ON MODELLING AND SYNTHESIS OF CONVERSATIONAL SPEECH

Eva Strangert
Rolf Carlson

1 Introduction

In this paper we discuss some preliminaries for modelling and synthesis of conversational speech basing ourselves first and foremost on insights gained within the Swedish GROG project (Carlson et al., 2002).

Speech synthesis systems developed so far generate speech mostly modelled on read-aloud data. The next step, to synthesize conversational speech is a great challenge and an important endeavour which requires an understanding of how speech, and prosody in particular, is continuously produced. One important application of such knowledge is a more appropriate model of how speech synthesis should be generated in speech based 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 and their commit-and-restore model (Clark and Wasow, 1998) outlining these interactions has had a great influence in speech research. Besides the current GROG project this interest in speech and language processing is also reflected in the Swedish research project The role of function words in spontaneous speech processing (see Horne et al., 2003, 2004).

2 Modelling 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

1 http://www.speech.kth.se/grog/
2 http://www.ling.lu.se/projects/ProSeg.html
words, make the speaker produce syntactically less well-formed speech than in reading aloud. At first hand many of these characteristics of conversational speech may seem haphazard, although, when looked upon in more detail, they are not.

This is substantiated in the commit-and-restore model and supported by data testing its predictions (Clark and Wasow 1998). Initially the model predicts 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 signallng 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 modelling of conversational speech, predictions such as those above, if substantiated, should play a significant role. In the following, focus will be on observations on boundaries and groupings in Swedish made within the GROG project. The findings will be discussed in relation to the commit-and-restore model. After a short description of the database and the analysis procedures, we give an overview of earlier observations on the syntactic and prosodic aspects of chunking – the grouping of words into constituents. These findings have previously been more fully accounted for in Heldner and Megyesi (2003), Carlson et al. (2004), and in particular, in Strangert (2004). Next, we extend the analyses to deal with the prosodic-temporal domain as it relates to different boundary categories, function words and prosodic focus. We also discuss the implications for modelling and finally demonstrate some synthesis efforts.

3 The GROG data base and analysis procedures

The conversational data and analyses presented here stem from a Swedish Radio interview with a well-known female politician. The interview is part of a larger GROG database also including samples and analyses of read speech. The interview, about 25 minutes long and including 4102 words, was annotated for perceived boundaries. Three annotators marked each word as followed by a strong (/\), a weak (/), or no (\(\emptyset\)) boundary and, based on a majority vote, 211 strong, 407 weak, and 3459 no boundaries resulted. There were, in addition, 25 unclear cases reflecting total disagreement between the three annotators. Moreover, perceived focus on words was annotated (to be able to control for interactions between prosodic focus on the one hand and boundaries and groupings on the other). Also in this case, the same majority vote was used as for boundaries.
The material was further segmented and temporal data were extracted. Measurements included word and word-final-rhyme (last vowel and all following consonants) 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 normalized durations. The normalized measures were calculated as the average z-score normalized segment durations across the word and word-final rhyme, respectively.

Data moreover, included linguistic descriptions of the transcribed speech. The linguistic features used to classify the words were: content-function word, part of speech and phrase structure. For a description of the interview data as well as some descriptive statistics, see also Heldner and Megyesi (2003).

4 Summarizing findings of chunking in the interview data

In this section we will shortly summarize the most important findings of chunking – the grouping of words between boundaries – in the interview data.

4.1 Chunk length

The chunks were predominantly very short. Of a total of 618 chunks, 211 were followed by a strong boundary (//) and 407 by a weak boundary (/). There was a preponderance of chunks with 2-4 words, but single-word chunks were also 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 had its maximum at chunks with 7 words and a preponderance for chunks with 3-9 words.

4.2 Continuous delivery

Almost 80% of all chunks had a continuous delivery, while the remaining 20% violated continuity by the occurrence of a boundary placed in syntactically unmotivated positions. Here continuous delivery (= syntactically motivated) means (a) occurrence of a boundary between, rather than within constituents and (b) before constituent initial words rather than after.

4.3 Violation and weak boundaries

Cases of violation predominantly included boundaries occurring within prepositional and noun phrases and after subjunctions and conjunctions. Thus 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). In accordance with the hypothesis of initial commitment, suspension mainly occurred after initial function words, the most frequent being prepositions, and clause-initial conjunctions and subjunctions. Also, the shortest chunks (1-3
words) had the highest incidence of violation, the one-word chunks being the most extreme.

Almost all cases of violation occurred at boundaries judged as weak. Of the total of 117 cases of violation, 112 occurred at weak boundaries. Once again, the differences are striking, comparing these figures with read-aloud data (Strangert, 1991), where less than 1% ended non-syntactically.

4.4 Silent intervals
Silent intervals varied between strong and weak boundaries, but were unaffected by size of chunk. At strong boundaries they were about twice as long as at weak boundaries (see also Heldner and Megyesi, 2003).

4.5 Durations
Word and in particular word-final- rhyme durations were affected by boundary type and size of chunk. For chunks ending before a weak boundary, there was a tendency of decreasing durations when the size of chunk increased. One-word chunks in particular stood out as having extreme durations. For chunks before a strong boundary, there was a similar tendency, although weaker, but the one-word chunks did not have similar excessively long durations. Also, the durations were generally longer before weak boundaries.

Thus, combining the observations of silent intervals (Section 4.4) and word and word-final rhyme durations, a trading relationship emerged: More lengthening combined with shorter silent intervals (at weak boundaries), and less lengthening combined with longer silent intervals (at strong boundaries).

We believe this pattern has to do with the conversational style studied here and that its primary basis is the slowness that follows from hesitations and other kinds of disfluencies, a slowness that often also affects the word as a whole. That is, we seem to deal here with something quite different from the final lengthening occurring primarily in more structured speech.

4.6 Excessive lengthening
There was more excessive lengthening upon violation of continuity and less, when syntax was respected. Silent intervals, however, were unaffected by syntactic structuring.

5 Extending the analyses
The observations summarized above demonstrate very specific strategies, strategies that we ascribe to the on-line planning of speech in conversation. In the following we extend the analysis of the interview material to strengthen the basis for modelling conversational style. We analyze prosodic focus and how it relates
to different boundary categories. We also deal with patterns of lengthening within and across chunks as well as the behaviour of some function words.

5.1 Prosodic focus/non-focus, boundaries and lengthenings

In Swedish default focus position is held to be on the phrase final word, while emphasized and contrastively focused words may occur in any position in the phrase. Here we present basic interview data on the occurrence of focused words. For reference we also include comparative data on non-focused words. Table 1 shows the distribution of focused and non-focused words in different boundary positions as well as mean z-score normalized word-final-rhyme durations. The 25 cases of total disagreement among the annotators is left out in this table.

Table 1. Focused and non-focused words in different boundary positions and mean final-rhyme durations (z-scores).

<table>
<thead>
<tr>
<th>Focused words before</th>
<th>Number of occurrences</th>
<th>Final rhyme duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>strong boundary (/)</td>
<td>104</td>
<td>0,72</td>
</tr>
<tr>
<td>weak boundary (/)</td>
<td>157</td>
<td>1,12</td>
</tr>
<tr>
<td>no boundary (Ø)</td>
<td>269</td>
<td>0,31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-focused words before</td>
<td></td>
<td></td>
</tr>
<tr>
<td>strong boundary (/)</td>
<td>107</td>
<td>0,24</td>
</tr>
<tr>
<td>weak boundary (/)</td>
<td>250</td>
<td>0,86</td>
</tr>
<tr>
<td>no boundary (Ø)</td>
<td>3190</td>
<td>-0,19</td>
</tr>
</tbody>
</table>

As is evident focused words occur about equally often at chunk endings, that is before a boundary, as within chunks. Also, more focused words occur before weak boundaries than before strong. Recall however, that the total number of weak boundaries in the material is about double that of the number of strong boundaries. In total 530 of the 4102 words in the interview (13%) are focused.

Generally, focused words are lengthened (lengthening defined here as a positive mean z-score duration), while the non-focused have generally shorter durations and in addition appear to fall into two groups, lengthened and non-lengthened. However, looked upon in more detail the pattern is more complex. This is demonstrated for focused and non-focused words in Figure 1a and 1b, which show cumulative distributions of z-score word-final-rhyme durations. The three curves in each figure represent the durations of words before weak, strong and no boundaries, respectively and the vertical line at 0 (y-axis) marks the boundary between non-lengthened and lengthened words.

In spite of the general pattern of lengthening of focused words, quite a number (183 out of the 530 focused words, i.e. 35 %) have negative z-score durations. As
focused words are generally held to be prolonged in Swedish (at least in well structured read speech, see Heldner, 2001), the relatively high percentage here is somewhat surprising, but could as well reflect a more general pattern, as to our knowledge observations are sparse here both concerning spontaneous and (non-laboratory) read speech.

Figure 1. Final rhyme duration in different boundary positions

Lengthenings, in addition, affect non-focused words, not only words at chunk endings (cf. Table 1), which is to be expected, but also words within chunks, that is, words not followed by a boundary. This is illustrated in Figure 1b (“no-break” category). 25% (805 of the 3190 non-focused words within chunks) fall into this category. This rather high percentage we look upon as a conversational characteristic which, in combination with the abundance of breaks/perceived boundaries and resulting short chunks, set this style of speech aside from read speech. In the following we will look at these lengthenings in more detail.

5.2 Lengthenings across chunks

Preliminary analyses indicated that lengthenings often affected several words in a row in the interview. It was therefore motivated to find out to which extent whole sequences of words were lengthened. In the following analysis, sequences of lengthened words spanning single strong and weak boundaries are studied as well as sequences spanning two or more boundaries. In doing so, we restrict ourselves to reporting only word-final-rhyme durations, even though there is ample evidence that lengthenings span the entire word in our material. As before (see Section 5.1) we define lengthening as equal to a positive mean z-score duration.

While there are cases of lengthenings spanning up to four boundaries in the interview, lengthening across one single boundary or two boundaries is most frequent. Table 2, which gives an overview of the extent of cross-boundary lengthenings, shows that in the majority of cases lengthenings affected sequences of two words. But occasionally, more than five adjacent words were lengthened.
For lengthenings spanning just one boundary, a division is made between strong and weak cases. As is evident, the weak boundaries are more frequently crossed by lengthened words and also by longer sequences of lengthenings than strong boundaries (cf. Section 4.5 above).

Table 2. Sequences of lengthened words spanning boundaries.

<table>
<thead>
<tr>
<th>No of boundaries spanned</th>
<th>Occurrences</th>
<th>Occurrences separate for different no of words affected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1-strong</td>
<td>41</td>
<td>26</td>
</tr>
<tr>
<td>1-weak</td>
<td>74</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>63</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

In addition to the lengthenings crossing one or several boundaries, sequences of lengthened words occur in other positions too: before boundaries, after boundaries and frequently also in other positions in chunks. These lengthenings as well as those covered in Table 2 have to be accounted for in modelling conversation. They form part of the “disfluency spectrum” besides silent and filled pauses, restarts, repetitions etc. However, although disfluencies are a very apparent characteristic of conversational speech, speaking spontaneously also involves long fluent periods - recall that 80% of the chunks in the interview had a continuous delivery. The challenge thus is to model the mixture of fluent and disfluent stretches of conversational speech.

5.3 Function words and lengthenings

As noted in Section 4.3, violations often occurred after phrase- and clause-initial function words. This motivated a more detailed analysis of such words. However, due to the restricted material in the interview, most function words have few tokens. Therefore, we focus here on the two most frequently occurring words, \textit{att} ‘that’ (148 tokens) and \textit{och} ‘and’ (128 tokens). Figure 2 shows cumulative distributions of the two words, the most apparent difference between them being a wider range of durations for \textit{och} than for \textit{att}.

The same words have previously been examined in great detail by Horne et al. (2003, 2004). Among other aspects, they analyzed durations of these words in disfluent (before a perceived pause) and fluent environments based on a considerably larger corpus (in total 1279 cases of \textit{att} and 644 cases of \textit{och}) than in our interview. Comparing these environments they observed both \textit{att} and \textit{och} to be considerably longer (130 ms) when disfluent. The percentages of disfluent words, however, differed (38% for \textit{att} and 68% for \textit{och}).
These figures are dramatically different from those of our data. First, there are few cases of att ‘that’ and och ‘and’ before perceived boundaries. For och there is just one case and for att 16 (with a difference in duration of +85 ms as compared to the cases not followed by a boundary). The lengthenings we find in our material then are not generally tied to the pre-boundary position.

However, this type of lengthening is not restricted to att and och, but occurs for other function words and for content words as well (cf. Eklund, 2004). As they go unnoticed by listeners (see the “no-focus”, “no-boundary” cases in Figure 1) they most reasonably reflect planning problems rather than the type of listener-oriented strategy, assumed in the initial-commitment hypothesis. For a discussion of the intentionality of disfluency see Nicholson et al. (2003).

We attribute these divergences between our findings and those of Horne et al. (2003, 2004) to differences of corpora. Horne and colleagues analyzed a monologue style of non-professional speech produced primarily in private homes. Our interview, though spontaneous, was produced in a radio studio and both parts were more or less professional speakers. The interviewed person was a politician, used to speak in public. In these situations a more fluent style of spontaneous speech is required. The high fluency rate – and the low percentage of violations of continuity – may very well be ascribed to this public style of conversational speech. This dimension of formality of spontaneous speech has to be considered also in speech modelling.

6. Synthesis

Data so far have demonstrated specific strategies to handle planning problems in speech production. These results have to be accounted for in a synthesis model that reflects human processing in natural situations. As an example of spontaneous speech, Figure 3 shows a spectrogram of a sample from the database.
(...) Sverige ska kräva / att / ...). The sample is taken from the passage: Jag tycker att Sverige ska kräva / att / bombningarna... ‘I think that Sweden should insist (on) / that / the bombardment ...’. The sample includes a focused word (in bold) and two weak boundaries of which one was caused by a disfluency.

Figure 3. Spectrogram of the utterance “..... Sverige ska kräva / att / ....”.

As a baseline version the passage was first synthesized using a text input including two types of boundary markers, a strong boundary (//) corresponding to “.”, and a weak (/) boundary corresponding to “,”. As a synthesis model we used Mbrola synthesis with a female voice and the default read speech rules in the KTH text-to-speech system. Figure 4 shows the synthesis of the same sample as in Figure 3.

Figure 4. Read speech synthesis of “..... Sverige ska kräva / att ....”

In the next step, we approached conversational synthesis. We used as input an orthographic transcription of the interview with focused words and perceived strong and weak boundaries marked. For the strong (//) and weak (/) boundary realizations the same rules were used as for “.” and “,” respectively, in the read speech transcription. A ”disfluency label” for weak boundaries in non-syntactic
positions (\#) was also used in the input transcription. This label marked a modified weak boundary realized with primarily extra lengthening of the entire word in the pre-boundary position and in addition an even higher factor of lengthening of the word final rhyme compared to the regular weak boundary. The \#-boundary was also modelled with a shorter silent interval and a very slight F0-fall instead of a rise. Figure 5 illustrates the resulting speech synthesis after these modifications.

![Figure 5](image.png)

Figure 5. Conversational synthesis “..... Sverige ska \textit{kräva} / att \# ....”

### 7. Final remarks

Informal listening tests indicated that the last synthesis version with disfluency characteristics had a higher degree of naturalness and actually had some similarity with our reference speaker. The modelling was done in line with the empirical findings summarized in the paper. This result encourages us to continue the work on speech synthesis modelling of disfluently speech. The reported work is just a pilot experiment which points to some of the most important factors which are needed to model conversational speech.

### Acknowledgement

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


