Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

# Off–Talk – A Problem for Human–Machine–Interaction?

Daniela Oppermann
Nicole Beringer
Florian Schiel
Silke Steininger

Ludwig–Maximilians–Universität München

**Report Nr. 6**
Juni 2001

Juni 2001

Daniela Oppermann
Nicole beringer
Florian Schiel
Silke Steininger

Ludwig Maximilans Universität München
Schellingstr. 3
80799 München

Tel.: (089) 2180−2806
FAX: (089) 2800362

E−Mail: daniela.oppermann@phonetik.uni−muenchen.de

**Dieses Technische Dokument gehört zu Teilprojekt 1: Modalitätsspezifische Analysatoren**

Das diesem Technischen Dokument zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01 IL 905 gefördert. Die Verantwortung für den Inhalt liegt beim Autor.

# Inhaltsverzeichnis

# Abstract

This paper is concerned with the definition and description of the phenomenon Off−Talk in human−machine−interaction. This phenomenon is considered to cause problems due to non−relevant information that is conveyed within these utterances. Besides the definition of Off−Talk our work aims to provide an analysis of transcribed audio data that is part of the SmartKom[1] data collection. In the search for features that could indicate the occurrence of Off−Talk we looked at several speech levels e.g. acoustics, lexicon and prosody. Due to the small amount of available data only three features were examined, as there are: *loudness, word frequency* and *filled pauses*. The analysis revealed that a correlation might exist between Off−Talk and all features, so that they may serve as indicators for this phenomenon.


# 1 Introduction

Human communication primarily consists of spoken language, which serves to convey information from one person to another. The speech style a speaker uses in a communicative situation depends on several factors, e.g. where the conversation takes place, how many speakers are involved, whether the speaker is giving a talk or is perhaps absorbed in soliloquy. A very particular form of communication is found in the interaction between a person and an artificial speech understanding system.

Two factors play an important role in human−machine−interaction and have an impact on the user behaviour as well as the speech he used.

- The first influence is that the user expects a certain service or wants to get information from the communication system.
- The second factor is that the user is aware of the fact that he is dealing with artificial intelligence.

The impact of both factors on the users' speech behaviour is, that it is less spontaneous than in human−human communication. To be understood by the system the user probably plans more carefully what to say. An explorative study of what kind of speaking style people choose while working with a computer was carried out by E. Kachelrieß [3]. In this study she found, that speech behaviour is dominated by clear articulation, short sentences and that many filled pauses occur, but very few self corrections.

These assumptions seem to make it quite easy for a dialogue system to recognise what the user intends and to react properly. But there still exists a large number of misunderstandings and inappropriate reactions on the part of the machine. Several reasons can be named for this misbehaviour. For example, the system has to deal with an unlimited number of unknown speakers and speaking styles, dialectal and accent variations, background noise and an infinite number of words.

In most cases the machine is "listening" permanently to what the user utters; that means that every kind of user input is processed. But in some cases utterances are not relevant for the system or are even not meant to be processed by the system. This kind of speech we refer to as Off−Talk.

In this paper we describe the phenomenon Off−Talk and the problems which arise for the speech understanding system. First, we give a definition of the phenomenon and a short overview of the different forms it may take shape. Next, we will show which of these forms of Off−Talk can be found in our data, which is part of the empirical data collection of the SmartKom project[2]. Finally, we give some numbers on occurrences and conclude with a discussion of why Off−Talk is a problem for human−machine−interaction and whether any features exist that allow it to be recognised and extracted automatically.

# 2 Definition

As mentioned above, communication between the system and a user may contain speech that is not always relevant for the system or even not meant to be "heard" by the system. We call this phenomenon Off−Talk. We define it as follows:

**We define Off−Talk to be every utterance that is not directed to the system as a question, a feedback utterance or as an instruction.**

---

1 This research was founded by the German Federal Ministry of Education and Research, grant no. 01 IL 905.

2 http://smartkom.dfki.de

The following is an open list of possible utterances which could be considered as Off−Talk in human−machine−interaction:

- soliloquy / thinking aloud
- swearing
- reading from displayed text aloud
- conversation with other person(s) present
- telephone conversation (e. g. with cellular phone)
- extrinsic speech (e. g. video player, TV set, etc.)

Utterances directed by a user to a dialogue system that fit into one of the mentioned categories do not deliver any information and therefore should be ignored. Trying to process such utterances may cause problems, leading to misinterpretations and may interrupt a smooth conversation.

To train a system to defeat Off−Talk the phenomenon must first be tagged in the data and features have to be extracted which make it possible to identify Off−Talk automatically. In search for such features we made an analysis of the transcribed audio data that were collected up to now in the SmartKom project.

# 3 Data

The SmartKom project aims to develop a multi−modal communication interface where the user can interact with the system via multi−modal input channels. He may choose to communicate with SmartKom by speech or with the additional help of gestures. The system is able to react adequately with several output modalities (speech synthesis, text and graphic output).

The data collection for this system is done via Wizard−of−Oz−experiments at the Institut of Phonetics and Speech Communication of the University of Munich. Naive test persons have to fulfil a task (e.g. reserve tickets for the cinema in town for tonight) through the SmartKom system. The whole dialogue is recorded with several microphones and digital cameras. The audio channels are transcribed separately without regarding the information of the video stream.

The conventions for transcription [1] are widely based on the handbook for the annotation of spontaneous speech [2] which were developed within the Verbmobil project [6]. The handbook had been modified and amplified according to the WOZ−recordings. Adaptations in the annotation handbook were necessary because of the special communicative situation, the subjects being expected to behave less spontaneously (as mentioned before). What further complicates the communication is the ability to provide additional information to the system via gestures. Aside from the changes there are also several new phenomena which had to be added to the handbook. A symbol which had to be added and what is characteristic for the human−machine−interaction is the phenomenon Off−Talk. It will be discussed in detail in the following.

## 3.1 Annotation rules

The transcription of Off−Talk is divided into two categories (corresponding to the categories we found in our data)

- read Off−Talk −> ROT
- other (forms of) Off−Talk −> OOT

and are labelled on several levels in the transcribed audio data:

- Dialogue level
- Turn level
- Word level

### *3.1.1 Dialogue level*

At the first step the general occurrence of Off−Talk in a WOZ dialogue is noted in the header of the transcription file. The header of a transcription file conveys overall information about the subject, date of recording, transcriber ID, etc. The amount of Off−Talk is noted in three categories.

- none −> no Off−Talk occurs at all
- little −> less than 10 % of all words are Off−Talk
- much −> more than 10% of the words are Off−Talk

#### *3.1.2 Turn level*

Within the transcription there are two possible ways to annotate Off−Talk. In the case that a whole turn fulfils the criteria of Off−Talk it is marked by the tags <t*ROT> or <t*OOT> at the beginning of a turn annotation.

#### *3.1.3 Word level*

In cases where only parts of the turn are concerned, every word of this part is tagged by the corresponding symbols for "Read" (<ROT>) or "Other Off−Talk" (<OOT>).

### 3.2 Examples

Here are given some examples of the transcriptions to illustrate the annotation rules and to show some typical cases we found in our data.

In the WOZ recording situation the subject is left alone with the communication interface and is not allowed to keep any technical equipment which could disturb the experimental surroundings (such as cellular phone etc.). This is done to avoid changing recording conditions. Consequently only three of the possible Off−Talk categories can be found in this data. Phenomena such as "extrinsic speech", "telephone conversation" or "conversation with another person" do not occur. The only category which can be found in our data easily are parts which are read aloud by the subject (<ROT>). Parts where the subject thinks aloud / speaks to himself or is swearing were not differentiated any further and were therefore grouped together into one category (<OOT>).

#### *3.2.1 Read Off−Talk*

Examples of Off−Talk that belong to the category where the subject read aloud[3]:

- ***parts of a turn:***

gut . <P> in welchem Kino l"auft das ? in welchem Kino l"auft ~Armageddon ? <P> **Kino<ROT> ~Hoelldobler<ROT> , ~Wilhelm−Blum−Stra"se<ROT> .**

(*english version*: okay, in which cinema is that shown? In which cinema Armageddon is shown? **Cinema Hoelldobler, Wilhelm−Blum−Street.**)

- ***whole turn:***

**<*tROT>** <"ah> **Regie ~Mike ~Newell , mit** <P> **~Annie ~MacDowell . ~Vier+Hochzeiten+und+ein−+Todesfall ist die Geschichte von #acht Freunden , #f"unf Pfarrern** , <A> **#elf Hochzeitskleidern , #sechzehn Schwiegereltern , zw=** <*T>t

(*english version*: **directed by Mike Newell, with Annie MacDowell. Four−weddings−and−a−funeral is the story of eight friends, five priests, eleven wedding dresses, sixteen parents in law, tw=**)

#### *3.2.2 Other Off−Talk*

Examples of Off−Talk where the subject thinks aloud or speaks to himself:

- ***parts of a turn:***

gibt 's 'n paar nette Restaurants , italienisch , m"oglichst , in der Innenstadt ? <P> <A> **au<OOT> , das<OOT> sieht<OOT> ja<OOT> schon<OOT> gut<OOT> aus<OOT>** .

(*english version*: are there any nice restaurants , italian if possible, in the city centre? **Oh, that is already looking quite good.**)

- ***whole turn:***

**<t*OOT> ich mu"s den !KEYAladdin ber"uhren , aha . jetz' geht das**.

(*english version*: **I have to touch Aladdin, okay. Now it works.**)

---

3 Off−Talk parts in the given examples are highlighted with bold letters and most of other annotated phenomena are filtered out of the text to make it readable.

# 4 Analysis

At the writing of this paper not all dialogues had yet been transcribed completely, so that our analysis is based on only 81 transcriptions; These were corrected by at least 3 different transcribers[4].

## 4.1 General occurrence

First we counted the entries in the dialogue header regarding how many Off–Talk occurrences appear in the dialogues. The results are shown in Table 1.

*Table 1: General occurrence of Off–Talk*

| ***Off–Talk*** | ***Percentage*** |
|---|---|
| none | 34.6 |
| little | 58 |
| much | 7.4 |

About two thirds (65.4%) of the subjects  use Off–Talk while communicating with SmartKom.

## 4.2 Words affected

At the next step we looked at the distribution of that phenomenon within the dialogues. We counted the affected words in the subjects' turns (only the dialogues in which Off–Talk was present were considered).

*Table 2: Percentage of Off–Talk–Words of all words*

| ***Off–Talk*** | ***number*** | ***%*** |
|---|---|---|
| OOT | 543 | 6.1 |
| ROT | 290 | 3.2 |
| all | 833 | 9.3 |

A total of 8947 words spoken by subjects, 833 were not directed to the system, i. e. roughly 10 % of irrelevant information where the subject mostly speaks to himself / thinks aloud (6, 1%) or is reading the displayed text (3,2%).

In a next step we tried to find out whether there exist special characteristics of the phenomenon Off–Talk which possibly makes it detectable by a dialogue system like SmartKom. We took a closer look at the Off–Talk annotated passages in the transcripts in search of possible features.

## 4.3 Possible Off–Talk features

We analysed the following levels in the data:

- acoustic level
- lexical level
- prosodic level

Because of the small amount of data the analysis can not cover all features of these levels entirely.

### *4.3.1 Acoustic level*

Within the acoustic level we concentrated on the feature loudness in the Off–Talk segments. Whenever the speaker was lowering his voice or his speech was not identifiable anymore a special comment was made in the annotations.

*Table 3: Percentage of low voice in Off–Talk words*

| ***Low voice comments in general*** | ***%*** | ***Low voice comments in Off–Talk words*** | ***%*** |
|---|---|---|---|
| all | 0.7 | OOT | 16.6 |
| | | ROT | 9.6 |
| | | all | 10.8 |

4 Several correction passes are necessary to secure high quality transcriptions. For an analysis on the necessity of correction passes and the avoidance of transcription errors see [4].

The numbers show that only about 11% of the Off−Talk words were uttered with lower voice or were less understandable. But in comparison only 53 words (0.7%) had a comment on low voice in non Off−Talk words (8114). Therefore, the feature *loudness* may be considered as a significant indicator for Off−Talk.

### *4.3.2 Lexical level*

Next we looked at the lexical level to find possible indicators for Off−Talk. We hypothesised that there may be some keywords which could be typical for Off−Talk. We compared the 8 most frequent words of all words and those words which were marked as Off−Talk.

*Table 4: Percentage of most frequent words in of all turns of the subject and in Off−Talk passages*

| ***Non−Off−Talk*** | *%* | ***Off−Talk*** | *%* |
|---|---|---|---|
| ich (−> I) | 5.5 | mhm (−> uhu) | 5.4 |
| ja (−> yes) | 2.3 | ja (−> yes) | 3.4 |
| das (−> the, neutral) | 2.1 | gut (−> good) | 2.2 |
| Kino (−> cinema) | 1.9 | ah (−> oh) | 2 |
| in (−> in) | 1.6 | das (−> the, neutr.) | 1.9 |
| der (−> the, masculine) | 1.5 | so (−> so) | 1.8 |
| du (−> you) | 1.3 | ich (−> I) | 1.8 |
| die (−> the, feminine) | 1.3 | ist (−> is) | 1.4 |

The comparison of the word frequencies shows that the vocabulary subjects use in the Off−Talk mode differs quite obviously from normal talk. While using the word *ich* most in general, it is used significantly less in Off−Talk (1.8%). Conversely, *mhm* is the most frequent word in Off−Talk passages, in non−Off−Talk speech it occurs significantly less with 0.8% of all words (rank 26).

### *4.3.3 Prosodic level*

On the prosodic level special features like filled pauses[5] could also be seen as indicators for Off−Talk. Filled pauses can be considered as a signal for planning sentences or thinking processes [5] (therefore they have a similar function to <OOT>). Table 5 shows how often a filled pause co−occurred with Off−Talk.

*Table 5: Percentage of turns with filled pausesplus Off−Talk[6]*

| ***Turns with filled pauses*** | *%* | ***Off−Talk−turns with filled pauses*** | *%* |
|---|---|---|---|
| all | 34.7 | OOT | 44.5 |
| | | ROT | 20.2 |
| | | all | 54.3 |

To get a better view of the co−occurrence of the phenomenon Off−Talk and filled pauses we analysed the annotations on the turn level. In 286 turns of all 824 turns the subjects produced filled pauses (34.7%). Next we counted the percentage of Off−Talk turns (21% of all turns) which contain filled pauses. In more than half of the turns filled pauses occurred as well (54.3%). The results show a significantly higher percentage of filled pauses in Off−Talk. A separation of the two categories shows that filled pauses are more associated with "Other Off−Talk"; this was expected due to the similarity in the function they share during communication.

# 5 Discussion

First we have to state that in about two thirds of the WOZ−dialogues Off−Talk occurs and that 10% of all spoken words are included in this phenomenon. This confirms our impression that Off−Talk is worth analysing further.

The results of the analysis show that on all levels a significant difference exist between the general occurrence of the feature and their occurrence in Off−Talk passages.

5 In the transcripts four forms of filled pauses were annotated: general hesitation <h"as>, two kinds of vocalic hesitation <"ah> <"ahm> and non−vocalic hesitation <hm>.

6 The results of "all" in table 5 is not the summary of the separate categories, because there are a number of turns where <ROT> as well as <OOT> occur.

Although being highly significant, the fact that only about 10% of Off−Talk words were uttered with lower voice unfortunately queries the function of the feature loudness in being an indicator. On the other hand we were restricted to the examination of annotation comments; additional acoustic measurements should be done to confirm the results.

The analysis of word frequency distributions also revealed a significant difference between Off−Talk and relevant speech parts. The main difference arises from the <OOT> parts while people were thinking aloud or speak to themselves. The words in the <ROT> parts did not affect the word order because most of the read passages consisted of proper names or dates which do not occur very often. The frequency of the word *Kino* (english *cinema*) in the data can be explained by the topic the subjects had to talk about Therefore it can not be taken as an indicator. Further, it should be questioned, if single words could function as an indicator themselves or whether word combinations or even whole phrases, which have a probability of near zero in the classical language modal, are more reliable.

The reason for the significantly higher occurrence of filled pauses in Off−Talk turns may lie in their similarity in function for planning processes in speech, as mentioned before. The artificial quality of the communicative situation may induce the user to make turn−holding procedures "visible" for the system. In thinking aloud he tries to indicate that he has not finished his turn, yet [5]. As expected the difference is higher in the category "Other−Off−Talk". However, the difference of 54% to 34% probably is still to small to be considered as a practicable indicator for Off−Talk.

# 6 Conclusion

We have to state that the phenomenon Off−Talk, no matter what kind of Off−Talk, represents a problem for a speech understanding system to handle it adequately. Our work represents a pilot study of the phenomenon Off−Talk. Our analysis showed that none of the features taken for themselves serve as an adequate indicator, but a statistically weighted combination of different features is more likely to yield a confidence measure to avoid the processing of non−relevant speech input. The different levels considered could well indicate the right direction for further examinations in this field. Even with the small size of the annotated cases and the small number of features the results still look promising. Currently, the annotated corpus is too small and the features too less to get robust results. The correlation with multi−modal features like gestures should not be neglected and could be another way to find processible features for the automatic extraction of Off−Talk.

# 7 References

[1] Beringer, N., Oppermann, D., Burger, S., *Transliteration spontansprachlicher Daten − Lexikon der Transliterationskonventionen − SmartKom*, SmartKom TechDok−Nr−02, 2001.

[2] Burger, S., *Transliteration spontansprachlicher Daten − Lexikon der Transliterationskonventionen − Verbmobil II*, Verbmobil TechDok.56−97, 1997.

[3] Kachelrieß, E., *Computerbezogene Sprache,* Verlag Dr. Kovac, Hamburg, 1999.

[4] Oppermann, D., Burger, S., Weilhammer, K., *What are transcription errors and Why are they made?,* Proceedings of LREC 2000, Athen.

[5] Sacks, H., Schegloff, E.A., Jefferson, G., *A simplest systematics for the organization of turn−taking in conversation*. Language 50.4, 1974, p: 696−735.

[6] Wahlster, W. (Editor), *Verbmobil: Foundations of Speech−to−Speech Translation,* Springer; Berlin / Heidelberg, 2000.