

Boundaries and groupings - the structuring of speech in different communicative situations: a description of the GROG project

Rolf Carlson² Björn Granström², Mattias Heldner^{1,2}, David House², Beata Megyesi², Eva Strangert¹, Marc Swerts³*

¹Department of Philosophy and Linguistics, Umeå University; ²Centre for Speech Technology, KTH;

³Antwerp University and IPO; *Names in alphabetic order

Abstract

The goal of the project is to model the prosodic structuring of speech in terms of boundaries and groupings. The modeling will include different communicative situations and be based on existing as well as new speech corpora. Production and perception studies will be used in parallel with automatic methods developed for analysis, modeling and prediction of prosody. The model will be perceptually evaluated using synthetic speech.

Introduction

The objective of the GROG project¹ is to model the structuring of speech in terms of prosodic boundaries and groupings. This structuring is fundamental for spoken communication in that it reflects the speaker's internal organization of the information, and facilitates the listener's processing of the message (e.g. Gee & Grosjean, 1983). The later aspect is important for various speech technological applications. Moreover, the structuring of speech is known to vary in different communicative situations. Terken (2001) suggests that communicative situation can be characterized in terms of at least three dimensions: *the content, the speaker, and the communicative context*.

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. Moreover, the knowledge of how boundaries and groupings are signaled and perceived, and the relationship between those is far from complete. This is reflected in the multitude of prosodic transcription systems available; ToBI

(Beckman & Ayers Elam, 1997) is perhaps the best known, but there are several alternative systems.

Second, we need an *acoustic modeling of prosodic boundaries* tuned to different communicative situations. It is widely known that prosodic boundaries and groupings are signaled mainly by features such as silent or filled pauses, breathing, F0 movements, preboundary lengthening, speaking rate, intensity variation, voice quality and reduction phenomena. We also know that there is a complex interplay between e.g. tonal and temporal features in this signaling (Bruce et al., 1993a).

Third, several researchers have shown that morphosyntactic structure and prosody are related to each other, although this relationship is not clearly understood (e.g. Strangert et al., 1993; Ostendorf & Veilleux, 1994). The same holds for the relationship between discourse structure and prosody (e.g. Swerts & Geluykens, 1994). An area of specific interest (see *Production and perception studies* below) concerns the relations between discourse structure and prosody, where an interplay can be expected between prominence (focus) and boundary phenomena. Thus, we need to explore what kinds of linguistic features and what detail of linguistic analysis that are needed for making correct predictions about prosody. We also need to study *the relation between morpho-syntax, discourse and prosody*.

¹ GROG: *Gräns och gruppering – Strukturering av talet i olika kommunikativa situationer*, supported by The Swedish Research Council (VR) 2002–2004.

Previous research on Swedish

The areas outlined above have all been addressed in previous research on Swedish. Bruce (1995) includes a classification system for prosodic boundaries based on auditive analysis. Studies concerning prosodic characteristics in the vicinity of boundaries include Fant & Kruckenberg (1989), Bruce et al. (1993a). An intonation model suggested by Fant & Kruckenberg (2002) 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 by Strangert (1990) and Strangert et al. (1993), Gustafson-Capkova & Megyesi (2002) and the relationship between prosodic structure and prosody by Horne et al. (1995). Horne et al. (2001) concerns prosodic correlates related to topic structure of spoken discourse. Finally, speech style variation related to boundaries and grouping has been investigated by Strangert (1993) and Bruce (1995).

Methodology

The modeling aims at a structured and optimized description of the relations between the prosodic/perceptual annotations, on the one hand, and the acoustic and linguistic structure on the other. Thus, the model predicts *where* boundaries occur as well as *how* they can be realized in a given context. Such models can be evaluated in perceptual experiments with e. g. synthetic speech.

To achieve this we will rely on data from detailed studies of specific areas within the boundary/grouping framework. E.g. leaning on previous work within the research group we have begun studying pausing as related to syntactic and information structure, see the next section. Another approach is automatic methods. In this case sufficiently large prosodically annotated materials are needed. Furthermore, the automatic methods are required (or at least preferable) for efficient acoustic and linguistic analyses of these data sets.

Existing speech corpora (e.g. from spoken dialogue systems) will be used alongside new material supplied by the Swedish Radio. The new material will include speech from different communicative contexts spoken by a variety of speakers; read-aloud speech, spontaneous monologues and dialogues.

The analyses include: (i) auditive analysis and prosodic transcription of boundaries (and groupings), (ii) acoustic analysis in boundary regions, (iii) linguistic analysis, primarily in the form of automatic syntactic and discourse analysis.

The auditive analysis is fundamental given the goal of modeling naturally sounding speech. The auditive analysis will in principal consist of prosodic annotation using a transcription system for boundaries and groupings in Swedish based on perceptual criteria.

The automatic methods approach will lean on the development of a prosodic feature vector (PFV) combined with a linguistic feature vector (LFV). The PFV will contain the acoustic analysis of boundary and grouping phenomena. In general a PFV is supposed to cover durational phenomena such as preboundary (or final) and accentual lengthening, duration of pauses, and speaking rate. In addition, F0 and energy features will be used for describing phenomena such as boundary tones, F0 resets across boundaries, and voice quality etc. Similarly, the linguistic feature vector (LFV) will be developed for the linguistic analyses. The LFV will include information about the morphological, syntactic and discourse structure of the transcribed speech material and also measures related to breathing and planning.

The three types of analyses of the material (the prosodic transcriptions, the PFV, and the LFV) will be combined into an integrative model describing the relationship between perceived boundaries (and groupings), and their acoustic and linguistic features. Automatic methods using machine learning will be used for the selection and ranking of features in the target model. These procedures as well as a more detailed account for the development of the prosodic and linguistic feature vectors are described in the next section.

Work in progress

Prosodic annotation

The perceptually based transcription system for prosodic boundaries and groupings will borrow details from various other systems (e.g. Bruce et al., 1993b; Beckman & Ayers Elam, 1997; Buhmann et al., 2002). This system has to fulfil a number of requirements. The first is that the annotation system should be applicable to both non-spontaneous and spontaneous speech. Thus, the categories should include information about

various boundary strengths, as well as hesitations and disfluencies. Second, the system must be detailed enough to cover the phenomena of interest and at the same time reasonably efficient in terms of annotation speed.

Prosodic feature vector

The acoustic analysis will be performed and stored in the prosodic feature vector (PFV). A PFV approach enables a selection of various features in terms of importance for the signaling of boundary and grouping phenomena. In addition, the PFV might serve as an important component in a system for automatic classification and transcription of prosodic categories. The analysis in terms of a PFV is meant to be a general tool for automatic extraction of basic acoustic features from a speech signal and its associated segmentation. It is intended for analysis of a wide range of prosodic phenomena, including, but not limited to, prosodic boundaries and groupings.

In the development of a PFV for Swedish many details and features will be borrowed from previous work, and especially from the work on the use of prosodic feature vectors for automatic classification of accents and boundaries in English by Wightman & Ostendorf (1994). In addition to normalized duration features, such vectors typically also include various F0, pause and energy features.

Linguistic feature vector

A linguistic feature vector (LFV) will be used for the linguistic description of the transcribed speech materials. The LFV will include features that have been shown to be relevant for prosodic structuring. For example, in text-to-speech systems, phrase breaks are often predicted on the basis of content/function words, part-of-speech (PoS), or detailed but incomplete syntactic constituent structure (e.g. Wang & Hirschberg, 1992; Ostendorf & Veilleux, 1994; Taylor & Black, 1998) indicating that there is some relation between the morpho-syntactic and prosodic structure. Furthermore, it has also been shown that prosody and discourse structure are related to each other (e.g. Hirschberg, 2001). For example, phrases introducing new topics are often relatively louder and slower, and demarcated by some combination of silent pauses, low boundary tones and/or pitch range resets. Thus, discourse as well as the morpho-syntactic structure affects the realization of

prosodic boundaries. Therefore, the feature set will include information about morphology, syntax, and discourse. Different modules will take care of the various linguistic analyses. Each module will be built automatically using data-driven methods for efficient processing. The modules necessary for automatic linguistic analysis include a part-of-speech (PoS) tagger (Megyesi, 2001), a parser (Megyesi, 2002), a clause boundary detector, a grammatical function assigner, and possibly also a topic boundary detector.

Production and perception studies

In speech read aloud breaks reveal the phrasing of utterances by dividing the speech stream into structural chunks. In spontaneous speech, breaks may reflect the human planning and lexical retrieval process, when they occur before content words. Both the syntactic breaks in read speech and the non-syntactic breaks in spontaneous speech may be advantageous for the listener, giving clues to the structure of the utterance and the important words, respectively (Strangert, 1993).

The listener perspective is in the foreground also when pausing is used intentionally to draw attention to an important part of an utterance. Pausing here relates to new-given information; breaking up the speech stream by inserting a pause can be looked upon as a means to focus words. This kind of semantically governed pausing has been observed to occur in professional news reading (Strangert, 1993). The pauses occurred after grammatical function words and before the semantically heavy (new) words thereby helping to give them extra emphasis. Preliminary observations of communicative situations where speakers had to clarify misunderstandings by emphasizing words indicate that pausing may also occur *after* the emphasized word. Further, Selkirk (2002) reports phrase breaks (most of them with a pause) after words with contrastive focus.

The occurrence of pausing in relation to focus is one of the areas in which we want to go into detail. The work will be based on the previous observations and new data using an experimental dialogue paradigm (Ericson & Lehiste, 1995) in which subjects repeatedly have to respond to misunderstandings. In addition, spontaneous dialogue will be analyzed. The study aims at a better understanding of the conditions governing the use of pausing for

focusing and emphasis, in particular in cases of clarification in human dialogue.

Concluding remarks

The models, feature vectors and prosodically labeled corpora under development within the GROG project open possibilities for investigations of many kinds, both of a theoretical and of a practical nature. The models provide insight into the factors that govern the human structuring of speech. Detailed studies and studies based on automatic methods should contribute to this. In addition, the models may also serve as input for speech technology applications such as to model dialogue, to predict boundaries from input texts for speech synthesis, or to predict boundaries from input speech for automatic speech recognition and understanding. The prosodic and linguistic feature vectors allow investigations of interactions of features within and between the prosodic and linguistic feature sets.

References

- Beckman M E & Ayers Elam G. (1997). *Guidelines for ToBI labelling*. http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/.
- Bruce G (1995). Modelling Swedish intonation for read and spontaneous speech. In: *Proceedings ICPHS 95*. Stockholm, Sweden, 28-35.
- Bruce G, Granström B, Gustafson K & House D (1993a). Interaction of F0 and duration in the perception of prosodic phrasing in Swedish. In: *Nordic Prosody VI*. Stockholm, 7-22.
- Bruce G, Granström B, Gustafson K & House D (1993b). Phrasing strategies in prosodic parsing and speech synthesis. In: *Proceedings Eurospeech '93*. Berlin, Germany: ESCA, 1205-1208.
- Buhmann J, Caspers J, van Heuven V J, Hoekstra H, Martens J-P & Swerts M (2002). Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. In: *Proceedings LREC*. Las Palmas.
- Ericson D & Lehiste I (1995). Contrastive emphasis in elicited dialogue: Durational compensation. In: *Proceedings ICPHS 95*. Stockholm, 352-355.
- Fant G & Kruckenberg A (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR*: 1-83.
- Fant G & Kruckenberg A (2002). A new approach to intonation analysis and synthesis of Swedish. In: *Speech Prosody 2002*. Aix-en-Provence, France, 283-286.
- Gee J P & Grosjean F (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15: 411-458.
- Gustafson-Capkova S & Megyesi B (2002). Silence and discourse context in read speech and dialogues in Swedish. In: *Speech Prosody 2002*. Aix-en-Provence, France.
- Hirschberg J (2001). Communication and Prosody: Functional Aspects of Prosody. *Speech Communication: Special Issue on Dialogue and Prosody*.
- Horne M, Hansson P, Bruce G, Frid J & Filipsson M (2001). Cue words and the topic structure of spoken discourse: The case of Swedish *men* 'but'. *Journal of Pragmatics*, 33: 1061-1081.
- Horne M, Strangert E & Heldner M (1995). Prosodic boundary strength in Swedish: Final lengthening and silent interval duration. In: *Proceedings ICPHS 95*. Stockholm, Sweden, 170-173.
- Megyesi B (2001). Comparing data-driven learning algorithms for PoS tagging of Swedish. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*. Pittsburgh, PA, USA, 151-158.
- Megyesi B (2002). Shallow parsing with PoS taggers and linguistic features. *Journal of Machine Learning Research: Special Issue on Shallow Parsing*, 2: 639-668.
- Ostendorf M & Veilleux N (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics*, 20: 27-54.
- Selkirk E O (2002). Contrastive FOCUS vs. presentational focus: Prosodic evidence from right node raising in English. In: *Speech Prosody 2002*. Aix-en-Provence, 643-646.
- Strangert E (1990). Pauses, Syntax, and Prosody. In: K Wiik & I Raimo (Eds.), *Nordic Prosody V*, 294-305.
- Strangert E (1993). Speaking style and pausing. In: *PHONUM 2*. Umeå: Department of Phonetics, University of Umeå. 121-137.
- Strangert E, Ejerhed E & Huber D (1993). Clause structure and prosodic segmentation. In: *Proceedings of the Seventh Swedish Phonetics Conference*. Uppsala, 81-84.
- Swerts M & Geluykens R (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech*, 37: 21-43.
- Taylor P & Black A W (1998). Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language*, 12: 99-117.
- Terken J (2001). Variability and speaking styles in speech synthesis. In: E Keller & G Bailly & A Monaghan & J Terken & M Huckvale Eds, *Improvements in Speech Synthesis*. Chichester, UK: John Wiley & Sons, 199-203.
- Wang M O & Hirschberg J (1992). Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6.
- Wightman C W & Ostendorf M (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2: 469-481.