Acted and Spontaneous Conversational Prosody — Same or Different?

Petra Wagner[1,2], *Andreas Windmann*[1]

[1]Bielefeld University, Faculty of Linguistics and Literary Studies, Germany
[2]Center of Excellence for Cognitive Interaction Technology

petra.wagner@uni-bielefeld.de, andreas.windmann@uni-bielefeld.de

Abstract

Previous work has shown that read and spontaneous monologues differ prosodically both in production and perception. In this paper, we examine whether similar effects can be found between spontaneous and read, or rather acted, dialogues. It is possible that speakers can mimic conversational prosody very well. Alternatively, they might use prosodic resources more than the conversational situation actually requires (overacting). Another possibility is that in acted dialogues, prosody is actually used less as a communicative device, as there is no need to establish a common ground or to organize the floor between interlocutors. In our study, we examined spontaneous and read dialogues of equal verbal content. The task-oriented dialogues contained a communicative situation implicitly demanding for for a higher speaking rate (time pressure). Our results show that globally, speakers met this conversational demand of increased speaking rate both in the acted and in the spontaneous situation, although we find different global speaking rates between the spontaneous and the acted condition. Also, read speech exhibits a lower F0 minimum and, consequently, a larger F0 range than read speech, which may be explicable by a lack of active turn taking organization. Summing up, acted conversational prosody resembles many features of spontaneous interaction, but also shows systematic differences.

Index Terms: speaking styles, conversational prosody, speaking rate, read speech, spontaneous speech

1. Introduction

In recent years, the usage of read speech as the sole ground truth against which our models and theories should be tested has been called into question [1]. One main concern about the predominant usage of read speech lies in the circumstance that read speech has been shown to differ from spontaneous speech on various segmental and prosodic parameters. Some of these differences are easily explicable as the result of different modes of speech production, with read speech typically being less disfluent [2]. Based on these results, the usage of read speech in many of our experiments may actually be regarded as an advantage, as no "disfluency noise" caused by incremental speech planning interferes with the production result. Another obvious advantage of using read speech is that it gives us maximal control of our experimental variables. This point is strengthened even further by [2] in that listeners may at times be unable to differentiate between read and spontaneous utterances, if speakers are able to truthfully re-enact spontaneous dialogues with an appropriate "quasi-spontaneous" prosodic structure. They therefore argue that the prosody of read speech may in fact be an adequate model of spontaneously produced utterances, but only if great care is taken that the discourse context is modeled appropriately, as to enable readers to modify their speech according to the pragmatic needs.

However, the presence or absence of disfluencies appears to be not the sole prosodic difference between read and spontaneous speech: Spontaneous speech has been found to be have a higher articulation rate (syls/s), less fundamental frequency variability [3, 4, 5], and fewerer pitch accents [2] and more rising boundary tones [6, 7]. Contrary to most other studies, [2] found a lower rate in spontaneous speech, measured as *perceptual local speech rate*, a metric combining both phone and syllable rate [8]. Their study was conducted on map-task dialogues, while other research has focused on monologues. Most investigations found rather weak acoustic effects differentiating *clearly* between read and spontaneous speech productions, timing and speech rate being the best predictors of style.

Furthermore, [6, 7] provide evidence that de-accentuation of given referents occurs much more predictably in read compared to spontaneous speech, and no designated pitch accent types are used to systematically distinguish between new and accessible referents. [6] suggests that similarly to the more frequent occurrence of hesitations in spontaneous speech, some of the differences in information status marking may be explicable by different resources needed for speech planning across the two modes. However, she also suggests that due to communicative needs, contrastive accents may at times override the accent structure predicted by accessibility and novelty constraints, i.e. the pragmatic needs shaping prosody may not always be fully accessible in a reading task, e.g. due to montioring the listener's level of attention based on his or her feedback behaviour [9, 10].

This assumption is in line with an argument by [1], who suggested that due to its authenticity and real communicative needs, spontaneous interactions may actually lead to different communicative (and prosodic) behavior and reactions than read interactions.

In our corpus study, we take up this last point and examine whether the communicative behavior intended to actively change the prosodic behavior of a discourse partner leads to different behaviors and reactions depending on whether there is a spontaneous or re-enacted interaction where communicative needs need not be fulfilled to a similar degree.

2. Methods

In order to find out whether the communicative needs of spontaneous interaction lead to different prosodic reactions on the side of a discourse partner, we examined a corpus of German task-oriented face-to-face dialogue interactions. The recordings resemble a tourist information scenario, with one speaker being equipped with a set of information items about a fictitious popular tourist destination typically inquired after in a tourist information: accommodation, time tables for public transport, theater and concert programs, activities for children, hiking trails

and seasonal highlights (carnival, winter sports). The other interlocutor is a confederate, who systematically inquires after various possible recreational activities for her family holidays. After half of the information has been retrieved by the confederate, she claims to be in a hurry, as she has another appointment about to start, but continues to ask further questions. She expresses her time pressure by actively raising her f0, speaking faster and giving considerably more backchannel signals, throughout the interlocutors' utterances. As backchannels tend to be interpreted as a signal to continue speaking [11], an increased backchannel frequency may lead to a certain time pressure on speech productions. At no point in the interaction does the confederate openly ask the interlocutor to speak faster. The recordings made under time-pressure will henceforth be referred to as *fast*, independently of the speech rate actually produced, the recordings made during the first half of the conversation will be henceforth referred to as *slow*.

Several weeks after these recordings had been made and transcribed orthographically, the same speakers and the confederate were asked to repeat their conversations, this time reading their previous interactions based on the orthographic transcriptions. In order to simplify the reading task, disfluencies had been deleted and ungrammatical sentences had been repaired in the reading material. We will continue to refer to the data from the original interaction task and the subsequent reading session as *spontaneous* and *read* condition, respectively.

As it is possible that the prosodic realizations are influenced by the segmental and grammatical structure rather than the speaking condition (slow vs. fast), the order in which the inquiries were made by the confederate was balanced: in half of the conversations, the inquiries that were made in the beginning (slow condition), were made in the other half during the fast condition under time pressure and vice versa. That way, it was ensured that the text material was distributed more or less equally across the fast and the slow conditions.

All recordings were made in a sound-treated recording studio at the former Institute of Communication Sciences and Phonetics at the University of Bonn using high-quality studio equipment.

In total, 6 male and 6 female speakers were recorded in both recording conditions, resulting in 24 dialogues under study. For all dialogues, the information-givers' productions were transcribed and annotated manually using a narrow transcription. Furthermore, syllable and intonation phrase boundaries were annotated manually likewise. Most dialogues had a duration between 7 to 8 minutes, resulting in roughly 3 hours of analyzed speech.

3. Results

3.1. Speech Timing

We analyzed segmental durations from the information givers' speech in order to assess whether the experimental manipulations had an effect on overall speaking rate. We excluded segments from phrase-final syllables and segments with a duration of more than 500 ms, as these are likely to result from annotation errors or obvious disfluencies. The remaining segment durations were z-normalized within cells defined by the label of the segment itself, manner of articulation and voicing status of the preceding and following segment, and the speaker. This approach facilitated control of the well-known effects of phonological identity and environment on segment duration [12], and also allowed us to factor out between-speaker variation. Analysis was restricted to data from cells with at least 50 observations, so as to obtain stable mean and standard deviation estimates for the z-score normalization. After all exclusions, approximately 45000 observations were analyzed. The data were analyzed using linear mixed effects models as implemented in the *lme4* package [13] in R [14]. Analysis was conducted by first fitting a null model without any fixed effects but with by-speaker random intercepts and by-speaker random slopes for the variables textsccondition (spontaneous/read) and RATE (fast/slow). We assume that random effects for items are not necessary due to the wide variety of lexical items in the corpus. We then assessed whether the factors CONDITION (spontaneous/read) and RATE (fast/slow) or their interaction had any effect on segmental durations, by fitting models including them as fixed factors and comparing these models to the null model using likelihood ratio tests. The analysis yielded significant main effects of both factors (CONDITION: F(1)=28.50; RATE: F(1)=30.43) on z-normalized segment durations. All likelihood ratio tests yielded $p-$values< 0.0001. No evidence for an interaction was found.

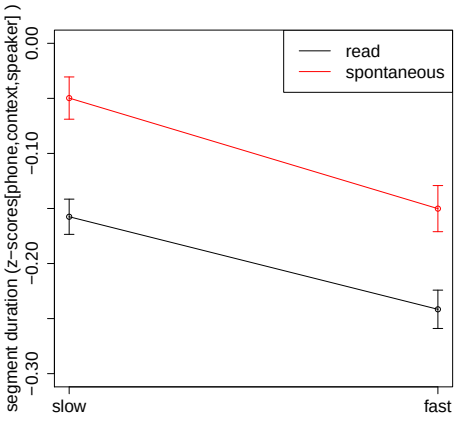

Figure 1: *Median segment durations (z-scores) by rate and condition. See text for details.*

Thus, information-givers were influenced in their speaking rate by the behavior of the confederate in the spontaneous conversation and, interestingly, also in the subsequent reading, talking faster when prompted by the confederate's indication of time pressure. Surprisingly, the read condition was also overall *faster* than the spontaneous condition, as is also evident from Figure 1. One caveat is that prosodic prominence was not controlled, and effects were on the whole also rather subtle, amounting to a few milliseconds on average in absolute terms.

3.2. Intonation

F0 contours were extracted using Praat's [15] autocorrelation algorithm with the default settings, i.e., a floor of 75 Hz and a ceiling of 600 Hz. After inspection of the data, we decided to exclude observations lying outside 1.5 times the interquartile range of the respective speaker as likely tracking errors. The remaining contours were smoothed using a three-point moving average filter and converted to semitones according to the following formula:

$$st = 12 * log_2 \frac{F_0}{F_{0_base}} \quad (1)$$

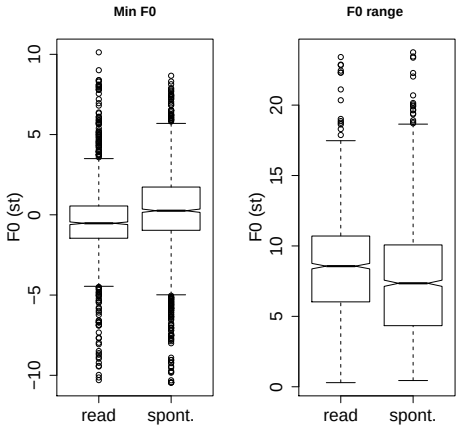

Figure 2: *F0 minimum and range (semitones) by condition. See text for details.*

In accordance with [16], $F_{0_{base}}$ was defined as the 5th percentile of all F_0 measurements for one speaker.

We extracted F0 mean, maximum, minimum and range (defined as the minimum subtracted from the maximum) for the individual phrases in the data, restricting analysis to phrases with at least ten measurements. These dependent variables were entered into linear mixed effects models. A similar analysis protocol was followed as in the timing analysis: for each of the four variables, we set up a separate null model with by-speaker random intercepts and by-speaker random slopes for the variables textsccondition (spontaneous/read) and RATE (fast/slow) as well as for gender effects. For each of the dependent variables, we then built models including each of the two experimental variables as a fixed effect and compared these models to the null models using likelihood ratio tests, in order to determine whether the experimental variables contributed significantly to the overall variance. We found significant effects only for two combinations of experimental and dependent variables: speakers had a lower F0 minimum (F(1)=12.51), and – as a consequence – also a larger F0 range (F(1)=15.98) in the read than in the spontaneous condition (see Figure 2). The likelihood ratio tests against the respective null models yielded p-values < 0.003125 in either case (Bonferroni correction; $\alpha = 0.05/16$) , hence the effects appear to be robust even though they are rather subtle in absolute terms.

Inspection of the data suggested a tentative interpretation of the difference in F0 minimum, and, hence, range. In the spontaneous conversations, the information givers' task was to present the confederate with selections of items for different categories, such as hotels or leisure activities. This frequently resulted in a kind of list intonation with final rises, as shown in Figure 3 for a phrase-final sequence from a spontaneous conversation. Pragmatically, these final rises may function as turn holding cues, signaling that further information items will be presented. In the read condition, this communicative strategy was not necessary because the structure of the turn succession was clear from the printed transcripts of the dialogues. In this situation, speakers mostly produced F0 contours with final lowering towards the ends of phrases. This can be seen in Figure 4, which displays the read counterpart by the same speaker of the utterance in Figure 3. In many cases, the final lowering will have marked the lowest point in the F0 contour of a phrase, leading to lower F0 minima in the read than in the spontaneous condition.

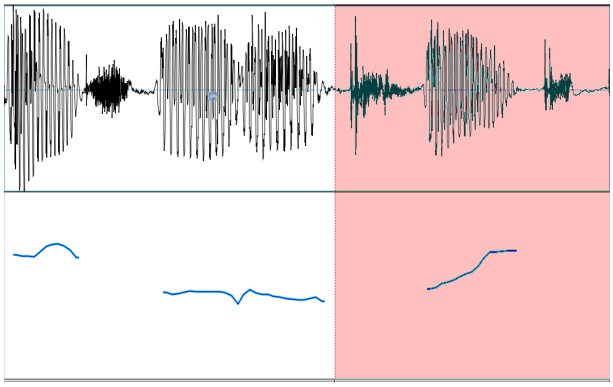

Figure 3: *Waveform and F0 trajectory of the phrase-final sequence "Hundsmühler Krug" (a fictitious restaurant name) produced by male speaker ada in the spontaneous condition. The marked section comprises the final syllable, /kRu:k/.*

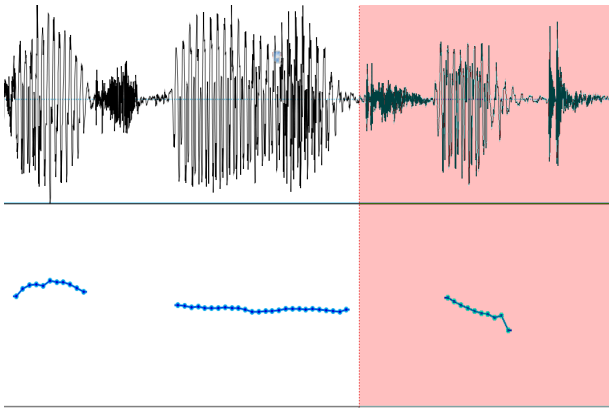

Figure 4: *Waveform and F0 trajectory of the phrase-final sequence "Hundsmühler Krug" (a fictitious restaurant name) produced by male speaker ada in the read condition. The marked section comprises the final syllable, /kRu:k/.*

Preliminary evidence for this interpretation was found in a re-analysis of the F0 contours from phrase-final syllables in the data. Data processing and statistical analysis was carried out as above, assessing effects of RATE and CONDITION on F0 mean, maximum, minimum, and, this time slope of the phrase-final F0 contours. Slopes were computed by fitting ordinary linear regression lines to the F0 trajectories. We found significant main effects of CONDITION on F0 mean (F(1)=16.97) and maximum (F(1)=18.70; p-values from likelihood ratio tests against null models < 0.0001 for both F0 mean and maximum), both of which were slightly higher in the spontaneous than in the read condition (cf. Figure 5). Effects of CONDITION at p-values < 0.05 were found for F0 minimum and slope as well, but these may not reliable enough for interpretation due to the multiple comparisons problem. The overall picture tentatively suggest that speakers indeed had a disposition towards attaining higher F0 targets phrase-finally in the spontaneous compared to the read condition. The failure to find a more robust effect for F0 slope in particular may be due to cases where the actual rise

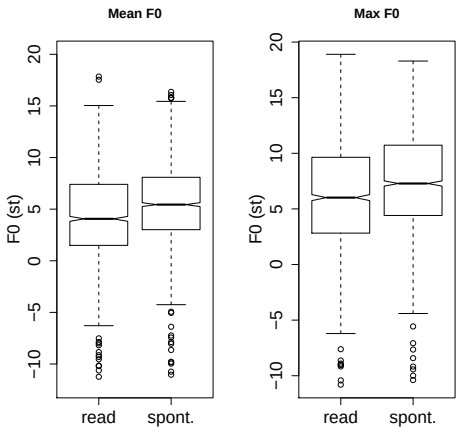

Figure 5: *Phrase-final F0 mean and maximum (semitones) by condition. See text for details.*

or fall happened prior to the phrase-final syllable.

4. Discussion

Our investigation has replicated previous comparisons of read and spontaneous speech insofar as some small but systematic differences were discovered. However, some of the results did not show the direction expected from former work: Rather unexpectedly, our spontaneous speech productions turned out to be systematically *slower* than the comparable read speech productions. Similar results were reported by [2] and on map task dialogues, while most other comparative studies have been concerned with monologues. This may lead us to conclude that the one-dimensional stylistic distinction between read and spontaneous speech can be misleading unless the complex communicative setting (here: dialogue vs. monologue) is also taken into account. There are two possible reasons for the effect observed here: on obvious explanation is the presence of hesitations and disfluency phenomena in spontaneous as opposed to read speech (cf. [17]). Another reason may be the presence of shorter intonation phrases in the spontaneous condition, resulting in an increased frequency of final lengthening phenomena, although we tried to control for both disfluencies and phrase-final lengthening to some degree by excluding phrase final and overly long syllables from our analysis. However, as final lengthening may be affecting more than the ultimate syllable in a phrase [18], this control may not have been sufficient. But even in this light, the simple rules of $spontaneous = fast$ and $read = slow$ appears to be overly simplistic. Also, even if shorter intonation phrases are to some extent responsible for this finding, it remains unclear why this pattern is exclusive to dialogues, as spontaneous productions in monologues are also subject to speech planning constraints.

While we did find a style-specific tempo effect, we also provided evidence that readers were indeed able to reproduce some of the prosodic adaptations due to communicative needs quite well, by increasing their speech tempo under the acted impression of time pressure. This finding is in line with the argument by [2], that re-enacted speech can be quite authentic if the pragmatic context is made fully explicit.

Another stable difference were the lower pitch floor and smaller pitch range in the read condition. As discussed above, one explanation for this may be the comparatively more frequent production of high boundary tones in spontaneous speech, and the lack of low boundary tones often representing the f0 minimum within an utterance. An obvious pragmatic reason for this is its usage as a turn holding device, which is simply not necessary to use in the reading condition, as speaker changes are entirely predictable by the orthography and there is not need for floor management in the ongoing discourse. This explanation would show that even if sufficient pragmatic context cues are available to the speakers (here: speaker changes), they fail to adapt their prosody accordingly, by failing to use the turn holding strategies of their corresponding spontaneous productions. However, it should be kept in mind that previous research also found an increased frequency of high boundary tones in spontaneous monologues (cf. Introduction), so the reason for these may be independent of floor management, and rather a consequence of ongoing speech planning processes. In either case, the speakers failed to indicate these production related of floor management related usages of prosodic marking. Thus, some of the richness of prosodic functions was not adequately reproduced by the speakers, despite their having access to as many contextual cues as can be possibly provided in a reading task. We suspect that this systematic difference between read and acted dialogues can be at least partly explained by the lack of necessity to indicate floor management or ongpoing production planning in read interactions.

For future work, it would be interesting to investigate the shape of the boundary tones further, as recent results indicate a systematic difference between floor holding and other high boundary tones [19].

It is interesting that *some* communicative adaptations of prosody, i.e. the change of speaking rate under simulated time pressure, remained in the acted condition, while others could not be replicated, i.e. prosodic adaptations probably related to floor management. Perhaps, the verbally expressed time pressure demands were obvious enough to lead to a prosodic reaction in the acted condition, while the not openly stated turn taking demands were too subtle to transfer to the prosody in acted interactions.

5. Conclusions

Our work reproduces previous results on subtle but stable differences between spontaneous and read speech, concentrating on dialogues. We could show that spontaneous speech should not be treated as a uniform speaking style, as it can be either slower or faster than read speech, depending on the task. We also found evidence that *some* pragmatically induced prosodic adaptations were indeed transferred from spontaneous to read interactions, while others were not. We therefore conclude that some but not all pragmatic uses of prosody can be reliably reproduced and studied in read interactions. Despite its obvious advantages, investigations on the pragmatic functions of prosody should therefore never be entirely restricted to read speech.

6. Acknowledgements

The authors would like to thank Julia Abresch and Felicitas Haas for planning and carrying out the recordings, transcriptions and annotations at the former Institute of Communiciation Sciences and Phonetics (IKP) at the University of Bonn.

7. References

[1] P. Wagner, J. Trouvain, and F. Zimmerer, "In defense of stylistic diversity in speech research," *Journal of Phonetics*, vol. 48, pp. 1–12, 2015.

[2] H.-J. Mixdorff and H. Pfitzinger, "Analysing fundamental frequency contours and local speech rate in map task dialogs," *Speech Communication*, vol. 46, pp. 310–325, 2005.

[3] V. Dellwo, A. Leeman, and M.-J. Kolly, "The recognition of read and spontaneous speech in local vernacular: The case of zurich german," *Journal of Phonetics*, vol. 48, pp. 13–28, 2015.

[4] G. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," *Speech Communication*, vol. 22, pp. 43–65, 1997.

[5] M. Eskénazi, "Changing speech styles: Strategies in read speech and casual and careful spontaneous speech," in *Second International Conference on Spoken Language Processing*, Banff, Alberta, Canada, 1992.

[6] L. de Ruiter, "Information status marking in spontaneous vs. read speech in story-telling tasks - evidence from intonation analysis using GToBI," *Journal of Phonetics*, vol. 48, pp. 29–44, 2015.

[7] S. Baumann and A. Riester, "Referential and lexical givenness: Semantic, prosodic and cognitive aspects," *Lingua*, vol. 136, pp. 16–37, 2013.

[8] H. Tillmann and H. Pfitzinger, "Local speech rate: Relationships between articulation and speech acoustics," in *Proceedings of ICPhS 2003*, vol. 3, Barcelona, Spain, 2003, pp. 3177–3180.

[9] H. Buschmeier, Z. Malisz, M. Wlodarczak, S. Kopp, and P. Wagner, "'are you sure you're paying attention?' – 'uh-huh'. communicating understanding as a marker of attentiveness," ser. Proceedings of INTERSPEECH 2011. International Speech Communication Association, 2011, pp. 2057–2060.

[10] Z. Malisz, M. Wlodarczak, H. Buschmeier, S. Kopp, and P. Wagner, "Prosodic characteristics of feedback expressions in distracted and non-distracted listeners," ser. Proceedings of The Listening Talker. An Interdisciplinary Workshop on Natural and Synthetic Modification of Speech in Response to Listening Conditions, 2012, pp. 36–39.

[11] V. Yngve, "On getting a word in edgewise," in *Proceedings of the Sixth regional Meeting of the Chicago Linguistic Society*, Chicago, IL, 1970.

[12] J. P. Van Santen, "Contextual effects on vowel duration," *Speech communication*, vol. 11, no. 6, pp. 513–546, 1992.

[13] D. Bates, M. Maechler, B. Bolker, and S. Walker, "lme4: Linear mixed-effects models using eigen and s4," *R package version*, vol. 1, no. 4, 2013.

[14] R. C. Team, "R language definition," 2000.

[15] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," 2001.

[16] J. Yuan and M. Liberman, "F0 declination in english and mandarin broadcast news speech." in *INTERSPEECH*. Citeseer, 2010, pp. 134–137.

[17] D. O'Shaughnessy, "Timing patterns in fluent and disfluent spontaneous speech," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 600–603.

[18] A. Turk and S. Shattuck-Hufnagel, "Multiple targets of phrase-final lengthening in american english words," *Journal of Phonetics*, vol. 35, no. 4, pp. 445–472, 2007.

[19] O. Niebuhr, K. Grs, and E. Graupe, "Speech reduction, intensity, and f0 shape are cues to turn-taking," in *Proceedings of SIGDIAL 2013 Conference*, Metz, France, 2013, p. 261269.