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.

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↓intonone ('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 contour's 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
The outcomes of the study are as follows: subject f02 has a significantly higher mean value than subjects m01, f01, f02, and f03. In these cases, a more conspicuous 'spontaneous' version of the speech was observed, but only the female subjects exhibited this behavior. One could have expected a larger standard deviation in the data, but the minor differences between spontaneous and read versions were not significant. The two speaking conditions were compared, and there were only minor differences found.

In the discussion, it was noted that the results are consistent with previous studies. For example, in Figure 1, center, there are a few examples of a question, which, however, is also marked syntactically ("haste da irgendwas anders?"). There are also occasional instances of non-terminal boundaries marked by rises to a mid-level of F0, like, for instance, in Figure 1, center, ("oben rum..."), and bottom ("runter...", "links...").

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. The results are presented in Figure 1, top, and Table 2. The data show that F0 contours are influenced by 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 anders?"). There are also occasional instances of non-terminal boundaries marked by rises to a mid-level of F0, like, for instance, in Figure 1, center, ("oben rum..."), and bottom ("runter...", "links...").

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. The results are presented in Figure 1, top, and Table 2. The data show that F0 contours are influenced by 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 anders?").
“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](image)

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 dialogues 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