ISCA Archive
http://www.isca-speech.org/archive

ITRW on Prosody in
Speech Recognition and Understanding
Molly Pitcher Inn, Red Bank, NJ, USA
October 22-24, 2001

USING PROSODIC FEATURES TO CHARACTERIZE OFF-TALK IN HUMAN-COMPUTER INTERACTION

Rolf Siepmann[+], *Anton Batliner*[*], *Daniela Oppermann*[+]

[+]University of Munich, Institute of Phonetics and Speech Communication, Germany
[*]University of Erlangen-Nueremberg, Chair for Pattern Recognition, Germany
{daniela.oppermann, rolf.siepmann}@phonetik.uni-muenchen.de, batliner@informatik.uni-erlangen.de

Abstract

This paper provides a prosodic analysis of so-called Off-Talk in spoken German in human-computer interaction. Off-Talk consists of user utterances, which are not directed to the automatic speech processing system. These utterances have to be mastered automatically as far as possible. The data collection in the SmartKom project is described and problems with the consistent annotation of Off-Talk are discussed. Different forms of Off-Talk are distinguished and compared prosodically with each other as well as with speech, which is directed to the system. The analysis includes various perceptual and acoustic prosodic features. There are clear differences in the distribution of prominent accents and phrase boundaries found. F0 range turned out to be a further relevant feature. In future, a refined definition of Off-Talk has to be applied, which fulfills our requirements for consistent and efficient annotation.

1. Introduction

There is a growing need in exploring human-computer interaction from a conversational point of view. At present, automatic dialogue systems are designed with language facilities, which ideally should enable natural dialogues between the user and the system. In order to realize these systems, human linguistic behavior has to be studied thoroughly in this context.

A difficult task of dialogue systems consists in the automatic processing of user utterances being not a part of what might be called a coherent and coordinated human-computer dialogue. Phenomena like Barge-In, fragmentary utterances and different kinds of repairs like repetitions or reformulations are usually discussed in this context. In this paper we carry out a prosodic analysis of the hardly studied phenomenon Off-Talk in spoken German in human-computer interaction. Off-Talk is a special type of conversation. It is defined preliminary as users utterances in human-computer-interaction, which are not directed to the automatic speech processing system [1]. The automatic recognition of problematic user utterances is moreover of crucial importance for the analysis of the users' (emotional) state, i.e. his actual positive or negative attitude towards the chosen dialogue system [2]. An automatic analysis of these states might initiate an implemented dialogue strategy, which prevents the current dialogue from breakdown. We believe, that an early recognition of (increasing) Off-Talk will help to detect critical phases in human-computer interaction.

2. The SmartKom Corpus

The data presented in this study are based on a corpus created within the German research project SmartKom [3]. The project aims to develop a multimodal and easy-to-use dialogue system which simultaneously processes the linguistic utterances, the gestures and even the facial expressions of the user. In SmartKom, human-computer interaction involves a life-like animated virtual communication assistant placed on a graphical display. The assistant communicates with the user and carries out certain instructions like providing cinema information on his graphical display. There are three versions of the SmartKom dialogue system planned, each including specific services, namely a public version for airport and train launches, a mobile version for use in cars and for pedestrians, and a version for use at home and/or at office.

In general, SmartKom has the ambitious goal to recognize user states like anger, irritation, joy, surprise and so on in order to react immediately to misguided dialogues. For instance, in case of a user, who is getting more and more angry because SmartKom does not behave in the desired way, the system should interrupt and ask directly for (new) instructions. To provide such kind of functionality, SmartKom must be able to analyze user states like anger automatically in the verbal and non-verbal behavior of the user.

2.1. Wizard-of-Oz Experiments

The Institute of Phonetics and Speech Communication of the University of Munich is at present conducting Wizard-of-Oz experiments with the aim of collecting relevant empirical data both for the design of SmartKom and for its statistical training. The subjects of the experiments do not know that they are working with a simulated system. Moreover, the Wizard-of-Oz on purpose works sometimes incorrectly in a certain way, so that reactions to such system errors are elicited.

The task of the subjects during a session of the experiment is to use the system for several services, e.g. reserving tickets for a cinema film. The subjects are performing this task by verbal instructions to the virtual communication assistant and/or by pointing at certain textual or graphical information presented at the display. For the sake of human-computer interaction, the output facility of the manually driven simulated system consists of a speech synthesis module and of the screen just mentioned. The various sessions are recorded with different cameras and microphones. As the SmartKom corpus should contain a broad variety of speakers, foreign German speaking subjects are recorded as well. In our project a database of approximately 300 sessions will be created, so that the results presented in this paper are in fact based on a constantly increasing empirical basis.

2.2. Levels of Annotation

The recorded sessions of the experiments are annotated manually on various levels. First the linguistic utterances are

transliterated with a broad orthographic transcription scheme. This scheme includes some additional symbols for information like pauses or hesitations. Next prosodic events are transcribed quite broadly too, i.e. essentially primary and secondary accentuation are marked in a phrase. Further parts of the prosodic annotation are the slope of the pitch curve at the end of a phrase as well as certain boundary symbols. Both prosodic and orthographic transcription are based profoundly on experiences gained in the long-termed former project Verbmobil [4]. All transliterations are checked several times by different persons during the annotation process.

Furthermore, in SmartKom the gestures of the subjects interacting with the graphical display and their different user states during a session are annotated. The actual user state is referenced to facial expression and/or speech, which means, that the state is coded in the facial expression, in the linguistic utterances of the user or in both information sources. Consequently, an annotation process has been developed, mainly inspired by the second author, which consists of three different and deliberately disconnected phases. The three annotation phases are realized separately by different working groups. The first phase produces a so-called holistic annotation involving both information sources. In this case, multimedia videos are inspected containing exclusively the facial display of the subjects as well as their acoustic utterances. The classification of the different user states is grounded on an experimental set of assumed states, which are assigned in a subjective manner by the annotating person. In other words, formal criteria like raising eyebrows as a sign of surprise are on purpose not operationalized. The unique difference between the first and the second phase is the missing availability of the acoustic utterances, i.e. in the second phase solely the facial displays are annotated without any audible utterances. In the third phase a prosodic analysis of the linguistic utterances is accomplished using formal parameters like syllable lengthening or hyperarticulated speech. These perceptual prosodic parameters of emotional speech were also developed within the Verbmobil project [2].

3. Types of Off-Talk

Our prosodic analysis of Off-Talk is based in the first place on a classification of the turns in those with Off-Talk and in turns containing no Off-Talk (no OT). At the second place the turns with Off-Talk are subdivided in read Off-Talk (ROT), other Off-Talk (OOT) and not Off-Talk (NOT). There are turns completely marked as OOT and/or ROT.

ROT and OOT are classified during the broad orthographic transliteration of the recorded dialogues. ROT is transliterated, if the subjects are reading a displayed text aloud in a way, that their reading is not meant to be part of their verbal interaction with the (simulated) dialogue system. The occurrence of ROT obviously depends on the specific design of the SmartKom dialogue system, namely its use of a textual display. All other forms of Off-Talk are classified as OOT; further details of these types of Off-Talk are discussed in [1].

The classification of Off-Talk results from an interpretation of the verbal behavior of the subjects. In practice, a homogeneous interpretation and accordingly consistent annotation of Off-Talk, especially of OOT, turned out to be a difficult task. The following literal translated dialogue from German illustrates the problem.

SmartKom: *This is the location of the cinema Studio Europe, which shows the film Mars Attacks.*
User: *Would you like to zoom in here [↑], I can hardly recognize it.[...]. Are there any buses going there?*

Remember, that the user is acting in front of a display, where certain kinds of his gestures are analyzed too; accordingly [↑] stands for a gesture in this case on a displayed map to the location of the named cinema (cf. [3] for this notation). The brackets with the dots should be replaced by one of the following sentences:

a) You zoomed in fine, stop it now.
b) Let me think a moment, I should work tonight.
c) I see, this doesn't work.

For the sake of our discussion, let us assume in case of a) and b), that the system still zooms in the map when the user utters one of these sentences. Contrarily, in case c) the system only highlighted the cinema on focus or something like this. What follows is, that in case of b), the system has to proceed zooming, while in case of a) it has to stop zooming immediately. In case c) on the other hand the user adopts himself to a system malfunction; now the system should present information about the relevant bus lines. Thus the system has to process a) and to ignore b) and c).

Now, longer utterances instead of b) or c) are quite easy to imagine in the above context. These utterances have to be classified as Off-Talk too. This presupposes fundamentally a sophisticated understanding of dialogue pragmatics in human-computer interaction [5]. As already mentioned, (increasing) Off-Talk should be estimated also as an indication of a (growing) negative user state and/ or of system malfunctions. In order to improve future annotations of OOT, we propose to distinguish at least between out of conversation Off-Talk (cf. case b) and out of dialogue Off-Talk (cf. case c).

4. Off-Talk Analysis

In our SmartKom corpus we analyzed 152 dialogues having a duration about four to five minutes each. The dialogues consist of 18847 words in total and are realized by 78 native speakers of German. Most of our subjects performed a recording session, which consisted of two dialogues between the subject and the (simulated) system as interlocutors. In each dialogue a certain task has to be performed. Nearly all dialogues (138) are recorded with background street noise, which simulates a real-life environment; cf. the planned different versions of the SmartKom dialogue system mentioned above.

Chronologically, from dialogue 80 onwards a somewhat different scenario has been used, which resulted in a decrease of Off-Talk. In the case of ROT, the decrease is explained by a change of the use of the display, i.e. less textual information has been presented. The impact on OOT is most likely explained by an improved guidance of the subjects during the recordings sessions. The following diagram I contains the frequencies of the words marked as OOT or ROT. The frequency is expressed as percentage of the number of words marked as Off-Talk in the dialogues 1 to 152 (x-axis). The quadratic polynomial trend lines of the diagram show, that there is a stronger decrease of OOT than of ROT at about dialogue 80. Moreover, there exists a marked variability between the

speakers in the amount of produced Off-Talk.

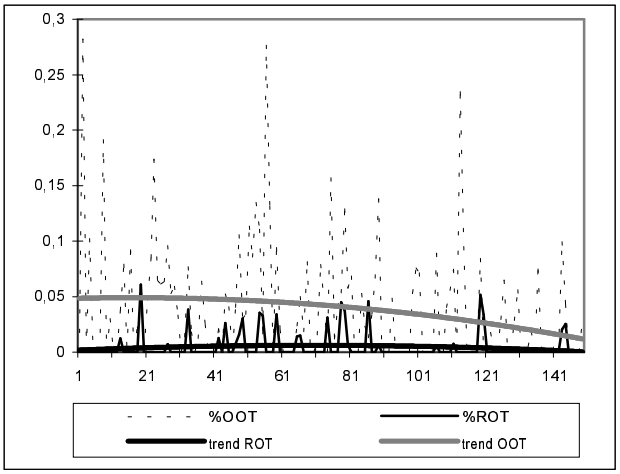

Diagram I: Frequency of Off-Talk

In view of the individual dialogue histories, OOT occurs mostly in the first and in the second quarter of a dialogue. Diagram II represents on the x-axis the positions of the turns in the dialogues. On the y-axis the number of OOT words per turn divided by the total number of words in the turn are listed. The turn positions on the x-axis are normalized by the total turn number of a dialogue, e.g. the second turn in a dialogue with 10 turns of a subject is represented on the x-axis as 2/10. Only turns with OOT are considered in diagram II, while the position of a single turn depends on the total number of turns of the subject in the dialogue. The x - values are sorted in ascending order on the x-axis.

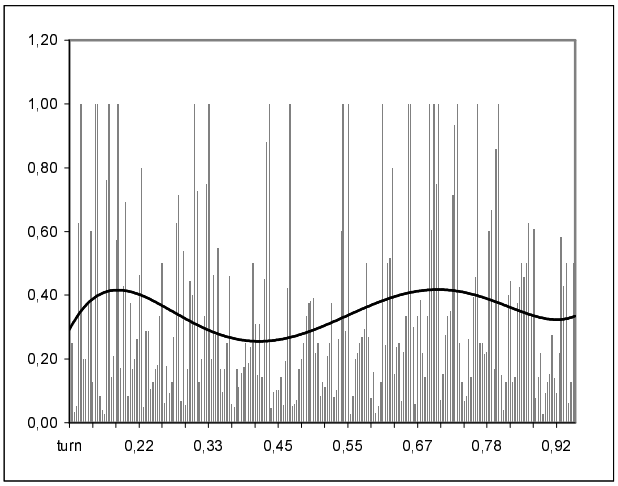

Diagram II: OOT in Dialogue History

As already mentioned, diagram II shows accumulations of OOT in the first and in the third quarter of the dialogues. The accumulations are best visualized in the diagram by a polynomial trend line of the degree of four. The course of the trend line is explained by the experimental design of the recorded dialogues. It is a possible interpretation of diagram II, that the subjects need some time for orientation at the beginning of a dialogue. Afterwards they perform their scenario specific task, which is in the fourth quarter connected with some purposely initiated communication problems. In this phase, the Wizard-of-Oz reacts for instance with system outputs like "I did not understand you" or "This function is not available" (literal translation). The first and the fourth phase of the dialogues might be characterized therefore as phases of deliberation, which are of course predestinated for the generation of Off-Talk.

In the 152 analyzed dialogues 68 contain OOT and/ or ROT. On the average, each part of a subject of a dialogue consists of 131.0 words and of 14.7 turns. On average, in each turn about 8.8 words are spoken by a subject. Counting the turns containing Off-Talk reveals, that 9,7% of the turns exhibit OOT and 2.2% ROT. These numbers are not additive, because some turns contain both OOT and ROT. In relation to the number of words in the turns containing Off-Talk, there are 34.6% OOT and 22.0% ROT words. So, an amount of roughly 10% turns with Off-Talk containing more than 30% words marked as OOT is worth to be investigated prosodically.

4.1. Prosodic Labeling

The broad prosodic transcription symbols used in our annotation process should be described in more detail [6]. We distinguish four types of word-based boundary labels and again four different types of syllable-based accent labels. The accent labels are mapped onto word-based labels denoting if a word is accentuated or not. The number of accents label within a phrase is not restricted. A normal word boundary is not labeled explicitly, so in practice the following symbols are used:

- PA: prominent phrase accent
- NA: secondary phrase accent
- EK: emphatic phrase accent
- B3 cont(inuation): full intonational boundary with strong intonational marking, combined with a final level intonation contour
- B3 rise: see above, with a final rising contour
- B3 fall: see above, with a falling rising contour
- B2: minor phrase boundary with weak intonational marking
- B9: irregular prosodic boundary (e.g. hesitation or repair)

Table 1 describes the distribution of the prosodic labels of the words, which are marked as no OT, NOT, OOT or ROT. The numbers express the percentage of the assigned prosodic label relative to the number of no OT, NOT, OOT and/or ROT words in a turn. In order to prove the significance of the data underlying table 1, we conducted eight non-parametric statistical tests (Friedman) for the prosodic label. The variables of the tests were the four columns of table 1. Assuming a level of significance $\alpha = 0.05$ and following the formula $1 - (1- \alpha)^{1/c}$ (c being the number of tests, in our case 8) provides an adjusted level $\alpha = 0.0063$. Each test yields $p = 0.000$ and thus significant differences.

In Table 1 no OT and NOT are quite close together, so that we will concentrate on NOT, OOT and ROT. Firstly, there is an increase of prominent accents from NOT to OOT and finally to ROT. Secondary accentuation is almost similar and emphatic accents are nearly not found in the SmartKom corpus. Secondly, there are more B3 cont and B3 fall boundaries

in OOT and ROT, while there are more B3 rise boundaries in NOT. Minor phrase boundaries B2 are less realized in OOT in comparison to NOT and ROT. The differences between the occurrences of B9 boundaries are quite similar to those for B3 rise, i.e. B9 boundaries are realized most in NOT.

Table 1: Perceptual Data

	no OT	NOT	OOT	ROT
PA	24.0	23.3	26.5	30.7
NA	19.9	20.5	18.9	21.7
EK	0.2	0.2	0	0
B3 cont	5.9	6.3	10.6	10.3
B3 rise	3.8	4.3	0.7	2.1
B3 fall	10.9	9.7	17.1	13.5
B2	9.2	9.9	5.6	9.7
B9	2.1	2.9	0.9	1.0

4.2. Acoustic Measurements

The acoustics of NOT, OOT and ROT is measured with Praat [7], whereby an autocorrelation method was used for the computation of f0 [8]. Due to the background noise in most of our recordings the voicing threshold was set to 0.7. In the computed pitch contours octave jumps were deleted and finally the contours were smoothed by 10 Hertz.

Table 2 contains the mean f0 minimum and maximum (Hertz), the mean minimum and maximum intensity (db) and the mean duration (sec) values again of the no OT, NOT, OOT and ROT parts of the investigated dialogues. Furthermore, the values of mean standard deviation of f0 and intensity are specified. The range of f0 and of db is computed from the min and the max values.

No OT is now separated from NOT, OOT and ROT, though it should be noted, that no OT utterances are in the mean much more longer than other utterances (48.2 sec). There is also a clear distinction in the f0 ranges of no OT on the one hand and of OOT and ROT on the other hand. In Table 2 the db range decreases from the left to the right column.

Table 2: Acoustic Data

	no OT	NOT	OOT	ROT
f0 min	93.1	102.3	119.0	116.4
f0 max	285.0	186.2	153.5	152.3
f0 range	**191.9**	**83.9**	**34.5**	**35.9**
f0 dev	46.2	21.3	9.7	9.6
db min	36.3	36.7	39.8	27.3
db max	60.6	56.5	55.9	38.8
db range	**24.3**	**19.8**	**16.1**	**11.5**
db dev	5	4.8	4.5	3.0
duration	48.2	6.5	1.7	1.6

5. Discussion and Conclusion

The distribution of prominent accents correlates probably with the occurrence of content words [9]. We assume, that in ROT in contrast to OOT and NOT more content words like cinema or film names are realized, which preferably carry a prominent accent. Moreover, Off-Talk is primarily realized with the prosodic boundaries B3 cont and/or B3 fall, while B3 rises are found most in NOT. B3 rises are connected usually with questions. As mentioned above, Off-Talk should be conceptualized as out of dialogue and/or out of conversation, so that the realization of questions is hardly motivated in Off-Talk. The amount of prosodic boundaries correlates in general with the number of syntactic phrases. Hence, Off-Talk is realized with more syntactic phrases than NOT.

In future, our hypotheses have to be investigated further. Currently a part-of-speech annotation is produced, so that the assumed correlation of prominent accents and part-of-speech classes will be explored profoundly. Furthermore, a richer syntactic boundary annotation is created along the lines of [6]. This enables the investigation of the specific character of the phrases in Off-Talk. The annotations of the user states mentioned above will result in an even better understanding of Off-Talk. We believe, that user states and Off-Talk correlate in a way, i.e. in the case of 'out of dialogue' Off-Talk the user will comment the systems behavior in a positive or negative up to a certain extent. Of course, more dialogues will also be recorded and annotated in future, so that our database increases constantly.

Acknowledgments

We would like to thank Susen Rabold for valuable information about the transliteration process and Victoria Penide Lopez for her skilful Perl programming. This work was founded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the SmartKom project under Grant no. 01 IL 905. The responsibility for the contents of this study lies with the authors.

6. References

[1] Oppermann,D., Schiel,F., Steininger,S. and Beringer,N., Off-Talk – a Problem for Human-Machine-Interaction. Proc. of EUROSPEECH, vol. 3, pp. 2197-2200, Aalborg, 2001.

[2] Batliner,A., Fischer,K., Huber,R., Spilker,J. and Nöth,E., How to Find Trouble in Communication. To appear in Speech Communication.

[3] Wahlster,W., Reithinger,N. and Blocher,A., SmartKom: Multimodal Communication with a Life-Like Character. Proc. of EUROSPEECH, vol. 3, pp. 1547-1550, Aalborg, 2001.

[4] Burger,S., Transliteration spontansprachlicher Daten. Verbmobil Technisches Dokument 56, 1997.

[5] Bunt,H., Dialogue pragmatics and context specification. In: Bunt,H. and Black,W. (eds.), Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics. Amsterdam: Benjamins, 2000, vol. 1, pp. 81-150.

[6] Batliner,A., Kompe,R., Kießling,A., Mast,M., Niemann, H. and Nöth,E., Syntax + Prosody: A syntactic labeling scheme for large spontaneous speech databases. Speech Communication 25, 1998, pp. 193-222.

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

[8] Boersma,P., Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound. Institute of Phonetic Sciences. University of Amsterdam. Proceedings 17, 1993, pp. 97-110.

[9] Batliner,A., Nöth,E., Buckow,J., Huber,R., Warnke,V., Niemann,H., Duration features in Prosodic Classification: Why Normalization Comes Second, and what they really Encode. This volume.