Analysing fundamental frequency contours and local speech rate in map task dialogs

Hansjörg Mixdorff a,*, Hartmut R. Pfitzinger b

a Department of Computer Sciences and Media, TFH Berlin University of Applied Sciences, Luxemburger Str. 10, 13353 Berlin, Germany
b Department of Phonetics and Speech Communication, University of Munich, Schellingstr. 3, 80799 Munich, Germany

Received 1 September 2004; received in revised form 19 January 2005; accepted 19 February 2005

Abstract

The current paper reports the first results from the analysis of task-oriented dialogs using a Fujisaki model-based parameterization of F0 contours, as well as a model of the perceptual local speech rate. Two versions of map task style dialogs were examined: (1) the recordings made during the map task proper, (2) readings from scripts of the original dialogs by the same subjects. The first part of this paper presents an analysis of phrase boundaries with respect to form and function. A second issue is the problem of processing fillers, hesitations and repairs within the framework of the Fujisaki model-based analysis. The second part of the paper describes the comparative analysis of spontaneous and read versions of the same dialog fragments with respect to Fujisaki model parameters, contours of the perceptual local speech rate, and other features. In a perception test we asked listeners to identify the speaking style of dialog fragments. Apparently this was possible only for part of the data. Analysis of accent commands and perceptual local speech rate contours still suggested differences between the two speaking styles. The number of accented syllables, the associated accent commands’ amplitudes, and the perceptual local speech rate were generally higher in the read than in the spontaneous utterances. These results were almost significant despite the fact that the read version had been well re-enacted by the subjects and therefore did not exactly exhibit typical reading style characteristics. Despite this drawback, the methodology presented here has strong potential for further comparative prosodic studies of speaking styles.

Keywords: Fujisaki model; Perceptual local speech rate; F0 contours; Map task

1. Introduction

The study of prosody in spontaneous speech as produced e.g. in dialogs still poses a challenge in speech research. While prosodic regularities in
read speech have been investigated for many decades (Isačenko and Schädlich, 1964) the transition to analysing less formal speaking styles is still in progress. Concerning the vague notion of ‘speaking style’, Eskenazi (1993) suggested abandoning any description approach along a single axis named e.g. ‘formality’ or ‘casualness’, in favour of a multi-dimensional characterization scheme. Nevertheless, the distinction between read and spontaneous speech seems to be of pivotal relevance to the identification of features which constitute the speaking style. According to Blaauw (1995) and Hirschberg (2000), intonation and timing are generally among these features. Thus, any quantitative analysis of the effect of different speaking styles on prosody should at least consider intonation as well as timing.

Mixdorff and Jokisch (2001) developed an integrated model of German prosody (henceforth IGM) anchoring prosodic features such as F0, duration, and intensity to the syllable as a basic unit of speech rhythm. In the framework of IGM, following the works of Isačenko and Schädlich (1964) and Stock and Zacharias (1982), a given F0 contour is described as a sequence of linguistically motivated tone switches, major transitions of the F0 contour connected to accented syllables, or by the so-called boundary tones before prosodic boundaries. Tone switches can be thought of as the phonetic realization of phonologically distinct intonational elements, the so-called ‘intonemes’. In the original formulation by Stock, depending on their communicative function, three classes of intonemes are distinguished, namely the N$\uparrow$ intoneme (‘non-terminal intoneme’ signaling continuation, rising tone switch), I$\uparrow$ intoneme (‘information intoneme’ at declarative-final accents, falling tone switch), and the C$\uparrow$ intoneme (‘contact intoneme’ associated, for instance, with question-final accents, rising tone switch, establishing contact). Hence intonemes in the original sense mainly distinguish sentence modality, although there exists a variant of the I$\downarrow$ intoneme, I(E)$\downarrow$ which denotes emphatic accentuation and occurs in contrastive environments, for instance. Intonemes for reading style speech are predictable by applying a set of phonological rules to a string of text as to word accentability and the accent group formation. In order to quantify the interval and timing of the tone switches and boundary tones with respect to the syllabic grid, IGM adopts the well-known quantitative Fujisaki formula for parameterizing F0 contours (Fujisaki and Hirose, 1984). The Fujisaki model reproduces a given F0 contour by superimposing three components: A speaker-individual base frequency $F_b$, a phrase component and an accent component. The phrase component results from impulse responses to impulse-wise phrase commands associated with prosodic breaks. Phrase commands are described by their onset time $T_0$, magnitude $A_p$ and time constant $\alpha$. The accent component results from step-wise accent commands associated with accented syllables. Accent commands are described by on- and offset times $T_1$ and $T_2$, amplitude $A_a$ and time constant $\beta$.

In a perception study (Mixdorff and Fujisaki, 1995) employing synthetic stimuli of identical wording but varying F0 contours it was shown that non-terminal intonation was identified by tone switches to the mid-range of the subject and plateau-like continuation up to the phrase boundary, whereas questions required F0 transitions to span a total interval of more than 10 semitones. In the latter cases a rising tone switch on the last accented syllable is typically followed by another tone switch associated with the question-final rise. Two prototypical stimuli corresponding to N$\uparrow$ and C$\uparrow$ intonemes are displayed in Fig. 1. As can be seen, the C$\uparrow$ intoneme is followed by a high boundary tone modeled with a separate accent command when the accented syllable is not the last in the phrase.

When F0 contours can be extracted directly from the speech signal with reasonable error rates, timing information usually requires extensive phone, syllable, and word level segmentation of the speech signals. Furthermore, to achieve a meaningful timing contour similar to F0 contours, knowledge is necessary. E.g. in the case of Z-score-based duration contours (see e.g. Campbell, 2000) prototypical mean durations and standard deviations of all segments are essential. Pfitzinger (1998, 1999) developed a model for deriving perceptual local speech rate (PLSR) directly from the speech signal with sufficient accuracy. The
model is based on a linear combination of the local syllable rate and the local phone rate, since his investigations strongly suggest that neither the syllable rate nor the phone rate on its own represent perceptual speech rate sufficiently:

$$\text{PLSR} = 7.83 \cdot \text{Syllables/s} + 4.06 \cdot \text{Phones/s} + 1.41$$  (1)

A window of 625 ms duration is used to extract macro-prosodic rate information and ignore micro-prosodic or segmental duration variation. The size of the window is a result of perception experiments and avoids perceptual overshoot which occurs when using smaller windows, and the presence of rate changes within the window which occur more frequently when using longer windows. The model was developed on the basis of German read and spontaneous speech and yields a smooth contour of local speech rate values aligned with the speech signal (see e.g. PLSR contour in Fig. 2). A value of 100\% means that subjects asked to assess the corresponding stretch of speech would perceive a typical average speech rate. The predicted values roughly deviate 10\% from the perceived values which is small enough to be used for spoken language research.

In the current study we examine whether results from modeling intonemes in read speech also hold for more spontaneous speaking styles, that is, whether sentence modality (declarative, question, non-terminal) is signalled by similar prosodic configurations. Furthermore we address the general issue of whether the framework of IGM can be readily applied to non-reading speaking styles such as the style observed in task-oriented dialogs, for instance. In this context, the so-called ‘map tasks’ present a well-known paradigm and have been studied and documented for a large number of languages Brown et al. (1984): Two talkers are given slightly different maps. One of them—the so-called ‘giver’—receives a map with a path drawn on it and is requested to explain the route to the ‘follower’ who in turn will try to reproduce it as closely as possible on his/her own map. Since the two versions of the map do not feature exactly the same landmarks and may also vary with regard to the names of certain landmarks, the talkers have to interact verbally in order to solve the task. The map task can be performed with or without eye-contact. In her typology of spontaneous speech Beckman (1997) denominates the resulting speaking style instruction dialogues.

2. Speech materials and methods of analysis

In the experiment reported in this paper, talkers did not have eye-contact, but were sitting in the same room only separated by a screen. Recordings were made using head-worn microphones (Audio-Technica ATM73a) and stored on a mini-disc recorder. Later on the speech data were transferred to a computer at 16 kHz/16 bit. Subjects were 14 students of Media Computer Science at T FH Berlin in their last year, 5 females and 9 males. All subjects attended a class on Speech Communication and most of them were familiar with each other. Each of the subjects participated in two different tasks, once as a giver and once as a follower.
Four different pairs of maps originally created by Claßen (2000) were used in the experiment. All dialogs were first annotated on the word level and then segmented into moves, following the coding scheme developed at HCRC (Anderson et al., 1991). Punctuation marks were used to indicate phrase boundaries of turns and modality, with ‘...’ being used for non-terminal ones, and facilitated the following automatic generation of a dialog script. After the scripts had been created and revised, the same talkers read the sentences in the sequence in which they had produced them spontaneously. It must be noted in this context that certain disfluencies and repairs were removed from the scripts to make them more easily readable. From the resulting corpus of 28 dialogs (14 spontaneous and 14 read ones) of durations between 3 and 10 min, four were selected for the current study. Due to the nature of the task the givers generally speak more than the followers. Therefore utterances by four givers (three female, one male, henceforth \(f_{01}, f_{02}, f_{03}, \) and \(m_{01}\)) of a total duration of 10 min and 34 s were selected for detailed analysis. The main criterion for the selection was a close match of segmental content between spontaneous and read versions, followed by general agreement in the sentence modality produced. It will be discussed later that the latter requirement was often not met. Utterances with extensive repairs that had been deleted from the script were also excluded, whereas those in which hesitations and fillers occurred were not. The resulting subcorpus contains 122 dialog fragments of durations between 2 and 12 s.

F0 contours were calculated at 10 ms intervals using the Praat\(^1\) default pitch estimation. Contours were checked and corrected within the Praat PitchEditor. Then Fujisaki parameters were estimated automatically (Mixdorff, 2000) and if necessary corrected using the interactive FujiParaEditor.\(^2\) Time constants \(a\) and \(b\) were generally set to 2/s and 20/s, respectively. \(Fb\) values chosen for the four subjects are listed in the rightmost column of Table 1. Utterances were segmented on the phone and syllable levels by means of automatic forced alignment. Then, segmentations were manually corrected. Prosodic breaks were labeled with respect to cues involved (\(F0\) rises or falls, pauses, lengthening, and vocal fry) as well as with respect to the underlying sentence mode (statement-final, question-final, non-terminal) and discourse context (move-non-final, move-final).

<table>
<thead>
<tr>
<th>Subject</th>
<th>Speaking style</th>
<th>F0 mean</th>
<th>F0 s.d.</th>
<th>(Fb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(f_{01})</td>
<td>Read</td>
<td>247.2</td>
<td>45.6</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>Spontaneous</td>
<td>251.9</td>
<td>50.0</td>
<td>170</td>
</tr>
<tr>
<td>(f_{02})</td>
<td>Read</td>
<td>231.0</td>
<td>38.3</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>Spontaneous</td>
<td>254.9</td>
<td>42.7</td>
<td>180</td>
</tr>
<tr>
<td>(f_{03})</td>
<td>Read</td>
<td>232.6</td>
<td>61.1</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>Spontaneous</td>
<td>232.1</td>
<td>62.7</td>
<td>160</td>
</tr>
<tr>
<td>(m_{01})</td>
<td>Read</td>
<td>132.2</td>
<td>25.5</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>Spontaneous</td>
<td>128.5</td>
<td>25.5</td>
<td>95</td>
</tr>
</tbody>
</table>

Figs. 2–6 illustrate F0 analyses of some spontaneous utterances produced by subject \(f_{03}\). Each figure displays from top to bottom: The speech waveform, the \(F0\) contour (extracted: +/-signs, model-based: solid line), the contour of the perceptual local speech rate, and the amplitudes \(A_p\) and \(A_a\) of phrase and accent commands, respectively, which underline the \(F0\) contour.

In Fig. 2 we see two inter-pause stretches, the first one marked as incomplete by an \(F0\) rise to the upper limit of the subjects range and the second one, an elliptic question (“yours as well?”) marked by a similar rise to a slightly lower level. This is a typical example for the observation that the distinction between incompleteness and question is not marked by the span of the \(F0\) interval at the right hand boundary of the phrase. That is, if the ‘incomplete’ inter-pause stretch were listened to in isolation, it could as well be interpreted as an echo-question. This means that the distinction between question and non-terminal mode can only be made by drawing on the discourse con-

---

text, that is, the following utterances of the giver or
the reaction of the follower. Further examples of
very high F0 offsets at non-terminal boundaries
are displayed in Fig. 3 ("Blume ..."), and Fig. 4
("landest ..."). In contrast, Fig. 5 shows another
element of a question, which, however, is also
marked syntactically ("Haste da irgendwas an-
deres?"). There are nevertheless many instances
of non-terminal boundaries marked by rising to
a mid-level of F0, like, for instance, in Fig. 3
("oben rum ...", "Uhrzeigersinn ...") and Fig. 4
("runter ...", "links ..."), very much like the kind
of non-terminal boundaries found in read speech.
These occur often in the middle of an inter-pause
stretch, but also before pauses.
If one examines closely the contexts in which
mid-level (henceforth M) and high-level (hence-
forth H) offsets of F0 occur at non-terminal bound-
aries, a certain pattern emerges: Subject f03
typically uses mid-level boundaries for signaling
that the piece of information provided by her needs further clarification before the follower can complete the current transaction, that is, draw the current section of the path. Compare, for instance, Fig. 3: "und gehst dann oben rum im Uhrzeigersinn über die blühende Blume"—"and then you go round the top clockwise over the blooming flower"; and Fig. 4: "und dann geradeaus runter, so dass du links neben der goldenen Moschee landest..."—English translation: "and then straight down so that you arrive to the left of the golden mosque...".

Table 2

<table>
<thead>
<tr>
<th>Subject</th>
<th>Falling F0</th>
<th>Rising F0</th>
</tr>
</thead>
<tbody>
<tr>
<td>f01</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>f02</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>f03</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>m01</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

The discussion will now turn to some of the specific phenomena observed in spontaneous speech.
Fillers, hesitations and repairs, and how these can be taken into account when parameterizing F0 contours with the Fujisaki model. Fig. 2 shows the filler [E:] (SAMPA transcription) inserted into the sentence “Also der Startpunkt ist bei mir ... links oben ...”—“Well, my starting point is ... at the top left ...” As can be seen, the F0 contour during the filler follows the underlying phrase component. This kind of pattern can usually be observed when fillers occur within a phrase as in the example. In other locations, that is, between phrases, however, the F0 contour can sometimes be almost horizontal like in singing, indicating a phrase component close to zero. The dropping PLSR contour in Fig. 2 nicely reflects the slowing down that occurs between the two phrases.

In Fig. 5 we see an example of hesitation indicated by a short pause: “rechts ... dadrunter.”—“right ... below”. As it becomes clear, the phrase component continues across the pause, as well as the underlying accent command. During automatic parameter estimation, accent commands are not continued across pauses, so instances like these need to be corrected manually. Fig. 6 shows an example of repair with an almost completely flat F0 contour: “Ja dann sind das wahrscheinlich ... also da ist bei mir die goldene Moschee.”—“Well, then these are probably ... well, there's my golden mosque.” This example is also very conspicuous with respect to the relatively low underlying phrase and accent command amplitudes, $A_p$ and $A_a$. The fragment occurs at a point in the dialog where the talkers have realized that what is indicated as the 'golden mosque' on the giver's map is actually a second occurrence of 'blooming flowers' on the follower's. In the fragment, the giver basically restates this discovery without adding essentially new information. This corresponds to a CLARIFY move in the HCRC coding scheme. Furthermore, READY moves and ACKNOWLEDGMENT moves, typically consisting of single-word utterances such as ‘also’—‘well’, ‘okay’, and ‘ja’—‘yes’, ‘genau’—‘exactly’ are also associated with relatively small phrase commands (see head of utterance in Fig. 2).

4. Comparison of read and spontaneous speaking styles

4.1. Perception experiment

An identification experiment was conducted in order to examine whether the speaking style of the dialog fragments in the sub-corpus was classified correctly by native German listeners. The setting in which the recordings had been made were explained to them, as well as the fact that the read
385 utterances corresponded to the speech interactions that had spontaneously occurred during the map task. The utterances were randomized across subjects and speaking styles and played back to the listeners. They were presented a graphical user interface and had to perform a forced choice identification test with two possible answers: ‘spontaneous’ or ‘read’.

Although the number of listeners (six) was relatively small, certain tendencies could be observed. Listeners had considerable difficulties in identifying the correct speaking style. This could be due to the fact that the read utterances had wordings that arose from the map task interaction, and therefore did not correspond to typical reading style material. Furthermore, the subjects had returned into the mood of the original interaction as they re-enacted the map task from their scripts. Altogether, results differed depending on the particular subject as displayed in Table 3. The second column shows the total number of pairs of read/spontaneous utterances per subject. The third column contains the number of pairs for which each member had been correctly classified by over half of the listeners. The third column gives the percentage of votes ‘spontaneous’ across all stimuli pertaining to each subject, and therefore a measure of the degree of perceived ‘spontaneity’. The latter figures show that utterances by subject f03 were often mis-classified as ‘spontaneous’, whereas those by m01 were often falsely identified as read speech.

A first evaluation of the perception results indicated that fillers, such as [E:m] which exclusively occurred in the spontaneous utterances had facilitated the identification \( \left( \rho = 0.42, p < 0.01 \right) \) of the corresponding utterances. The duration of stimuli, however, did not have any significant effect with respect to correct identification.

### 4.2. Temporal characteristics

Despite the rather poor result of the perception test we conducted a quantitative comparison of the two sets of utterances in order to find significant differences, especially in cases where classification...
had been unanimous. For this purpose, a table was constructed in which the syllables of the two speaking styles were aligned. This facilitated the comparison of the temporal, as well as the intonational structure. Accent commands were associated with the nearest accented syllables or phrase-final syllables bearing high boundary tones. The aligned stretches of speech contain a net 1135 syllables for the read and 1106 for the spontaneous version, excluding hesitations and repair from the count, a reduction by only 2.6%, which only concerned unstressed function words as well as unstressed syllables in content words. The figure of actual phone elisions in the spontaneous version as compared to the read one was calculated as 4.3%. The spontaneous versions exhibit more pauses (230 against 212), but on the average these are not longer than those in the read versions (mean 0.40 s). We measured the syllable durations in the two versions, excluding fillers, and found a mean/standard deviation of 0.200 s/0.104 s in the spontaneous and 0.198/0.097 in the read versions, respectively. The figures were 0.260 s/0.116 s and 0.270 s/0.108 s for accented syllables only.

An analysis of the durations of corresponding phrases, that is, inter-pause stretches, yielded a mean/standard deviation of 1.882 s/1.273 s for the spontaneous, and 1.868 s/1.175 s for the read version. If we take the read version as a reference, the standard deviation of the duration of the corresponding spontaneous phrase increases with the duration of the read one, this means that there are more variations in the duration of spontaneous phrases as they grow longer.

4.3. Quantitative analysis of F0 contours

Table 1 lists means and standard deviations of raw F0 for all subjects and the two speaking styles. It is remarkable that there are only minor differences between spontaneous and read utterances. One could have expected a larger standard deviation in the ‘spontaneous’ versions, but only the female subjects have slightly higher values in these cases. A more conspicuous outcome is that subject f02 has a significantly higher mean F0 when speaking spontaneously, which had to be taken into account when Fb was chosen.

A statistical analysis of accents and boundary tones was conducted. In the spontaneous versions a total of 386 accent commands resulted from the analysis, 355 of which were associated with accented syllables and 31 with high phrase boundaries. The corresponding figures for the read version were a total of 406 accent commands, with 385 pertaining to accented syllables and 21 to boundary tones. Of the accented syllables 286 coincided in both versions, as well as 11 boundary tones. This indicates a considerable amount of difference in intonation between the read and spontaneous versions. Furthermore, the accent types, that is, the types of intonemes produced in the two versions do not always agree even when the accent locations do. Following the original definitions by Stock, we examined four types of intonemes altogether: (1) phrase-non-final N↑ intonemes, (2) phrase-final N↑↓ intonemes, (3) I↓ intonemes, and (4) C↓ intonemes. As explained in Section 1, the latter was associated with the last accented syllable of intonationally marked syntactic or functional questions. The distinction between types (1) and (2) was made in line with the results of an earlier study on read speech indicating significant differences between non-terminal accents occurring in phrase-non-final and phrase-final positions (Mixdorff and Jokisch, 2003).

As already indicated in Section 3.1, questions in the dialog often exhibit similar intonational marking as non-terminal phrase endings and can only be identified by their syntactic structure or the discourse context. Table 4 displays the mapping between intonemes realized in spontaneous (rows) and read (columns) utterances.

As can be seen there is a considerable number of rising N↑↓ intonemes replaced by falling I↓ intonemes.

<table>
<thead>
<tr>
<th></th>
<th>Phrase-non-final N↑</th>
<th>Phrase-final N↑↓</th>
<th>I↓</th>
<th>C↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase-non-final N↑</td>
<td>131</td>
<td>8</td>
<td>5</td>
<td>–</td>
</tr>
<tr>
<td>Phrase-final N↑↓</td>
<td>11</td>
<td>34</td>
<td>20</td>
<td>–</td>
</tr>
<tr>
<td>I↓</td>
<td>10</td>
<td>5</td>
<td>51</td>
<td>–</td>
</tr>
<tr>
<td>C↓</td>
<td>1</td>
<td>–</td>
<td>2</td>
<td>10</td>
</tr>
</tbody>
</table>
nemes and vice versa. This replacement occurs more frequently as we move from spontaneous to read condition. Examination of instances reveals that there are basically two reasons for this: Annotations which do not capture the correct sentence mode, as well as deviant productions despite correct annotations, the latter being the cause in over two-thirds of cases. This observation suggests that subjects found it inconvenient to produce a long sequence of utterances with a rising intonation under reading conditions, whereas in spontaneous mode the rises emerged naturally during the interaction. Furthermore, the annotation of non-terminality by ‘...’ might have irritated subjects as they read these text passages. As explained in Section 3.1, for subjects $f_01$, $f_02$, and $m_01$, the locations where the falling intonation in the read utterances occurred were usually at the end of a set of information leading to the next landmark, and this tendency became more prevalent in the read versions.

The accent command amplitudes $Aa$ were examined in cases where accent location as well as type of intoneme agreed. The correlation of $Aa$ measures between read and spontaneous material across these places was found to be 0.757. Since $Aa$ is a measure of the F0 interval spanned at an accented syllable, it is also a measure of the degree of prominence. Table 5 lists $Aa$ mean and standard deviation depending on the speaking style, the subject as well as the type of intoneme. The number of instances is listed in brackets only in the ‘read’ section, since it is the same for both conditions. The mean $Aa$ for read utterances is 0.42 as compared to 0.35 in spontaneous mode.

As can be seen, accent commands in read style were generally higher than in spontaneous mode, with the exception of subjects $f_01$ and $f_02$ in phrase-final N↑ intonemes. As expected, phrase-final non-terminal accents exhibit higher values of $Aa$ than phrase-non-final ones, since rising F0 is used for marking the end of a piece of information in the dialog and signals that more is to follow. $Aa$ for I↑ intonemes are generally close to the size of that for phrase-non-final N↑ intonemes which corresponds to the results of earlier studies. There are relatively few instances of matching C↑ intonemes which were all found with subjects $f_01$ and $f_03$. As expected they do not exhibit higher $Aa$ than those of phrase-final N↑ intonemes. This confirms that the formal distinction between non-terminal and contact intonemes cannot be maintained for the map task dialog. In the special case of subject $f_03$ discussed in Section 3.1 we examined the differences between transaction-final and transaction non-final phrase boundaries in the spontaneous version. We found an average $Aa$ of 1.12 for the former and 0.75 for the latter, suggesting that $f_03$ employed the range of the F0 interval to indicate the completion of a set of information.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Read speech</th>
<th>Spontaneous speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Phrase-final N↑</td>
<td>Phrase-final N↑</td>
</tr>
<tr>
<td>$f_01$</td>
<td>0.30/0.09 (39)</td>
<td>0.52/0.16 (10)</td>
</tr>
<tr>
<td>$f_02$</td>
<td>0.34/0.11 (22)</td>
<td>0.61/0.24 (3)</td>
</tr>
<tr>
<td>$f_03$</td>
<td>0.48/0.22 (42)</td>
<td>0.99/0.33 (16)</td>
</tr>
<tr>
<td>$m_01$</td>
<td>0.29/0.14 (28)</td>
<td>0.60/0.23 (5)</td>
</tr>
<tr>
<td>Total</td>
<td>0.36/0.18 (131)</td>
<td>0.76/0.34 (34)</td>
</tr>
</tbody>
</table>

Number of cases given in brackets.
Tables 6 and 7 show the results of two statistical analyses via General Linear Models for repeated measures. While the phrase command amplitudes $A_p$ are not significantly different with regard to speaking style, the accent command amplitudes $A_a$ are. Besides this global and conclusive result the data can only permit tentative detailed conclusions given in the small number of instances. Apparently what the data suggests is that subjects accent less frequently and assign less prominence during the spontaneous interaction than in the read utterances.

4.4. Quantitative analysis of perceptual local speech rate contours

The raw perceptual local speech rate (PLSR) values which were extracted at a step size of 10 ms from manually corrected syllable and phone boundaries of all utterances were submitted to two-way ANOVA using ‘subject’ and ‘speaking style’ as random factors. Although an analysis with fixed instead of random factors yielded extremely significant effects of both factors ($p < 0.001$) we used random factors in favour of the generalizability of the results. The mean PLSR values are shown in Fig. 7 and the outcome of the ANOVA is shown in Table 8. The subject has no significant influence on the mean PLSR while the speaking style is almost significant ($p = 0.085$). We decided to use the term “almost significant” and the symbol “(*)” for the 10% significant level because significant values which are only slightly above the 5% level indicate trends and tend to become significant on a 5% level if the sample size is extended from its original small size. The interaction between the subject and the speaking style is extremely significant ($p < 0.001$) indicating that different subjects use different speech rate strategies when changing the speaking style. Fig. 7 shows that subject f01 uses the same mean PLSR in read and spontaneous speech while the other three subjects have a lower mean PLSR in spontaneous speech. Their PLSR strategy is in accordance with the results of Hirschberg (2000) who consistently found a smaller number of syllables per second in spontaneous speech than in read speech.

The second statistical analysis is concerned with the amount of PLSR variation. For each utterance a standard deviation of the corresponding PLSR

Table 6
Statistical analysis of accent command amplitudes $A_a$ by means of a general linear model for repeated measures

<table>
<thead>
<tr>
<th></th>
<th>Degrees of freedom</th>
<th>$F$</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Style</td>
<td>1</td>
<td>4.705</td>
<td>0.030 *</td>
</tr>
<tr>
<td>Subject</td>
<td>3</td>
<td>7.255</td>
<td>0.000 ***</td>
</tr>
<tr>
<td>Style $\times$ Subject</td>
<td>3</td>
<td>0.706</td>
<td>0.548 n.s.</td>
</tr>
</tbody>
</table>

Table 7
Statistical analysis of phrase command amplitudes $A_p$ by means of a general linear model for repeated measures

<table>
<thead>
<tr>
<th></th>
<th>Degrees of freedom</th>
<th>$F$</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Style</td>
<td>1</td>
<td>0.832</td>
<td>0.362 n.s.</td>
</tr>
<tr>
<td>Subject</td>
<td>3</td>
<td>0.394</td>
<td>0.757 n.s.</td>
</tr>
<tr>
<td>Style $\times$ Subject</td>
<td>3</td>
<td>1.755</td>
<td>0.154 n.s.</td>
</tr>
</tbody>
</table>

Table 8
Statistical analysis of all perceptual local speech rate (PLSR) values by means of an ANOVA with two random factors (speaking style, subject)

<table>
<thead>
<tr>
<th></th>
<th>Degrees of freedom</th>
<th>$F$</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Style</td>
<td>1</td>
<td>6.381</td>
<td>0.085 (*)</td>
</tr>
<tr>
<td>Subject</td>
<td>3</td>
<td>4.503</td>
<td>0.124 n.s.</td>
</tr>
<tr>
<td>Style $\times$ Subject</td>
<td>3</td>
<td>47.324</td>
<td>0.000 ***</td>
</tr>
</tbody>
</table>
values was estimated as a representation of the amount of PLSR variation during an utterance. Next, General Linear Model statistical analysis with repeated measures was applied to test the influence of speaking style and subject on the standard deviations. The result is shown in Table 9: Only the speaking style has an almost significant influence on the amount of PLSR variation indicating that spontaneous speech shows a tendency towards greater PLSR variation.

Since means and standard deviations of the perceptual local speech rate were similar in the spontaneous and read versions of the data, we compared speech rate histograms for all four speakers in order to examine whether the shapes of the distributions differed. These histograms with occurrences displayed in percent are displayed in Fig. 9 in the top section of each of the panels. The bottom section of each panel shows the difference between the two distributions in percent. A positive value indicates a larger contribution of the respective interval in the read version, whereas a negative value shows a larger contribution in the spontaneous version. As can be seen the two distributions (top) show distinct differences which are clearly reflected by groups of neighboring intervals with either positive or negative values (bottom). The details of these patterns vary between speakers, but there is a tendency of intervals below the mean PLSR of a respective speaker to be found more often in the spontaneous versions than in the read ones, whereas intervals above the mean PLSR are more frequent in the read versions.

Compare, for instance, speaker f01 with means of about 90% for read and spontaneous versions. At the right edges of the distributions, very high values of PLSR are more frequent in the spontaneous versions. Subjects m01, f03, and f02 also show very low values of PLSR more often in the spontaneous mode. These findings suggest that although the read versions are on average faster, very slow as well as very fast chunks of speech are more typical of the spontaneous interaction in the map task dialog.

### 4.5. Read versions produced in isolation

In order to exclude the possibility that the strong similarities between read and spontaneous versions had been caused by the methodology of having subjects re-enact their original dialogs in pairs, we asked the four speakers examined in this study to produce an additional read version. The dialog fragments under investigation were split at sentence boundaries, randomized and displayed to the subjects by means of a slide show for reading in an isolated fashion. We thoroughly compared the resulting data with the corresponding read/re-enacted versions by repeated listening and partial quantitative analysis. As a result of our analysis we did not find any striking differences, because some of the utterances produced in isolation sounded more spontaneous than the re-enacted read versions and vice versa.

**Fig. 10** displays three versions of the utterance “und dann zwischen der Goldmine und der Villa Milano hoch”—“and then upwards between the goldmine and the Villa Milano”, produced by speaker f02, from top to bottom: (1) spontaneous, (2) read (re-enacted), and (3) read (produced in isolation).
Fig. 9. Distributions of PLSR values of the four subjects in read and spontaneous conditions (top), and resulting distribution differences (bottom). See text for discussion.
Fig. 10. Three versions of the utterance (subject f02): “und dann zwischen der Goldmine und der Villa Milano hoch”—English translation: “and then upwards between the goldmine and the Villa Milano”.

5. Discussion and conclusions

In this paper preliminary results from the quantitative prosodic analysis of map task dialogs have been presented. We are aware that the relatively small amount of data only permits tentative conclusions. Our findings suggest that prosodic categories emerging from the analysis of read speech which can be reproduced on isolated utterances cannot be applied to spontaneous speech without taking into account the pragmatics of the map task discourse. In particular, the functional distinction between non-terminal and contact intonemes as distinguishing sentence modality cannot be maintained for the map task dialogs, because F0 can reach approximately the same high values at the end of incomplete utterances as at the end of functional questions. If we interpret this result with respect to Stock’s original definition of N↑ intonemes and C↑ intonemes, the communicative functions of ‘non-terminality’ and ‘establishing contact’ of phrase-final accents collapse into one group, as they cease to be prosodically distinct—at least in the cases of subjects f01, f02, and m01. Only the phrase-non-final accents which exhibit lower accent command amplitudes correspond to the original class of non-terminal N↑ intonemes in read speech. The case might be different for subject f03 who uses very high transaction endings as compared to falling tone switches in the other subjects. A tentative conclusion that could be drawn from this observation is that high rises in the map task discourse actually fulfill two purposes: They signal incompleteness in the sense that more information will follow, and at the same time establish contact with the collocutor and receive his/her acknowledgement.

The results of our study also suggest that specific phenomena of spontaneous speech, such as fillers, hesitations, and repairs, can be modeled within the current framework. Subjects usually slow down and continue on the same phrase component line when they begin an utterance with a filler, for instance. Our perception experiment showed that speaking style identification was unanimous only for part of the data, indicating that the ‘read’ version had been well re-enacted by the subjects and therefore did not exactly exhibit typical reading style characteristics. In this context, an A/B-comparison rather than an identification test might have obtained a clearer separation.

The quantitative analysis of accent commands and perceptual local speech rate contours suggested relatively small differences between the two speaking styles. The number of accented syllables, the associated accent commands’ amplitudes, and the mean perceptual local speech rate, were generally higher in the read than in the spontane-
ous utterances. These results were found to be almost significant. It must be stated, however, that the differences between the two versions with respect to the accented words, phrasing, and prosodic realizations of phrase boundaries made comparison difficult. This is illustrated by the poor match of only 56% of accent command locations that coincided in the two versions.

In a continuum between read and spontaneous speaking styles, most of the read, re-enacted dialogs appeared to be closer to the ‘spontaneous’ side, if we take into account that their wordings were the same as in the original dialog, as well as the fact that the read versions were produced by subjects in an interaction based on the script. This proved to be a general drawback of the very specific utterances produced during the map task which predominantly take the form of instructions and orders.

As a consequence, prosodic features such as F0 and duration contours represent only a part of the properties of a speaking style. Specific syntax and lexical choices, as well as the presence of non-linguistic markers such as fillers are as characteristic. This is even more important when we take into account that listeners in our tests strongly drew on the latter cues for identifying the speaking style.

Despite the shortcomings of the data examined in this study we feel that the methodology presented is promising and should be applied to further comparative prosodic studies of speaking styles, as the parameters we use suitably describe the intonational and temporal structure of the discourse. Future activities will therefore be directed towards the comparison of a larger range of speaking styles.

References


868


