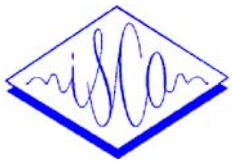

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Quantitative Analysis of Prosody in Task-Oriented Dialogs

Hansjörg Mixdorff

Faculty of Computer Science
TFH Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de

Abstract

The current paper reports first results from the analysis of task-oriented dialogs using a Fujisaki model based parameterization of *F0* contours. 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 dialog by the same speakers. In the scope of this paper an analysis of phrase boundaries with respect to form and function is presented. Results indicate, inter alia, that *F0* cues differ considerably from what has been observed in earlier studies on read speech. In particular, the strict functional distinction between non-terminal and contact intoneme which has been established through listening experiments cannot be maintained for the map task dialogs. Nevertheless speakers in the dialog make consistent use of *F0* cues associated with non-terminal and contact intonemes in read speech. A second issue touched on briefly in this paper is the problem of processing fillers, hesitations and repairs within in the framework of the Fujisaki model based analysis.

10.21437/SpeechProsody.2004-66

1. Introduction

Earlier work by the author and his co-workers was dedicated to an integrated model of German prosody [1] (henceforth IGM) anchoring prosodic features such as *F0*, duration and intensity to the syllable as a basic unit of rhythm. In the framework of IGM, following the works by Isačenko & Schädlich [2] and Stock & Zacharias [3], 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 so-called *boundary tones* before prosodic boundaries. Tone switches can be thought of the phonetic realization of phonologically distinct intonational elements, so-called 'intonemes'. In the original formulation by Stock, depending on their communicative function, three classes of intonemes are distinguished, namely the N↑intoneme ('non-terminal intoneme' signaling continuation, rising tone switch), I↓ intoneme ('information intoneme' at declarative-final accents, falling tone switch), and the C↑ 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↓ intoneme, I(E)↓ 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 accent group forming.

In order to quantify the interval and timing of the tone switches and boundary tones with respect to the syllabic grid, IGM employs the well-known quantitative Fujisaki formula for parameterizing the natural *F0* contours [4]. In an early perception study [5] 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 speaker 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.

In the current study we examine whether these results also hold for more spontaneous speaking styles, that is, whether sentence modality (declarative, question, non-terminal) is signaled by similar prosodic configurations as in read speech. 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 [6]: Two talkers are given slightly different maps. One of them - the so-called 'giver' - receives a map with a route drawn on it and is requested to explain this 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.

2. Speech Material and Method 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 was transferred to a computer at 16 kHz/16 bit. Subjects were 14 students of Media Computer Science at TFH Berlin in their last year, 9 males and 5 females. 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 [7] 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 [8]. Punctuation marks were used to indicate boundaries of turns and their associated sentence modality, 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 'readable'. From the resulting corpus of 28 dialogs (14 'spontaneous' and 14

read ones) of durations between 3 and 10 minutes, four were selected for the current study. Due to the nature of the task the givers generally talk more than the followers. Therefore utterances by the four givers (one male, three female, henceforth m01, f01, f02 and f03) of a total duration of 10 minutes and 34 seconds 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 agreement in the sentence modality produced. Utterances with repairs that had been deleted from the script were also excluded, whereas those in which hesitations and fillers (hmm…) occurred were not. The resulting sub-corpus contains 122 dialog fragments of durations between 2 and 12 seconds. *F0* contours were calculated at a step of 10 ms using the *Praat* (© P. Boersma [9]) pitch estimation. Contours were checked and corrected within the *Praat PitchEditor*. Then Fujisaki parameters were estimated automatically [10] and if necessary corrected in the interactive *FujiParaEditor* [11]. Time constants α and β were generally set to 2/s and 20/s, respectively. *Fb* values chosen for the four speakers are listed in the rightmost column of Table 1.

Table 1: *F0 stats and Fb for the four speakers in Hz.*

subject	condition	*F0* mean	*F0* s.d.	*Fb*
m01	read	132.2	25.5	95
	spontaneous	128.5	25.5	95
f01	read	247.2	45.6	170
	spontaneous	251.9	50.0	170
f02	read	231.0	38.3	160
	spontaneous	254.9	42.7	180
f03	read	232.6	61.1	160
	spontaneous	232.1	62.7	160

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-final, move-medial).

3. Results of Analysis

Table 1 lists mean and standard deviation of raw *F0* for the two speaking conditions. Interestingly there are only minor differences between spontaneous and read versions. 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* in the spontaneous version which had to be taken into account when *Fb* was chosen.

Figure 1 and Figure 2 show results of analysis from the spontaneous version by speaker f03. Each panel displays from top to bottom: The speech waveform, the *F0* contour (extracted: +-signs, model-based: solid line) and the underlying phrase and accent commands. In Figure 1 (top) we see two inter-pause stretches, the first one marked as incomplete by an *F0* rise to the upper limit of the speakers 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 taken for an echo-question. This means that the distinction between question and non-terminal mode can only be made by drawing on the discourse context, that is, the following utterances of the giver or the reaction of the follower. There are further examples of very high *F0* offsets at non-terminal boundaries in Figure 1, center ("Blume…"), and bottom ("landest…"). In contrast, Figure 2, top, shows another example of a question, which, however, is also marked syntactically ("haste da irgendwas anderes?"). There are nevertheless many instances of non-terminal boundaries marked by rises to a mid-level of *F0*, like, for instance, in Figure 1, center, ("oben rum…", "Uhrzeigersinn…") and bottom ("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 (henceforth ^{H}) offsets of *F0* occur at non-terminal boundaries, a certain pattern emerges: Mid-level boundaries typically signal that the piece of information provided by the giver needs further clarification before the follower can make the next move. Compare, for instance, Figure 1, center: "und gehst dann oben rum ^{M} im Uhrzeigersinn ^{M} über die blühende Blume ^{H}" – "and then you go round the top ^{M} clock-wise ^{M} over the blooming flower ^{H}"; and Figure 1, bottom: "und dann geradeaus runter ^{M}, so dass du links ^{M} neben der goldenen Moschee landest^{H}" - "and then straight down ^{M} so that you arrive to the left^{M} of the golden mosque^{H}". If one interprets this pattern from a communicative view, the two types of non-terminal boundaries have different functions: mid-level offsets of *F0* signal incompleteness with respect to the current move, whereas high offsets indicate that the giver is going to continue with a piece of information for the next move. In the framework of Stock's theory of German intonation [3], the former type would therefore rather correspond to the N↑intoneme whereas the latter could be regarded as a C↑intoneme which is employed for (re-)establishing contact.

Needless to say that the difference between the two types of boundary markers is also reflected by the amplitudes of the underlying accent commands as shown in Table 2.

Table 2: *Mean and standard deviation of accent command amplitude Aa associated with N↑ intonemes and C↑ intonemes.*

subject	*Aa*, ***N***↑ (μ/σ)	*Aa*, ***C***↑ (μ/σ)
m01	0.49 / 0.17	0.88 / 0.06
f01	0.41 / 0.11	0.77 / 0.19
f02	0.45 / 0.17	0.68 / 0.18
f03	0.66 / 0.13	1.09 / 0.23

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. Figure 1, top, 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. In Figure 2, top, we see an example of hesitation indicated by a short pause: "rechts… dadrunter."-

"right… below." As 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 [10]. Figure 2, bottom, shows an example of repair which is almost unnoticeable from the *F0* contour: "also 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. 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

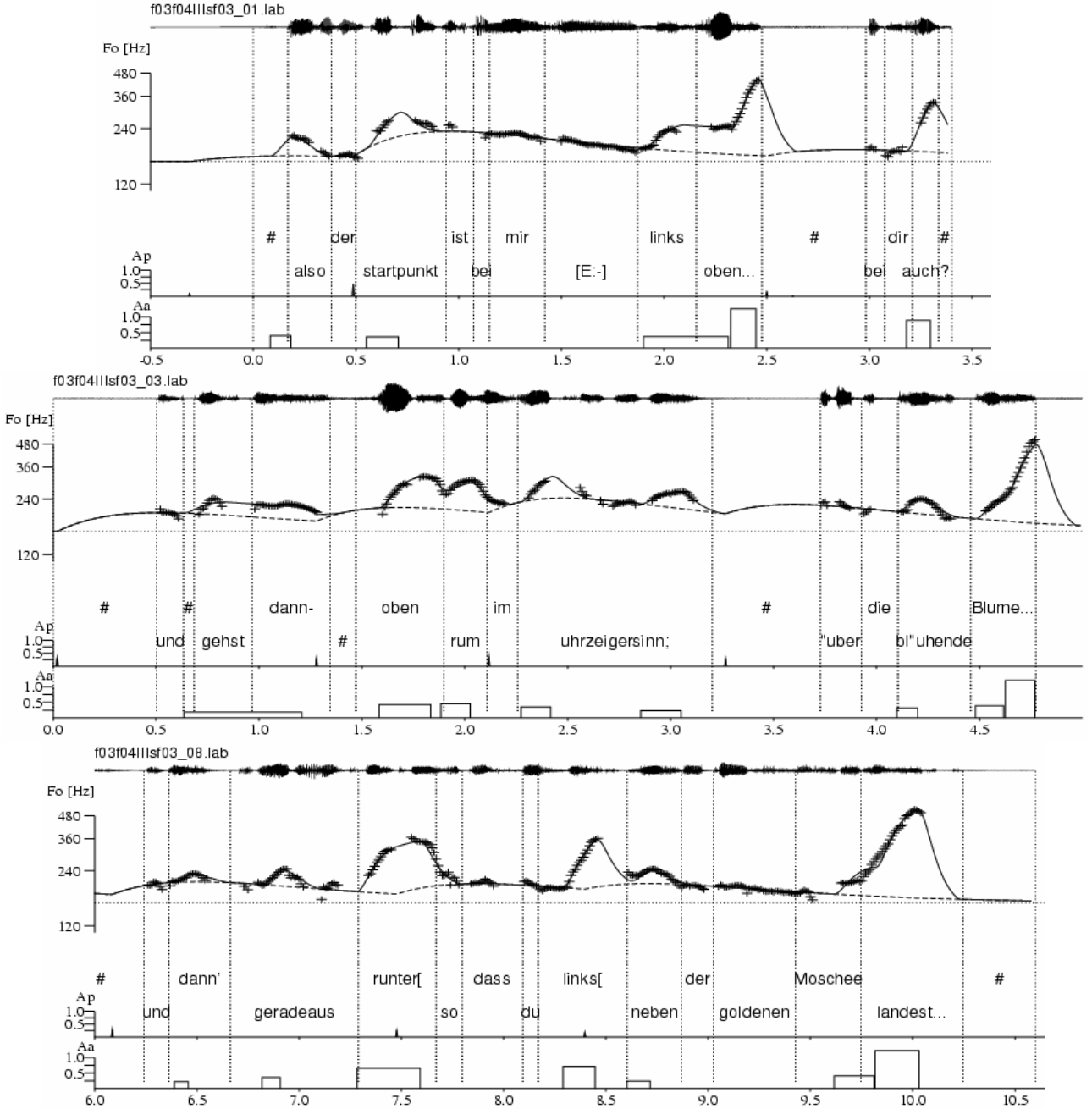

Figure 1: *Results of analysis, utterances by speaker f03, 'spontaneous' condition. Each panel displays from top to bottom: The speech waveform, the F0 contour (extracted: +-signs, model-based: solid line) and the underlying phrase and accent commands. Texts of utterances and English translations: (top) "Also der Startpunkt ist bei mir links oben… Bei dir auch?"-"Well, my starting point is at the top left… Yours as well?"; (center) "und gehst dann obenrum im Uhrzeigersinn über die blühende Blume…"- "and then (you) go round the top, clock-wise over the blooming flower…"; (bottom) "und dann geradeaus runter, so dass du links neben der goldenen Moschee landest…" - "and then straight down so that you arrive to the left of the golden mosque…". See text for discussion.*

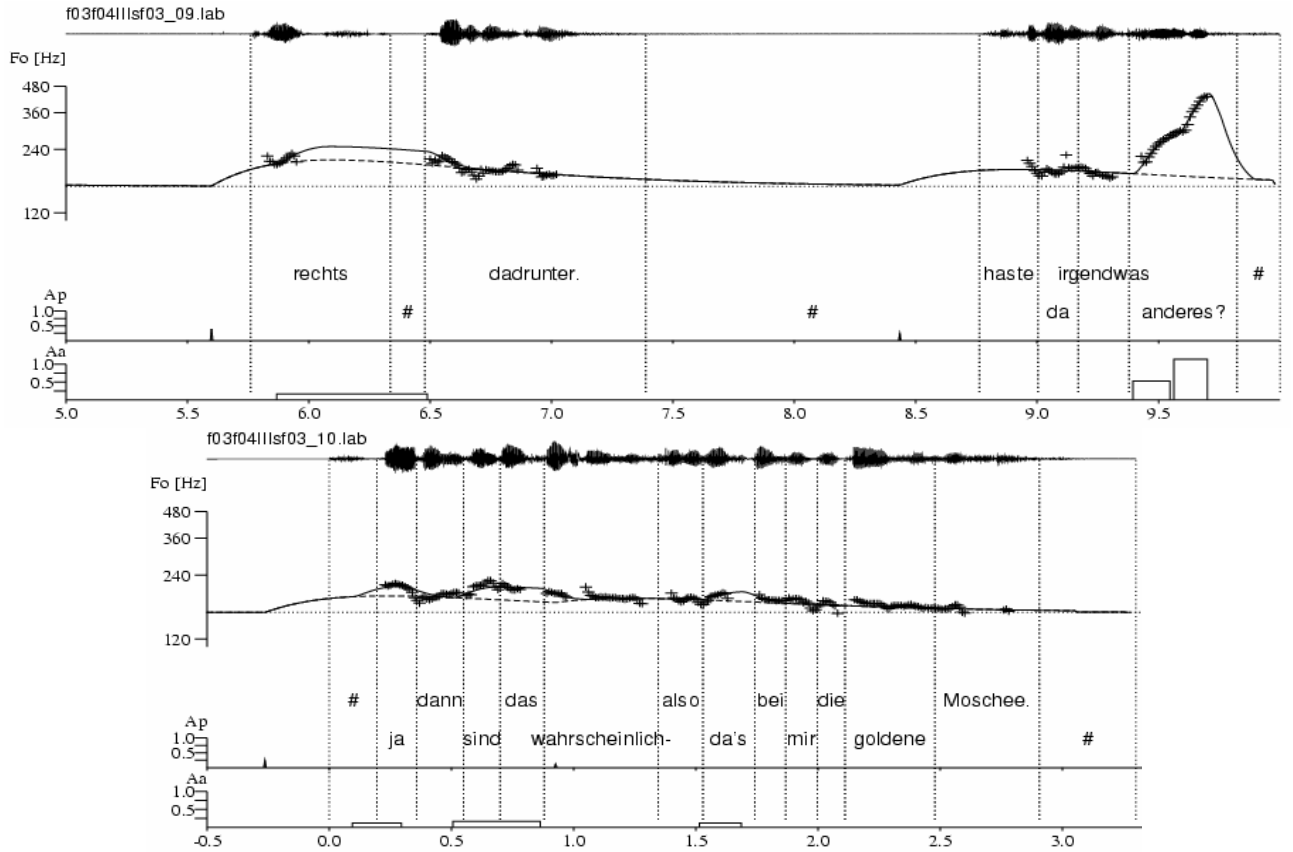

Figure 2: *Results of analysis, utterances by speaker f03, 'spontaneous' condition. Texts of utterances and English translations: (top) "rechts… dadrunter. Haste da irgendwas anderes?" " - "right… below. Do you have anything else there?"; (bottom)"also dann sind das wahrscheinlich… also da ist bei mir die goldene Moschee." - "Well, then these are probably…well, there's my golden mosque."*

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 Figure 1, top).

4. Discussion and Conclusions

In the current paper first results from the quantitative analysis of map task dialogs have been presented. As was shown for the prosodic marking of phrase boundaries, categories emerging from the analysis of read speech and reproducible in listening tests on isolated utterances cannot be directly applied to spontaneous speech without taking into account the pragmatics of the discourse. Still, Stock's N↑ intonemes and C↑ intonemes can be identified, though with more complex communicative functions than in read speech.

Preliminary results suggest that specific phenomena of spontaneous speech, such as fillers, hesitations and repairs can be modelled within the current framework. Future work will be dedicated to the detailed comparison of read and spontaneous versions of the map task and rhythmic analysis.

5. References

[1] Mixdorff, H. and O. Jokisch (2001): Building An Integrated Prosodic Model of German. In *Proceedings of Eurospeech 2001*, vol. 2, pp. 947-950, Aalborg, Denmark.

[2] A.V. Isačenko and H.J. Schädlich, 1964. Untersuchungen über die deutsche Satzintonation. Akademie-Verlag, Berlin.

[3] E. Stock and C. Zacharias, 1982. *Deutsche Satzintonation*. VEB Verlag Enzyklopädie, Leipzig.

[4] Fujisaki, H.; Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences *of Japanese," Journal of the Acoustical Society of Japan (E), 5(4),* 233-241.

[5] Mixdorff, H., Fujisaki, H. 1995: Production and Perception of Statement, Question and Non-Terminal Intonation in German. In: *Proceedings of the ICPhS '95*, Stockholm, Schweden, vol. 2, pages 410-413.

[6] Brown, G. Anderson, A.H., Yule, G., & Shillcock, R. 1984. *Teaching Talk*. Cambridge, England: Cambridge University Press

[7] Kathrin Claßen, 2000. Map Task – Eine Version für das Deutsche. AIMS, vol. 6 (4), pp. 65-83.

[8] Anderson, A.H., M. Bader et al., 1991. The HCRC Map Task Corpus. *Language and Speech 34*(4), pp. 351-366.

[9] http://www.praat.org

[10] Mixdorff, H., 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. *Proceedings ICASSP 2000*, vol. 3, Istanbul, Turkey, 1281-1284.

[11] http://www.tfh-berlin.de/~mixdorff/fujisaki_analysis.htm