Possible Lexical Indicators for Barge–In / Barge–Before in a multimodal Man–Machine–Communication

Nicole Beringer, Daniela Oppermann, Silke Steininger

Institute of Phonetics and Speech Communication

University of Munich

80799 Munich

+49 89 2180 5751

[beringer,daniela,kstein]@phonetik.uni–muenchen.de

1 Abstract

In the SmartKom project an intelligent computer–user interface which being deal with various kinds of oral or physical input is developed. The aim of SmartKom is to allow a natural form of communication within man–machine interaction.

One problem with natural computer–user dialogues is how to handle turntaking and –holding, especially when the user's intention is to stop the system output. The goal is to handle turntaking and –holding as in inter–human dialogues and to recognize if a system output is overlapped by another input of the user in order to react properly.

One method to deal with this phenomenon is to find out whether there are any lexical indicators for barge–in or barge–before in man–machine–communication.

This investigation gives a classification of wordclasses and lexemes which are significant in overlapping speech within the SmartKom project. This classification may be used in the system processing to handle barge–in / barge–before.

1.1 Keywords

lexical indicators, Barge–In, multimodality, SmartKom

2 Introduction

In the SmartKom project we are developing an intelligent computer–user interface which can deal with various kinds of input, e. g. speech, gestures or facial expression of the user. This interface allows computer experts as well as novices to communicate quasi–naturally with an intelligent multimodal system.

Of course, there are problems in dealing with man–machine–dialogues. One of them is how to deal with turnholding and turntaking [1], [2].

Even in inter–human dialogues there are situations where it is not clear whether the speaker will continue with his/her speech or whether the dialogue partner wants to take the next turn. Of course, regarding multimodal dialogues, we not only have speech for input and output but also gestural input or graphical output. In this paper we concentrate on the linguistic analysis of the user's linguistic input, namely spontaneous speech[1].

There are lots of situations where we can find overlapping speech because

1. turnholding / turntaking is not or too weakly signalized
2. turnholding / turntaking is misunderstood
3. turnholding / turntaking is not accepted

But what about overlapping speech in man–machine–interactions? Is there any? And if yes, how can it be handled?

This paper tries to give a definition for indicators of overlapping speech for German. The main goal here is to give a possibility for the system to distinguish between different user input by analyzing the orthographic transliterated experimental data.

We start with a short outline of the SmartKom project. Then after giving a short outline of the linguistic user behaviour in man–machine–interaction in section 4 we define in section 5 the technical terms **barge–in** and **barge–before** as well as transliteration conventions and frequency of overlapping speech in a man–machine–interaction.

Section 6 gives and clusters the found overlapped words with respect to their wordclass as well as the definition for lexical barge–in / barge–before indicators.

Technical strategies to handle barge–in / barge–before in the processing of the system are given in section 7.

Finally, results and future work are discussed in the last section.

3 The SmartKom Project

The aim of the SmartKom project (started in January 2000) is the development of a multimodal computer–user interface which allows the user to communicate naturally with an adaptive and self–explanatory machine.

---

1 Another conference contribution [12] analyzes the effects of gestural input in barge–in situations.

SmartKom is being developed for three application scenarios, each with different requirements:

1. SmartKom Home/Office to communicate and operate machines at home (e.g. TV, workstation, radio),
2. SmartKom Public to have a public access to the Internet and other public services, and
3. SmartKom Mobile as a mobile assistant.

The system understands input in the form of natural speech as well as in the form of gestures.

In order to "react" properly to the intentions of the user the emotional status is analyzed via the facial expression and the prosody of speech.

The output of the system is presented though with a graphical user interface – a computer screen projected onto a graphic tablet – and with synthesized language.

3.1 Multimodal Data Collection

In the first phase of the project we are collecting data for the following three purposes:

1. The training of speech, gesture and emotion recognizers.
2. The development of user–, language–, dialogue–models etc. and of a speech synthesis module.
3. The general evaluation of the behavior of the subjects in the interaction with the machine.

In each session of a Wizard–of–Oz experiment the spontaneous speech, the facial expression and the gestures of the subjects are recorded.

The technical equipment is as follows:

➔ Audio:

➔ a microphone array of 4 Sennheiser ME 104

➔ a directional microphone (Sennheiser ME66/KG) and

➔ (alternating) a headset or a clip–on–microphone (Sennheiser ME 104).

➔Video:

➔ a fixed positioned digital camera to capture the face of the subjects (facial expression).

➔ a second digital camera to capture the gestures in a side view of the subject (upper body) and

➔ an infrared camera (from a gesture recognizer: SIVIT/Siemens) to capture the hand gestures (2–dimensional) in the plane of the output projection.

➔Other:

➔ The coordinates of pointing gestures on the work space are recorded (SIVIT)

➔ as well as the inputs of a pen on the graphic tablet.

➔ Graphical user interface as video stream

3.2 Wizard Of Oz Data Collection

For training purposes of the system we are collecting data in Wizard of Oz experiments. Therefore, the subjects were instructed to test a new prototype of a dialogue system. It was suggested that the system could understand spoken language and gestures without giving a detailed explanation either on the functioning or on the gestures the subjects could use. The aim was to encourage the subjects to experiment with their gestures as well as with the system itself.

While working with the system for the first time they had to imagine being in Heidelberg (a German town). Their job was to solve a (certain) task like planning a trip to the cinema / restaurant in the evening, finding a hotel, programming their VCR at home or arranging a sightseeing tour. They also were encouraged to try different input modes. After the instruction two sessions were recorded (with a short break in between). Afterwards, in a different room, the subjects were questioned about problems that had appeared, which aspects of the system they approved of and which they resented, and if the system all in all gave more the impression of being a human or a machine.

For results of their inquiries please refer to [6].

All data is annotated according to speech [7], gestures [10] and emotions.

Whithin the annotation of speech we have several markers that can be combined content–freely but the combination has to follow the defined feature syntax [7].

4 Linguistic user behaviour in man–machine–interaction – how to interpret?

In a man–machine–dialogue the communication is characterized by the information retrieval. In general, the user wants to use a special offer of the computer system, e.g. timetables. The user bears in mind that he/she interacts with an artificial intelligence which causes a less spontaneous communication [9] than with a human dialogue partner.

In an inter–human dialogue situation the turntaking / –holding problem[1],[2],[3],[4],[5] other signals may be set, expected and recognized than in a man–machine–dialogue.

It is interesting to investigate if and what features may indicate to the system that the user is reacting during a system output in order to

➔ deny what is answered (i.e. for the system to stop the whole system output)

➔ ask for further information because some graphical output is given (i.e. for the system to stop the synthetic synthesis) information.

➔ confirm the output (backchannelling)

➔ read the output (Off–Talk [2])

➔ etc.

In our analysis of the recorded speech of the Wizard–

Of–Oz data collection we searched for any possible lexical indicators of the subject that could give hints on barge–in and barge–before.

5 Barge–In and Barge–Before

5.1 Definitions

Barge–In: Barge–in within SmartKom is used as a synonym for ‘‘user reaction (speech, gestures) during the system output (synthesis, display)’’.

Note: Within the transliterations there is no distinction between imprecise and precise reaction.

Barge–Before: Barge–before is defined as the user reaction during the system processing but before the system output.

Note: Audio data is not annotated according to possible system processing. Therefore, barge–before is not marked itself.

Barge–in can be recognized by the real system by overlapping speech, barge–before needs indicators to be recognized in time.

To give any hint for the processing of the real system, these indicators may be obtained by analysing lexical user input which overlaps synthesized speech in the Wizard–of–Oz collection.

Note that we can only analyze the most frequent lexemes in overlapped speech but not if this indicator leads to a correct reaction of the system because the ‘‘system’’ – the human wizard – is required not to react on any interruption!

5.2 Symbols for overlapping speech in the transliterated data

While producing the orthographic equivalent to the audio part of the recordings several inter–speech phenomena like pauses, breathing etc. are marked. Within an unprompted dialogue situation overlapping speech has to be marked as well. This is done by using the following symbols within SmartKom (see [7], [8]):

- ➔ ..n@ (passive overlapping of lexical units)
- ➔ @n.. (active overlapping of lexical units)
- ➔ ..n@> (passive overlapping of other observations, e.g. Pauses)
- ➔ <@n.. (active overlapping of other observations, e.g. pauses)

where *n* stands for the count of the overlapping part.

Transliterations with overlapped speech may look like the following example:

w001_pkd_002_AAA: I’d like to go to the cinema tonight. ~The fifth=1@

w001_pkw_003_SMA: @1here @1is @1an overview about what is currently showing at cinemas in ~Heidelberg.

w001_pkd_004_AAA: uhu. oh, ~Fight+Club, ~The+fifth+Element, ~Drum, ~Jakob+der+L"ugner. ah, I’d like to see ~Aimee+und+Jaguar. where’s that showing? ~Aimee+und+Jaguar, Germany, #ninety–nine2@.

w001_pkw_005_SMA: @2here @2you @2can @2see additional information on the film ~Aimee+und+Jaguar.

w001_pkd_006_AAA: uhu. under a double threat of bomb war and persecution #two young women experience the love of their lives. for one of them it turns into the decisive experience of her life, for the other one this love means hope. well, and at what cinema and where? ah, okay the ~Hol= ~Hoell_ _dobler cinema, the ~Kammer cinema, the ~Gloria cinema, the ~Schlo"s cinema, the ~Hengstler cinema. ah, and where am I right now? ’couse I don’t see the3@ cinemas3@ he=3@

w001_pkw_007_SMA: @3unfortunately @3this @3information is not available at the moment.

w001_pkd_008_AAA: I see. okay, in that case I will have to find it myself on this map here. I think I’ll take the closest one then. maybe ~Hauptstra"se #forty–two? ah, the ~Schlo"s cinema. at what time? ah, the film starts at #eight #thirty–five. at the ~Schlo"s4@ cinema4@.

w001_pkw_009_SMA: @4would @4you @4like to have tickets reserved?

The numbers within the marker show which active and passive overlapping parts belong together.

In principal, all markers can be combined. In the data ‘‘@2#eight’’ is possible.

5.3 Frequency of overlapping speech in a man–machine–interaction

Based on the transliterations we can compare the frequency of turns showing overlapped speech of unprompted inter–human dialogues [8] and man–machine dialogues of SmartKom.

Within 342 inter–human dialogues of the German Verbmobil Part Two [8] we can find turns out of (92.40%) including overlapping speech.

The frequency of dialogues with overlapping speech in unprompted man–machine–dialogues is around half of that.

The 74 analysed SmartKom dialogues show 36 with overlapping speech (48.65%).

Possible reasons for this phenomenon are:

- ➔ Interhuman dialogue partners may be characterized as more capable of adapting to their dialogue partners.
- ➔ Man–machine–dialogues are more adaptive on the human part, because the user does not expect an artificial intelligence equivalent to his own, and
- ➔ therefore only interrupts the output if there are long pauses in between.

Man–machine–communication is not as spontaneous as inter–human communications maybe because the artificial intelligence does not adapt to the speaking style of the user yet. Therefore, the dialogues are somewhat prompted and do not have as many overlaps as more spontaneous dialogues.

6 Classification for indicators for Barge–In

Having given some background information this section comes back to the goal of this paper **to find lexical indicators for barge–in.**

In the case of a multimodal system like SmartKom, it goes without saying that lexical indicators are not the only ones to be searched for.

For example, t can be a relation between barge–in and the use of gestures in man–machine–interaction [10] .

6.1 Classification and frequency of overlapped word categories

To find out whether there are any lexical indicators for barge–in / barge–before we analyzed which lexemes could be found in the overlapped regions and clustered them by word category.

Table 1 gives the distribution of word categories found in the overlapped parts (in total 265 overlapped words).

word category	probability for the types of lexemes found in the overlapped regions
Adjectives	0.81%
Adverbs	20.56%
Determinators	9.71%
proper nouns	2.82%
nouns (rest)	5.24%
pronouns	18.55%
verbs	15.73%
numbers	0.40%
conjunctions	2.02%
particles	20.97%
interjections	12.90%
prepositions	0.81%
interrupted words	4.44%
hesitations and Off–Talk	13.74%
overall number of overlapped words	265

Table 1: *Distribution of overlapped word categories*

Note that hesitations are per definitionem part of Off–Talk , i.e. "any utterance which is not directed to the system as a question, a feedback utterance or an instruction" [11].

6.2 Definition of lexical indicators

Having analyzed the lexemes of the word categories of table 1 we found the following categories (shown in table 2) highly probable for overlapping speech and therefore could signalize barge–in / before:

word category	Probability for occuring in the overlapping speech
Adverbs	20.56%
pronouns	18.55%
verbs	15.73%
particles	20.97%
interjections	12.90%
hesitations and Off–Talk	13.74%

Table 2: *Significant overlapped word categories*

Although interrupted words are not highly (4.44%) probable for barge–in in the analyzed corpus they should not be neglected as they could be interpreted as a (self–) interrupted turntaking.

Figure 1 shows the frequency of the most often used pronoun ''ich'' (I), particles ''ja'' (yes), ''okay'' and hesitation ''mhm'' in overlapped speech.

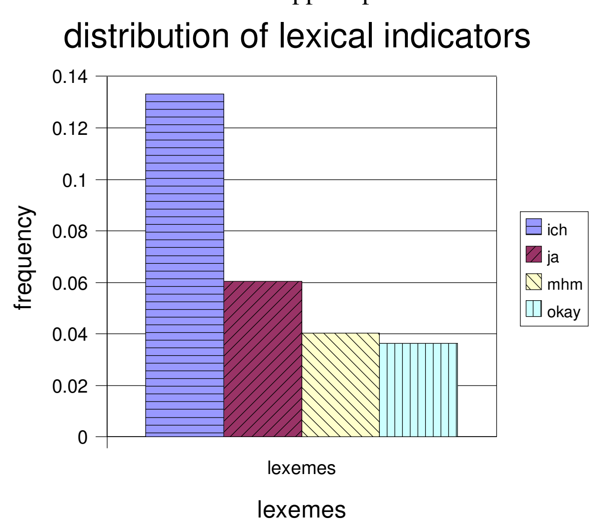

Fig. 1: *Distribution of lexical indicators*

Interpreting these results it can be said, that if the system recognizes words like ''ich'', ''ja'', ''okay'' and ''mhm'' during the processing of another retrieval or the system output, it is highly probable that one of the phenomena barge–in or barge–before is present and the system has to react accordingly.

That means the system has to interpret the input as confirmation (so called backchannelling) or contradiction and stop the whole output in case one. Only the synthesis has to be stopped in case two if other factors for barge–in / barge–before are positive as well.

7 Strategies to handle Barge–In / Barge–Before

Within the project there are some strategies being implemented to solve the barge–in / barge–before problem in the real system partly including the information from of this investigation.

The main idea here is to collect all relevant data in a separate pool (place for the interaction between the modules of the system) until the processing or the system output is finished.

One module – the recognition of intention – decides how to react to the different circumstances for barge–in / barge–before in each situation.

There are four possible reactions:

➔ the system has detected several negations: it has to stop the whole system output, because the user denies what is answered

➔ the system has detected a new task as a consequence of an already given graphical output. The new input is a demand for further information. The system has to stop the output of some information (e.g. the speech synthesis)

➔ the system has detected backchannelling. The planned output has to be shown.

➔ the system has to ignore the user, because he/she reads the output (Off–Talk [2]).

8 Conclusions

Dealing with multimodal systems like SmartKom the system has to cope with dialogue phenomena like turntaking and –holding.

In this paper we concentrated on linguistic input.

We defined situations where these phenomena are important for the processing of the system and gave examples of how the system should react properly.

The main goal of this paper was to show that there are – based on the data collection – lexical indicators for barge–in / before in SmartKom man–machine–dialogues, to classify them with respect to their word category, and to define the most significant ones.

Some strategies to handle the barge–in /barge–before phenomenon were presented.

It remains to be shown to what degree these barge–in / barge–before strategies improve the performance of the SmartKom system with regard to the users' acceptance and the overall technical innovation of the system.

An end–to–end evaluation of the internal system is under progress and will be presented soon.

9 Acknowledgements

This work was funded by the German Federal Ministry for Research and Technology (BMBF) in the framework of the SmartKom project (01IL905E/6).

0. REFERENCES

1. Sacks, H., Schegloff, E.A., Jefferson, G. (1974): Asimplest systematics for the organization of turn–taking in conversation. Language 50.4, 696–735
2. Rosenfeld, H.M. (1978): Conversational control functions of non–verbal behavior. In: Siegmann,A.W. und Feldstein, S.: Nonverbal behavior and Communication; Hillsdale, New Jersey (Bib: VII Sie 7,1)
3. Duncan,S.(1974): Some signals and rules for Taking speaking turns in conversation. In: Weitz, S.: Nonverbal Communication. Readings with a commentary; Oxford University Press, NY u.a.
4. Messing, L. und Campbell, R.(1999): Gesture, Speech and Sign; Oxford University Press
5. Goodwin, C. (1981): Conversational Organization. Interaction between Speakers and Hearers. Academic Press, NY
6. http://smartkom.dfki.de/index.html
7. N. Beringer, S. Burger, D. Oppermann(2000): Lexikon der Transliterationen. SmartKom Technisches Dokument 02–00.
8. S. Burger (1995): Transliterationslexikon. Verbmobil Technisches Dokument 36–95.
9. E. Kachelrieß (1999): Computerbezogene Sprache – Eine explorative Studie zur Untersuchung von Handlungsbegleitendem Sprechen in der Interaktion mit dem Computer. Verlag Dr. Kovac. Hamburg.
10. S. Steininger, Nicole Beringer, Daniela Oppermann (2001): Labeling von Gesten im Mensch–Maschine Dialog – Gesten – Kodierkonventionen SmartKom. SmartKom Technisches Dokument 14–01.
11. Daniela Oppermann (2000): OFF–TALK – Ein Problem für die Mensch–Maschine–Kommunikation? SmartKom Memo 04–00.
12. S. Steininger, F. Schiel, K. Louka (2001) : Gestures During Overlapping Speech in multimodal Human–Machine Dialogues. To appear in: International Workshop on Information Presentation and Natural Multimodal Dialogue, Verona, Italy, Dec. 2001