the basis of the distinction made between content and function words in spontaneous dialogues. 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.

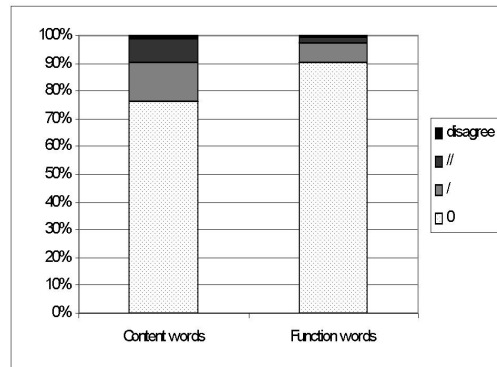


Figure 2: The distribution between different boundaries for content and function words.

The relationship between various parts-of-speech and boundary types is shown in Figure 3. Nouns (N), especially common nouns, and verbal particles (Q) are the categories that most frequently occur before weak or strong boundaries. Adjectives (A) (and above all participles), interjections (I), disfluencies (F), adverbs (R), and verbs (V) were followed by a boundary between 10% and 20% of the cases, while prepositions (S), conjunctions (C) (especially subjunctions), pronouns (P) (above all possessive pronouns), and determiners were only occasionally found in connection to a boundary (less than 10%). Numerals belong to the only category that never co-occurred with a boundary.

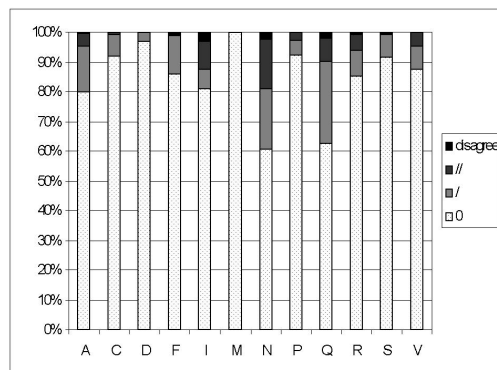


Figure 3: The relationship between parts-of-speech and various types of boundaries.

3.3 PREDICTING BOUNDARIES

To investigate the extent to which the acoustic and linguistic features model perceived boundaries, a number of prediction experiments was conducted using discriminant analysis. 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—content/function word distinction, the PoS as listed in (iv) in Section 2.5, and the following PoS—gave 66.3% correct results. Combining the best acoustic and linguistic features yielded slightly lower results, 85.3% correctly classified cases. Prediction using the feature set is thus slightly lower than the pairwise agreement among human transcribers.

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

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

4 CONCLUSION

The study provides insight into the factors that govern the structuring of speech. In addition, it is viewed as a first step towards a more general model with applications in speech technology; e.g. to predict boundaries from input texts for speech synthesis, and to predict boundaries from input speech for automatic speech recognition and understanding. For a perceptual evaluation of acoustic features see [16].

In the near future, we plan to extend the analyses with additional acoustic and linguistic features, and our speech material with other speaking styles. In particular, 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.

ACKNOWLEDGMENTS

The authors are grateful to the Swedish Radio for making the speech data and the transcriptions available, as well as to Kåre Sjölander for the aligner. This work was carried within the GROG project “Structuring of speech in different communicative situations” supported by The Swedish Research Council (VR).

REFERENCES

- [1] M. E. Beckman and G. Ayers Elam, “Guidelines for ToBI labelling,” <http://www.ling.ohio-state.edu/research/phonetics/E.ToBI/>, 1997.
- [2] M. Q. Wang and J. Hirschberg, “Automatic classification of intonational phrase boundaries,” *Computer Speech and Language*, vol. 6, pp. 175–196, 1992.
- [3] M. Ostendorf and N. M. Veilleux, “A hierarchical stochastic model for automatic prediction of prosodic boundary location,” *Computational Linguistics*, vol. 20, pp. 27–55, 1994.
- [4] P. Taylor and W. A. Black, “Assigning phrase breaks from part-of-speech sequences,” *Computer Speech and Language*, vol. 12, pp. 99–117, 1998.
- [5] G. Bruce, “Modelling Swedish intonation for read and spontaneous speech,” in *Proceedings of ICPHS-95*, 1995, vol. 2, pp. 28–35.
- [6] M. Horne, E. Strangert, and M. Heldner, “Prosodic boundary strength in Swedish: Final lengthening and silent interval duration,” in *Proceedings of ICPHS-95*, Stockholm, Sweden, 1995.
- [7] G. Fant and A. Kruckenberg, “Preliminaries to the study of Swedish prose reading and reading style,” in *STL-QPSR*, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden, 1989, vol. 2.
- [8] G. Bruce, B. Granström, K. Gustafson, and D. House, “Interaction of f0 and duration in the perception of prosodic phrasing in Swedish,” in *Nordic Prosody VI*, Stockholm, 1993, pp. 7–22.
- [9] G. Fant and A. Kruckenberg, “A new approach to intonation analysis and synthesis of Swedish,” in *Speech Prosody 2002*, Aix-en-Provence, France, 2002, pp. 283–286.
- [10] E. Strangert, “Pauses, syntax, and prosody,” in *Nordic Prosody*, K. Wiik and I. Raimo, Eds., 1990, vol. V, pp. 294–305.
- [11] S. Gustafson-Čapková and B. Megyesi, “Silence and discourse context in read speech and dialogues in Swedish,” in *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, 2002, pp. 363–366.
- [12] M. Poesio and R. Vieira, “A corpus-based investigation of definite description use,” *Computational Linguistics*, vol. 2, no. 24, pp. 183–216, 1998.
- [13] K. Sjölander, “Automatic alignment of phonetic segments,” in *Working Papers 49: Papers from Fonetik 2001*. Lund: Lund University, Dept. of Linguistics, 2001, pp. 140–143.
- [14] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, “Segmental durations in the vicinity of prosodic phrase boundaries,” *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1707–1717, 1992.
- [15] C. W. Wightman and M. Ostendorf, “Automatic recognition of prosodic phrases,” in *Proceedings of ISCAPS-91*, Toronto, 1991, IEEE, pp. 321–324.
- [16] R. Carlson and M. Swerts, “Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials,” in *Proceedings of ICPHS-03*, 2003.