PRODUCTION AND PERCEPTION OF PAUSES AND THEIR LINGUISTIC CONTEXT IN READ AND SPONTANEOUS SPEECH IN SWEDISH

Beáta Megyesi
Centre for Speech Technology (CTT)
Department of Speech, Music and Hearing
KTH, Stockholm, Sweden

Sofia Gustafson-Čapková
Computational Linguistics
Department of Linguistics
Stockholm University/GSLT, Sweden

ABSTRACT
We investigate the relationship between prosodic phrase boundaries in terms of pausing and the linguistic structure on morpho-syntactic and discourse levels in spontaneous dialogues as well as in read aloud speech in Swedish. Both the speakers’ production and the listeners’ perception of pauses are considered and mapped to the linguistic structure.

1. INTRODUCTION
To be able to analyze and generate the structure of the speech, we need not only prosodic features but also grammatical analysis on various linguistic levels. Several researchers have discussed how prosody, morpho-syntax and discourse structure are related to each other (e.g. [5], [18], [12], [17], [19], [13]), although this relationship is not clearly understood as was pointed out in e.g. [2], [15] and [12].

Linguistic structure, for example, has been shown to play an important role in pausing strategies which signal information flow of the utterance, thereby helping the listener to interpret the message uttered by the speaker [4]. Swerts and Geluykens [16] found that speakers in monologues use pauses of various lengths to signal information flow in terms of topic structure. Shriberg et al [13] reported that new topics are often realized by some combination of silent pauses, low boundary tones and/or pitch range resets in English. Also, Hirschberg [9] argued that phrases introducing a new topic can be characterized by an initially wider pitch range preceded by a longer pause, and on average they are louder and slower than other phrases. Van Donzel [19] studied prosodic features of discourse boundaries for Dutch on the basis of clause, sentence and paragraph division, as well as the prosodic features of information structure in terms of the New–Given taxonomy. She found that discourse boundaries in spontaneous speech are realized by silent pauses and high boundary tones. These studies show that there is a relationship between prosody and (at least) higher linguistic structure, such as discourse in terms of topic, theme, and New–Given taxonomies.

In addition, several researchers have investigated the relationship between morpho-syntactic structure and prosody. Most of the studies deal with the automatic prediction of prosodic phrase boundaries, given some linguistic information, used in text–to–speech systems. Some studies show that full syntactic analysis is not needed for the prediction of prosodic boundaries, while others claim the opposite. For example, in text–to–speech systems, phrase breaks are often predicted by distinguishing between content and function words. For prosodic phrase boundary detection, detailed but incomplete parses were used by Bachenko and Fitzpatrick [1] by implementing the Phi rule-based algorithm developed by Gee and Grosjean [4]. Wang and Hirschberg [18] as well as Ostendorf and Veilleux [12] used PoS and syntactic constituent structure together with some acoustic information (such as pitch accent, phrase duration, and position to the last break) to predict phrase breaks. However, Ostendorf and Veilleux reported that good phrase prediction can be achieved without using any detailed PoS, or syntactic information. Taylor and Black [17] assigned phrase breaks on the basis of part-of-speech sequences only, although they suggested that syntactic parsers giving reliable parse trees might facilitate phrase break assignment. These studies show that prosodic phrase boundaries do not necessarily correspond to syntactic phrase boundaries.

Most of the researchers agree that there is a relation between prosody and syntactic structure on one hand, and between prosody and discourse structure on the other hand. However, most of the studies performed on this topic investigate one of these relations either for non–spontaneous or for spontaneous speech.

The aim of this study is to investigate some aspects of the relation between the prosodic, syntactic and discourse structure in spontaneous as well as in non–spontaneous speech. Additionally, both the speakers’ production and the listeners’ perception of pausing are considered and mapped to the linguistic structure. For non–spontaneous speech, we use professional news announcement, and non–professional news reading, and for spontaneous speech we study elicited dialogs. Since prosodic phrase boundaries are often marked by, among other prosodic features, pauses, we investigate the pausing strategies in the speaking styles. Since various studies have shown that prosody might signal discourse structure in terms of topic structure ([9], [16]), we use the theme shift vs. theme continuation taxonomy for discourse segmentation.

Next, we will describe the data and method used for investigating the relationship between pausing and linguistic structure.

2. DATA AND METHOD
In this study, we use the same speech data for each speaking style as we used in our previously reported studies (see [10], and [6]). The material of read speech consists of recordings of Swedish Radio news [11] read by four professional and four non–professional readers. The spontaneous speech data [7] consists of recordings of two Swedish map task dialogs, each with two dialog participants. The data sets include 920 words each.

In order to investigate the duration, frequency and position of acoustic pauses, the speech data was processed automatically by
a pause detector. Silent intervals longer than or equal to 100 ms were defined as the acoustic correlate for pausing. Pauses may include natural physical phenomena such as breathing and swallowing. However, particles expressing feedback/back-channeling (e.g., mm, aaa, aha) in dialogs are not allowed inside pauses. The automatic detection was manually checked in order to obtain consistency.

To examine the frequency and location of the perceived pauses, eighteen subjects annotated the position of what they identified as a pause. They were asked to use different labels for long and short pauses, and also marks where they were uncertain. Of the eighteen subjects, there were eight females and ten males belonging to different age groups. Eleven subjects had some knowledge about linguistics but none of them had ever participated in a similar experiment.

In this study, for a definition of a discourse segment we use the notion of theme. Theme shift (TS) is the position where a new theme is introduced in the discourse. As the basis for our investigation of the discourse context regarding theme shift, we asked five new subjects to annotate the transliterated text materials for theme shift. In order to keep discourse and prosody apart, the annotators were not allowed to listen to the audio data. Inter-annotator agreement was computed for all materials and gave a kappa value of $K = 0.82$ for the news texts, and $K = 0.79$ for the dialogs. In both cases, the values indicate high inter-annotator agreement. However, the discourse boundaries can be treated as more or less strong since the annotators do not necessarily agree on the boundary. Therefore, where appropriate, we will refer to the continuum of discourse boundary in the notion of theme shift (TS), and the opposite term theme continuation (TC), as follows:

- No boundary – none of the five subjects labeled a TS, i.e. theme continuation (TC)
- Weak boundary – one or two subjects annotated a TS
- Strong boundary – three or four subjects labeled a TS
- Extra strong boundary – all five subjects agreed on a TS

Additionally, in order to examine the syntactic characteristics of the text materials, the words in each text material were automatically tagged with their part-of-speech tag including morphological information. Then, each sentence in the texts was automatically parsed on the basis of its phrasal constituent structure. The labels for the constituents include major phrase categories, e.g. adverb (AdvP), adjective (AP), noun (NP) and prepositional (PP) phrase, as well as maximal projections, e.g. coordinated noun phrases, or a noun phrase with a prepositional phrase attached to it.

3. RESULTS

In this section, the relationship between the production of pauses and the linguistic context in which pauses appear, as well as the perception of the pauses and the linguistic environment in which people actually perceive them will be presented for the three speaking styles: professional and non–professional news reading, and dialogs.

3.1. Production and linguistic features

The characteristics of the silent intervals detected in the three speaking styles will be described in this section with special attention directed to the discourse and syntactic context in which pauses appear.

Table 1 shows the mean duration and word/pause ratio of silent intervals. Although there are differences in the duration and frequency of pauses between the three speaking styles, the total duration of the speech material is approximately the same for each style. Hence, the time it takes to pronounce a word on average differs between the styles suggesting greater variation in speech tempo. Our results, indicating the variation of pausing patterns across speaking styles, are consistent with the results reported in [3], [8], [14].

<table>
<thead>
<tr>
<th>Speaking style</th>
<th>Pause Duration</th>
<th>Word/Pause Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional</td>
<td>271 ms</td>
<td>77 (920/12)</td>
</tr>
<tr>
<td>Non–professional</td>
<td>561 ms</td>
<td>8.4 (920/110)</td>
</tr>
<tr>
<td>Dialog</td>
<td>538 ms</td>
<td>5.5 (920/167)</td>
</tr>
</tbody>
</table>

In order to describe the correlation between silent intervals and theme shift in the three speaking styles, recall and precision rates were counted. Recall (R) describes the percentage of pauses that appear in strong or extra strong boundary positions, see Definition 1. Precision (P), on the other hand, gives the percentage of (extra) strong boundaries that corresponds to pauses in the speaking styles, see Definition 2.

\[ R = \frac{\text{# Pauses appearing at (extra) strong boundaries}}{\text{Total # of pauses}} \] (1)

\[ P = \frac{\text{# Pauses appearing at (extra) strong boundaries}}{\text{Total # of (extra) strong boundaries}} \] (2)

The correlation between silent intervals and discourse boundaries for each speaking style is shown in Figure 1. It is obvious that theme shift and pauses do not always coincide, and we find clear differences between the speaking styles. In the reading styles pauses often appear at strong discourse boundary positions (shown by the high recall rate), while in the dialogs the majority of pauses can be found in weak or non–existing discourse boundary positions (low recall). On the other hand, looking at the discourse boundaries and their acoustic correlate in terms of silent intervals (i.e. the precision rates) we find that non–professional readers and dialog participants mark the majority of discourse boundaries with a pause, although the non–professional readers do this to a greater extent (92% and 57% precision, respectively). Professional readers, on the other hand, pause only in few cases at discourse boundaries (13%).

When it comes to the linguistic context of the acoustic pauses, the majority of silent intervals can be found at sentence boundaries, and in the case of dialogs at turn taking. In professional reading, pauses appear entirely at sentence boundaries, while in non–professional news reading, pauses are located at sentence (62%) as well as clause boundaries (18%), and also in front of noun– or

---

1 In the case of dialogs, a sentence is defined as a turn.

2 100% recall means that all pauses appear at (extra) strong boundaries, 100% precision means that all (extra) strong boundaries are realized as pauses.

3 The reader should note that the number of pauses is considerably fewer in professional reading than in the other speaking styles.
prepositional phrases, or in case of topicalization in front of finite verbs. In dialogs, on the other hand, acoustic pauses appear mainly at turn taking (28%), but also between phrases, e.g. in front of noun phrases (16%), adverb phrases (10%), conjunctions (10%) and prepositional phrases (9%).

3.2. Perception and linguistic features

In order to investigate how often and in what linguistic context people actually perceive silent intervals, the frequency and position of the perceived pauses were examined.

The distribution of the perceived pauses, labeled by the eighteen subjects, are to a large extent evenly distributed across the speaking styles (as opposed to the distribution of silent intervals), see [10] for more detail. The average “words per perceived pauses ratio” is highest in the professional reading (12.2) followed by the dialogs (11.4), and lowest in the non-professional reading (8.2).

Concerning the relation between perceived pauses and discourse boundaries (TS), recall and precision rates are given in Figure 2, just as it was presented for the silent intervals and given in Definition 1 and 2. The results reported here are based on instances where at least 14 of the 18 subjects perceived a pause. In the reading styles the majority of the perceived pauses are located at (extra) strong boundaries (high recall), while in the dialogs we found the majority of perceived pauses at theme continuation, or in weak discourse boundary positions (low recall). From the perspective of discourse, the perceived pauses are rare at strong boundary positions in spontaneous speech, as shown by low precision, while more frequently occurring in read speech.

Additionally, in the reading styles pauses are perceived at sentence and clause boundaries entirely, while in dialogs they are located between phrases, e.g. in front of NP:s, AdvP:s, PP:s, conjunctions or verbs, as well as within phrases, e.g. in front of nouns.

Even though the linguistic context of the perceived pauses is similar to the context described for the silent intervals for the various speaking styles, the acoustic and perceived pauses do not necessarily overlap.

3.3. Production and perception of pausing

To give an overall picture of the correlation between the silent intervals and pauses perceived by each of the 18 subjects participating in the listening test, recall and precision are measured. Here, recall \( R \) describes the percentage of the acoustic pauses that were actually perceived (see Definition 3), while precision \( P \) gives the percentage of perceived pauses that corresponds to acoustic silence (see Definition 4).

\[
R = \frac{\text{Total \# of acoustic and perceived pauses}}{\text{Total \# of acoustic pauses}} \quad (3)
\]

\[
P = \frac{\text{Total \# of acoustic and perceived pauses}}{\text{Total \# of perceived pauses}} \quad (4)
\]

The average mean of the recall and precision rates, computed on the basis of each subject’s individual rates, for each style is shown in Figure 3. It is clear that the correlation of the acoustic and perceived pauses varies across the speaking styles. In the professional reading, a considerable number of pauses are perceived by the subjects, but many of the perceived pauses do not have any correlates in acoustic silence. In the non-professional reading, the majority of the acoustic pauses are perceived by the listeners, and many of the perceived pauses actually have an acoustic correlate. In the dialogs, on the other hand, few pauses are perceived but many of the perceived pauses correspond to acoustic silence.
4. CONCLUSION AND DISCUSSION

In this study, we investigated the relationship between pausing pattern, syntactic features and discourse boundaries in read aloud and spontaneous speech. We have shown that there are considerable differences between the speaking styles concerning both the production and perception of pauses, as well as the linguistic context in which pauses appear. In the reading styles, pauses are found in strong discourse boundary positions mainly between sentences, while in dialogs the majority of the pauses appear at weak boundary positions or at theme continuation between or within phrases. Further, in reading pauses are perceived by the majority of the listeners at strong boundaries, while in dialogs few pauses are perceived, and few perceived pauses are found in strong boundary positions. Thus, in the reading styles the relation between pauses and linguistic structure (syntactic and discourse) is evident, while in dialogs a clear relation cannot be established in our data. However, the strength of the relationship between pauses and discourse structure differs between the reading styles. The most evident relation is found in non-professional reading since many pauses appear at strong boundaries and many strong boundaries are marked as pauses. In professional reading, this relationship is unidirectional only, since many strong boundaries are not signaled by the readers as a pause. Therefore, we can assume that professional readers use other acoustic correlates than silence to signal the structure of the message.

The differences found between the spontaneous and non–spontaneous styles might be explained by the complexity of the planning process involved in order to structure the information. In reading aloud, the readers use pauses to signal theme shift in the already given message, while in dialogs the speakers need to use pauses in order to be able to plan the content and the form of the message.

Even though the results presented in this paper indicate that there is a relationship between prosody, syntax and discourse structure for various speaking styles, an exhaustive description of the relations cannot be given. One of the reasons is, of course, that we only examined the pausing pattern as a correlate to prosodic phrase boundaries. In future, it would be necessary to take into account other prosodic features such as F0 variations, duration, segment lengthening, speech tempo, etc. Secondly, as the discourse boundary is defined in this study as a binary categorization between theme shift vs. continuation, the hierarchical annotation of the text is not captured. Therefore it would be interesting to include a hierarchical categorization of the discourse and map it to the hierarchical syntactic and prosodic structure. Lastly, since the results described here are based on small corpora, it would be desirable to extend the investigation to large data sets in order to be able to model the structure of the speech.

Acknowledgments

We would like to thank Petur Helgason, the Swedish Radio and our subjects for the data, as well as Rolf Carlson, Mattias Heldner and David House for discussions and valuable comments. Beita Megyesi was supported by CTT’s industrial partners, KTH, VINNOVA, and Vetenskapsrådet, and Sofia Gustafson-Čapková by the Swedish National Graduate School of Language Technology.

5. REFERENCES