

On modeling of conversational speech

Eva Strangert

Department of Philosophy and Linguistics, Umeå University

Abstract

Data from the GROG project demonstrate specific strategies to handle planning problems in conversation, strategies that have to be taken into account upon modeling of naturally-sounding speech. The strategies are both structural – suspension of speech at positions revealing the syntactic form of the message – and prosodic – slowing down before suspension.

Background

Speech synthesis systems developed so far generate speech mostly modeled on read-aloud data. Synthesizing conversational speech is a far greater challenge and an important endeavor in order to understand how speech, and prosody in particular, is produced on-line and how speech synthesis should be generated in conversational systems. Research along this line would have to lean on insights on cognitive and linguistic processing as well as phonetics and speech technology. As far as cognitive-linguistic processing concerns, there are today various efforts aiming at a deeper understanding of the shaping of spontaneous monologue and dialogue speech. Clark and colleagues, for example, have studied the interactions between linguistic-prosodic and cognitive processing in a number of studies (Clark and Clark, 1977) and their *commit-and-restore model* (Clark and Wasow, 1998) outlining these interactions has had a great influence in speech research. In Sweden this interest in speech and language processing is reflected in research projects such as *Grammar in conversation: A study of Swedish*¹, *The role of function words in spontaneous speech processing*² (see Horne et al., 2003) and *Boundaries and groupings – The structuring of speech in different communicative situations (GROG)*³. The last one, see Carlson et al. (2002), and insights gained there form the primary basis for the discussion in this paper of some preliminaries for modeling of conversational speech.

Rules for conversational speech

We know that conversational speech differs in many respects from speech read aloud.

Planning problems result in hesitations, restarts and repetitions with drastic effects on prosody, and in particular on how boundaries, including pauses, are realized, and on their distribution as well. Interruptions, when searching for words, make the speaker produce syntactically less well-formed speech than in reading aloud. Though at first hand many of these characteristics of conversational speech may seem haphazard, they are not when looked upon in more detail.

This is substantiated in the commit-and-restore model and supported by data testing its predictions (Clark and Wasow 1998). The model predicts first that “speakers prefer to produce constituents with a continuous delivery”. (“Constituents” in this model primarily refer to noun, verb and prepositional phrases as well as to clauses and sentences.) That is, speakers aim at producing entire constituents without interrupting themselves. In cases where continuity *is* violated, which happens when speakers suspend speech within a constituent (as a result of planning problems, for example lexical search problems) speakers do so in a non-random way. According to the model, speakers make an initial commitment to what will follow, that is, they initiate the constituent before having decided on all of it. In doing so, they give clues to the listener about what kind of syntactic form the following message will have. This syntactic signaling occurs combined with pauses – silent or filled – as well as lengthened durations of the initial word(s). By such commitments the speaker signals that he/she is going to continue speaking.

When developing rules for modeling of conversational speech, predictions such as those above, if substantiated, should play a significant role. In the following, the focus will be of observations on boundaries and groupings in Swedish made within the GROG project. These observations have been more fully accounted for in Strangert (2004), Heldner and Megyesi (2003) and Carlson et al. (2004). The following brief overview concentrates on the syntactic and prosodic aspects of chunking – the grouping of words into constituents – occurring in conversational speech as reported on in Strangert (2004).

Swedish GROG data

The observations stem from a Swedish Radio interview with a well-known politician. The interview, about 25 minutes long and including about 4100 words, was annotated for *perceived* boundaries by marking each word as followed by a strong, a weak, or no boundary, giving 211 strong, 407 weak, and 3459 no boundaries, and in addition 25 unclear cases.

The material was further segmented and temporal data were extracted to capture prosodic boundary and pre-boundary characteristics. Measurements included word and word-final-rhyme durations as well as silent interval durations at boundary positions. The durations were given as absolute values and also, to be able to compare different words, as calculated average z-score normalized durations. (F0-data are presently extracted and will be included in the database in the near future.) Data moreover, included linguistic descriptions of the transcribed conversation. The linguistic features used to classify the words were: content-function word, part of speech and phrase structure. For a more detailed description of the database, including measurement procedures, see Heldner and Megyasi (2003).

Chunking

The chunks – sequences of words between boundaries – were predominantly short in the analyzed speech. This appears from Figure 1 containing the distribution for the entire conversation (618 chunks) with chunks ending with perceived strong (//) and weak boundaries (/) given separately.

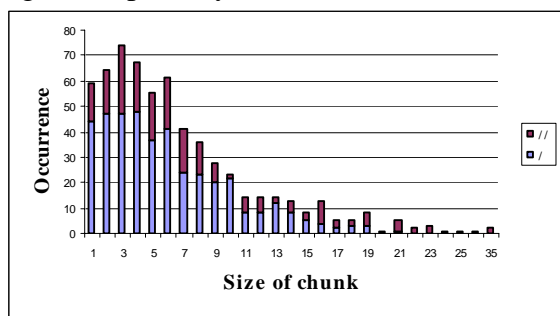


Figure 1. The distribution of size of chunk (number of words/chunk) separated for chunks ending with a strong (//) and weak boundary (/). Total number of chunks 618.

There is a preponderance of chunks with 2-4 words (with a maximum at 3 words), and even single-word chunks are very frequent. This

should be contrasted with a similar analysis of read speech (10 speakers, each reading a text of 810 words) in a material used to analyze pausing (Strangert, 1991). In the read speech, the distribution has its maximum at chunks with 7 words and also a preponderance for chunks with 3-9 words.

The longer chunks in read-aloud speech and the shorter in conversation can be looked upon in a continuity vs. violations-of-continuity perspective (Clark and Wasow, 1998). Thus, to find out to which extent chunks ended at “syntactically motivated” positions, that is, whether they had a continuous delivery or not (see introduction), perceived boundary positions were matched with parts-of-speech and phrase category markings. Here syntactically motivated (= continuous delivery) means (a) occurrence of a boundary *between*, rather than within constituents and (b) *before* constituent initial words rather than after.

Results showed that of the total of 618 chunks, almost 80% had endings coinciding with a syntactic boundary, while 117, slightly more than 20%, violated continuity by the occurrence of a boundary occurring in syntactically unmotivated positions. These positions and the frequency of occurrence of each appear from Table 1.

Table 1. Positions and occurrence of perceived boundaries in relation to syntactic structuring.

	Grammatical category	Occurrence
Within	Prepositional phrase	27
	Noun phrase	27
	Verb cluster	9
After	Subjunction	14
	Conjunction	11
	Infinitive mark	9
	Pronoun	8
	Adverbs	7
	Other	5

Quite apparently, when chunks end within a constituent, it happens close to the beginning in accordance with the hypothesis of initial commitment. Suspension mainly occurs after initial function words, the most frequent being prepositions, clause-initial conjunctions and subjunctions.

A more detailed analysis also showed that almost all cases of violation occurred at boundaries judged as weak (112 out of the 117 cases). Once again, the differences are striking comparing these with read-aloud data, where

less than 1% ended non-syntactically. Also, treating each size of chunk separately, the shortest chunks (1-3 words) have the highest incidence of violation. The shortest chunks often consist of just a single function word.

Prosody at chunk endings

Table 2 and 3 both show mean word and word-final-rhyme durations at the end of chunks as well as following silent intervals. All figures are given as average z-score normalized durations. The generally positive z-scores indicate longer than average durations – lengthening – at chunk endings. Though reflected in the word durations, the lengthening primarily occurs in the final part of words, as shown by the word-final-rhyme durations.

Table 2 shows the durations at the end of the differently-sized chunks. Chunks before weak and strong boundaries, respectively, are presented at the top and bottom of the table. For chunks ending before a weak boundary, there is a tendency of decreasing word and word final rhyme duration when the size of chunk increases. One-word chunks, in particular, stand out as having extreme durations. For chunks before a strong boundary, there is a similar tendency, although weaker, but the one-word chunks do not have similar excessively long durations. In addition, the durations are generally longer before weak boundaries than before strong. Thus, size of chunk as well as type of the following boundary affect the temporal structuring before the boundary. Silent intervals, on the other hand, appear to be unaffected by the size of chunks. Yet they differ consistently between strong and weak boundaries, being about half as long at weak as compared to strong boundaries. (See also Heldner and Megyesi, 2003.)

The extent to which the syntactic structuring affected prosody is demonstrated in Table 3 in which the durations for cases violating continuity is compared with cases of non-violation across all sizes of chunks. The speakers obviously behave differently in cases where chunk endings coincide with a syntactic boundary and when they do not. Word durations, and word-final-rhyme durations in particular, are longer in cases of violation. Silent intervals, on the other hand, are more or less unaffected.

Table 2. Mean duration of words and word-final rhymes before perceived boundaries and silent intervals for chunks of different size. Data (z-score normalized durations) given separately for weak (/) and strong (//) boundaries.

	Chunk size	Mean word dur	Mean word fin rhyme	Mean silence after	Occurrence, total
Before /	1	0,83	1,76	0,20	44
	2	0,51	0,95	0,15	47
	3	0,53	1,02	0,16	46
	4	0,70	1,18	0,20	48
	5	0,43	0,77	0,19	37
	6	0,61	1,14	0,21	41
	7	0,46	1,03	0,21	24
	8	0,34	0,69	0,21	24
	9	0,38	0,56	0,17	20
	10	0,25	0,64	0,24	22
	11	0,30	0,64	0,20	8
	12	0,21	0,73	0,27	8
	13	0,03	0,43	0,15	12
	14	0,21	0,16	0,13	9
	15	0,09	0,34	0,15	4
>15	0,06	0,48	0,17	14	
Before //	1	0,21	0,67	0,37	16
	2	0,25	0,59	0,39	17
	3	0,37	0,99	0,37	27
	4	0,48	0,71	0,35	20
	5	0,03	0,47	0,34	18
	6	0,11	0,36	0,45	21
	7	0,14	0,35	0,34	18
	8	0,13	0,31	0,34	11
	9	0,06	0,15	0,44	8
	10	-0,51	-0,73	0,03	1
	11	0,27	0,38	0,32	7
	12	0,19	0,61	0,41	6
	13	-0,11	0,37	0,35	3
	14	0,13	0,11	0,29	5
	15	-0,51	-0,34	0,20	3
>15	0,02	0,22	0,34	30	

Table 3. Mean duration of words and word final rhymes before perceived boundaries and silent intervals given separately for chunks with violation and non-violation of continuity. Data given as z-score normalized durations.

	Mean word dur	Mean word fin rhyme	Mean silence after	Occurrence
Violation	.83	1.26	.24	117
Non-violation	.28	.69	.25	501

Discussion and conclusions

The preponderance of chunks consisting of just a few words is characteristic for the material analyzed, setting conversational speech aside from read speech. This difference without doubt should be ascribed to the heavier

demands on on-line planning in conversational speech as compared to read.

Yet most of the chunks have the ideal continuous delivery assumed to be what speakers generally aim for. Also, when violations of continuity occur (in approximately 20% of the total number of chunks) they do not appear haphazardly, but rather in accordance with the strategies assumed by Clark and Wasow (1998). That is, suspensions primarily occur after initial subjunctions and conjunctions and in the initial part of phrases, primarily prepositional phrases and noun phrases. Also, most violations occur in chunks of 1-4 words with one-word chunks being the most affected.

The analysis here showed weak boundaries to be characteristically different from strong boundaries in that they had shorter silent intervals but at the same time longer word-final rhymes. That is, data reveal a trading relationship between lengthening and silent intervals (cf. Horne et al. 1995). This same pattern is evident across the different sizes of chunks. However, while silent intervals do not vary across the different sizes of chunks – although being consistently longer at strong as compared to weak boundaries – the lengthening of (final parts of) words are strongly affected by the size of chunk. There is a general trend, in particular at boundaries judged as weak, of increasing lengthening the less words in the chunk. Accordingly the one-word chunks again stand out from the rest, in this case by having the longest durations (most lengthening). Before weak boundaries, the one-word chunks even have extreme durations.

Cases of violation almost exclusively involved boundaries judged as weak, that is, boundaries with relatively short silent intervals but considerable final lengthening. Violations, moreover, predominated in chunks with just a few words, the chunks characterized by the most extreme lengthening. Thus planning problems resulting in suspensions of speech within constituents appeared to be characteristically signaled to the listener through excessively long durations before the suspension. Similar observations were made by Horne et al. (2003) in a study of disfluencies.

Thus, data so far have demonstrated very specific strategies to handle planning problems in speech production. In speech modeling these strategies have to be accounted for in order to

produce speech that reflects human processing in natural situations.

Acknowledgement

This work was supported by The Swedish Research Council (VR).

Notes

1. <http://www.tema.liu.se/Tema-K/gris/>
2. <http://www.ling.lu.se/projects/ProSeg.html>
3. <http://www.speech.kth.se/grog/>

References

- Carlson, R., Granström, B., Heldner, M., House D., Megyesi, B., Strangert, E., and Swerts, M. (2002) Boundaries and groupings – the structuring of speech in different communicative situations: A description of the GROG project. *Proc. Fonetik 2002, TMH-QPSR 44*, 65-68.
- Carlson, R., Swerts, M. and Hirschberg, J. (2004) Prediction of upcoming Swedish prosodic boundaries by Swedish and American listeners. *Proc. Speech Prosody 2004, Nara*, 329-332.
- Clark, H. H. and Clark, E.V. (1977) *Psychology and language: An introduction to psycho-linguistics*. New York: Harcourt Brace Jovanovich.
- Clark, H. H. and Wasow, T. (1998) Repeating words in spontaneous speech. *Cognitive Psychology 37*, 201-242.
- Heldner, M. and Megyesi, B. (2003) Exploring the prosody-syntax interface in conversations. *Proc. 15th International Congress of Phonetic Sciences, Barcelona*, 2501-2504.
- Horne, M., Frid, J., Lastow, B., Bruce, G. and Svensson, A. (2003). Hesitation disfluencies in Swedish: Prosodic and segmental correlates. *Proc. 15th International Congress. of Phonetic Sciences, Barcelona*, 2429-2432.
- Horne, M., Strangert, E. and Heldner, M. (1995) Prosodic boundary strength in Swedish: Final lengthening and silent interval duration. *Proc. 13th International Congress of Phonetic Sciences, Stockholm*, 170-173.
- Strangert, E. (1991) Pausing in texts read aloud. *Proc. XIIth International Congr. of Phone-tic Sciences, Aix-en-Provence*, 4, 238-241.
- Strangert, E. (2004) Speech chunks in conversation: Syntactic and prosodic aspects. *Proc. Speech Prosody 2004, Nara*, 305-308.