Gestures During Overlapping Speech in multimodal Human−Machine Dialogues

Silke Steininger, Florian Schiel, Katerina Louka

Institute of Phonetics and Speech Communication

Ludwig−Maximilians−University
80799 Munich
+49 89 2180 5751
[kstein][schiel][kalo]@phonetik.uni−muenchen.de

1 Abstract

A dialogue system has to deal with the problem of interruptions by the user, e.g. changes of requests (called »barge−in«). This contribution is concerned with this problem in the special case of the multimodal dialogue system SmartKom[1]. How are gestures used during such interruptions if they are utilized at all? To answer this question we analyzed a number of human−machine dialogues qualitatively. The analysis showed that most overlap situations were not accompanied by gestures at all. In the remaining instances the gestures were almost never "interactional" gestures, but mostly "unidentifiable" and "emotional" ones. We allocated the overlap situations accompanied by gestures to several subcategories of barge−in, pointed out the peculiarities of the gestures in the different cases and discussed their suitability as indicators for the dialogue system. Although from a small−scale in depth analysis no generalizations can be drawn, valuable insights for further investigation have been won. Most importantly it can be noted that the dynamic features of gestures seem far more promising as indicators of dialogue situations that need to be taken care of by the dialogue system than their static features.

1.1 Keywords

Multimodal Dialogue Systems, Gesture Input, Automatic Speech Processing, Barge−In, WOZ−Experiments.

2 Introduction

During a conversation it often happens that a speaker is interrupted by her listener. The speaker may stop talking or she may try to retain her turn and keep talking (for more information on inter human turn−taking see [1]). In most cases there will be a short timespan during which the speech of both participants overlaps. If one of the participants in the conversation is a dialogue system, these interruptions and periods of overlapping speech pose a special problem. Barge−in during human−machine interaction can be compared to turn−taking during human interaction [2], although the dialogue with an artificial assistant is presumably very different from a dialogue with a human [3], [4]. Gramkow [5] argues that turn−taking is not always "competitive and interruptive", but also "supportive and collaborative". We assume the same for barge−in (as reflected in the sub−categories we propose, see below). Therefore, for a system to be able to react properly to interruptions, it has to categorize them into those it should ignore and those it should attend to. If a reaction is called for, it has further to be decided what kind of reaction is appropriate. We are interested in the kind of barge−in situations that show up during dialogues between humans and a multimodal dialogue system. We want to know whether the analysis of the gestures may indicate the state of the user / her intention, especially whether she wants to make a request or not. To take a first step in answering this question we analyzed 78 human−machine dialogues that were recorded with the Wizard−of−Oz technique. After labeling the speech and gestures of the subjects, the speech annotations and the gesture label files were compared with regard to the periods where overlapping speech occurred. From a qualitative analysis of the data we derived descriptions of the gestures, speech and context for several categories of barge−in interactions. Peculiarities of the kind and usage of the gestures in the different categories are pointed out and can be used as starting point for further research. After defining what we understand as »barge−in«, we describe the data that served as basis for the analysis. We explain shortly how the data was labeled and analyzed. Finally we discuss the results and their possible practical implications for automatically detecting and categorizing barge−in.

3 Definition

We define as "barge−in" every user reaction (speech, gestures) during the system output (synthesis, display). Barge−in can be further divided into sub−categories[2]:

1.Abort: e.g. »stop!«; required action: Abort the processing/the presentation.

2.Premature request: e.g. »ok, this one«, before end of output; required action: Abort the presentation and fulfill the request.

3.Correction: e.g. »no, the right one«; required action: Change the request that is processed at the moment.

4.Successive request: e.g. »and another one too«; required action: Include the new information to the request that is processed at the moment .

[1] This research is being supported by the German Federal Ministry of Education and Research, grant no. 01 IL 905. We give our thanks to the SmartKom group of the Institute of Phonetics in Munich that provided the WOZ data. We especially thank the gesture−group: B. Lindemann, T. Paetzold, A. Glesner, P. Götz, C. Beiras−Cunqueiro, R. Schaller.

[2] Modified from a proposal of Tilman Becker, DFKI Saarbrücken, verbal communication.

5.Back−channeling: e.g. »ok«; No required action.

This contribution deals only with a subclass of barge−in, namely overlapping speech (verbal barge−in during speech synthesis). Other forms like verbal barge−in during graphical output or gesture barge−in during graphical output or speech synthesis are not considered. Additionally, the contribution deals with gestural indicators of overlapping speech[3].

4 The Data

4.1 The SmartKom Project

The analyzed data was collected for the SmartKom project[4]. The goal of this project is the development of an intelligent computer−user interface that allows almost natural communication with an adaptive and self−explanatory machine. The system does allow input in the form of spontaneous speech and in the form of gestures. Additionally the emotional state of the user is analyzed via prosody of speech and her facial expression. The output of the system comprises a graphic user interface (GUI) and synthesized speech. The GUI is realized as a computer screen that is projected onto a graph tablet. To explore how users interact with a machine, data is collected in so−called Wizard−of−Oz (WOZ) experiments: The subjects have to solve certain tasks with the help of the system (like planning a trip to the cinema). They are made believe that the system is already fully functional. Actually many functions are only simulated by two "wizards" that control the system from another room. In each WOZ−session spontaneous speech, facial expression and gestures of the subjects are recorded. For the gestures a digital camera is used which captures a side view of the subject (hip to head) and an infrared camera (SIVIT/Siemens) which captures the hand gestures (2−dimensional) in the plane of the graphical output. For the labeling these two streams are copied together with the beamer output of the display and a front view of the subject.

4.2 The Subjects and the task

The data analyzed consisted of 78 sessions of about 4.5 min length each. The voluntary, naive subjects were paid a small recompense. They were told that they had to test a new prototype of a dialogue system which could understand spoken language as well as gestures. It was not shown to them what sort of gestures, it was only pointed out that the system understood movements of the hand which were performed on or above the display. No subject reported or showed knowledge of the fact that the system was not real.

4.3 Coding Conventions for Gestures

Each gesture is identified with a *label*, that belongs to one of three superordinate *categories.* The label is complemented by several *modifiers* (e.g. reference word, reference zone). Each gesture is assigned to one of the three following categories: **I**nteractional gesture (I−gesture), s**U**pporting gesture (U−gesture)[5] and **R**esidual gesture (R−gesture) [6].

Interactional Gesture

The I−gesture is (possibly together with the verbal output) the means of the interaction with the computer. It can be a request, a confirmation or an answer. The following I−gestures exist:

•I−circle (+), I−circle (−): The circling of an object with (+) and without (−) touching the display.

•I−point (long +), I−point (long −): The pointing to an object for a longer duration (20+ frames) with (+) and without (−) touching the display.

•I−point (short +), I−point (short −): The pointing to an object for a short duration (up to 19 frames) with (+) and without (−) touching the display.

•I−free: All complex gestures above the display that signify a request like waving for »no« or »back«.

Supporting Gesture

A U−gesture occurs during the preparation of a request. It signifies the gestural support of a "solo−action" of the user (like reading or searching). The following U−gestures exist:

•U−continual (read), U−continual (count): The subject reads or counts and follows a line with the finger/hand.

•U−continual (search): A continual movement with the finger/hand. A request is clearly not made. The movement spans a large part of the display.

•U−continual (ponder): A continual movement with the finger/hand. A request is clearly not made. The movement takes place in only one reference zone.

•U−point (read), U−point (ponder): Like the respective continual gestures only that the movement is similar to a pointing gesture (hand/finger remains in one place).

Residual Gesture

This category subsumes all gestures that do not belong to one of the above categories. The few of the labeled gestures that take place outside of the space above the display belong to this category, too. A residual gesture does not prepare a request (at least not obviously) and is not a request or confirmation. A residual gesture is either an emotional gesture or an unidentifiable gesture. The following R−gestures exist:

•R−emotional (+ cubus), R−emotional (− cubus): A gesture that is connected to an emotional expression or to another interesting user state (e.g. pondering) of the subject on or over the display (+ cubus) or outside the room over the display (− cubus). Examples: Slapping the hand to the forehead, drumming with the fingers.

[3] With regard to lexical indicators of overlapping speech see Beringer, N.: “Possible Lexical Indicators for Barge−In / Barge−Before in a multimodal Man−Machine−Communication” at this workshop.

[4] http://smartkom.dfki.de/index.html

[5] We called the supporting gesture U−gesture for reasons of consistency with the German name "Unterstützende Geste".

•R−unidentifiable (+ cubus): Every movement that does not fit in the above categories.

A detailed description of the label−system can be found in [6].

4.4 Analysis

Since not much is known about gestures during barge−in situations we decided to first analyze a small number of cases in depth. From the corpus of 78 sessions we selected the sessions that showed a high number of overlapping events (eight or more). Nine sessions were eligible with regard to this criterion. The number and kind of the gestures in the nine sessions was analyzed. 86 overlap situations were found and compared with regard to the transliterations, the gesture coding, the audio and video streams. They were sorted into categories and the type and use of the gestures was noted. Essentially three broad categories were found: 1. Backchanneling, 2. Dissatisfaction/Helplessness, 3. Abort, Premature & Successive Request (see below).

In an second step the rest of the corpus was analyzed in a similar manner like the first nine sessions. The overlap situations with gestures were sorted into the three categories found in the first step. The overall picture was retained after the analysis of the full corpus, however, the overlap episodes with gestures were less frequent.

4.5 Results

Frequency of the overall gesture categories:

The total number of labeled gestures was 651. We decided to exclude a number of gestures because their classification was not possible without doubt[6].

430 interactional gestures (mostly pointing gestures)

121 supporting gestures (mostly pondering and searching)

100 emotional gestures

63 residual gestures (unidentifiable)

It can be assumed that the number of residual gestures normally would have been much higher. They make up about 25% of the corpus where the infrared video stream is not missing. The composition of the labeled gestures (without the residual gestures):

66,1% interactional gestures (mostly pointing gestures)

18,6% supporting gestures (mostly pondering and searching)

15,4% emotional gestures

Frequency of the gesture categories during speech overlaps:

Overall there were 222 overlap situations. In 25 sessions no overlaps occurred, in the other 53 sessions there were on average 9 overlaps (from 1 to 13).

In the nine sessions with the highest number of overlaps that were analyzed first, 41% of all overlaps were accompanied by a gesture. This figure lessened to 20% after analyzing all 78 sessions.

During co−occurring of speech overlap and a gesture[7] we found:

2% of the interactional gestures

14% of the supporting gestures

23% of the residual gestures (unidentifiable)

38% of the emotional gestures

Following is an example of the analysis for one subject:

Subject 77

1.U−continual (search):

SmartKom: diese Information ist momentan leider nicht verfügbar .

Sub77: [hey] , [@2ich @2muss (laughter) @2doch (laughter) @2wissen]

SmartKom: [kann@2 ich@2] Ihnen auch anders helfen ?

(SmartKom: This information is not available at the moment.

Sub77: hey , @2but @2I @2have @2to @2know

SmartKom: Can I help you with something else?)

Context: During »hey« the subject makes an R−emotional gesture. She is amused and confused and moves her hand palm up away from her (in a kind of exasperation gesture). The hey is not overlapped by synthesis.

U−continual (search): After this a searching gesture follows. The subject moves her hand fast over the display where a map is depicted that is the reference object of the conversation. She does not look there however but to the web persona, saying amused but reproachful to it [@2ich @2muss (laughter) @2doch (laughter) @2wissen]...

Barge In: The subject indicates that she is not content with the output. Her speech and gesture both show this protest/helplessness. We categorized this event as *dissatisfaction*.

2.R−unidentifiable

SmartKom: diese Information ist [momentan4@]

Sub77: [@4eah]

SmartKom: leider4@ nicht@4 verfügbar .

Sub77: @4auch @4nicht .

SmartKom: kann5@ ich5@ Ihnen5@ auch5@ anders helfen ?

Sub77: @5dann @5nützt @5mir @5aber @5doch @5der @5Plan @5nicht @5so @5viel .

(SmartKom: this information is [momentarily4@] unfortunately4@ not@4 available .

Subject: [@4ew] @4not @4also . @5but @5then @5after @5all @5the @5map @5does @5not @5help @5me @5a @5lot .

[6] In several sessions the data from the infrared camera was missing. This makes classification more difficult. We decided only to include cases that could be classificated without doubt. Most of the eliminated gestures were residual gestures.

[7] The percentages were calculated with respect to the frequency of the category, not with respect to the frequency of the gestures overall.

Context: The subject gets an unsatisfactory message and reacts with an emotional comment and makes a face (»ew«).

R−unidentifiable: During this she executes an unspecific movement with the right hand.

Barge In: This is a similar case as the first one. The subject shows her unhappiness with the output of the system (*dissatisfaction*). Differently from the first case is that the subject only makes a small movement that is not tied to the content of her speech (but the gesture is timed to the exclamation of dismay).

3.R−emotional −

SmartKom: diese Funktion ist momentan leider nicht verfügbar6@ .

Sub77: @6hm @6ja

SmartKom: [kann@6 ich Ihnen auch anders helfen ?]

Sub77: [@6okay] . Wie komme ich in das Kino ?

(SmartKom: This function unfortunately is not available6@ yet. Can@6 I help you with something else?

Sub77: @6okay . How do I get to the cinema ?)

Context: The subject gets an unsatisfactory message.

R−emotional: She reacts with an affirmation during which she grasps her chin pondering. The gesture to the chin ends when she begins her next question »How do I...«.

Barge In: *Backchanneling* − the system should not stop.

4.R−emotional −

SmartKom: diese [Information7@ ist momentan leider nicht verfügbar . Kann ich Ihnen auch anders helfen ? − pause]

Sub77: [@7ach . In welcher Richtung] ist das Kino ?

(SmartKom: this information7@ is momentarily not available. Can I help you with something else?

Sub77: @7sigh . In which direction is the cinema?)

R−emotional: The subject gets an unsatisfactory message and makes an appropriate exclamation. During this her hand moves to her face. She is obviously pondering. Halfway during the next question the hand moves away from the face.

Barge In: *Backchanneling* − although the subject reacts to a system output with an exclamation of unhappiness the system should not stop but go on.

Apart from the example the description of different typical speech overlap situations where gestures were used within the three categories cannot be given here in detail. Instead we explain the resulting categories and the typical characteristics of the gestures that showed up.

1. **Backchanneling**. The subject gives an affirmation to the output of the system, possibly thinking over her next step. Roughly 55% of all cases belong to this category. The system should not stop but go on. Examples are pondering gestures of the hand at the chin or unidentifiable gestures (changes in posture). The kinds of gestures that are used seem user−specific and therefore hard to exploit for automatic discrimination of satisfaction and dissatisfaction. It seems that they are less dynamic than the ones of other categories, but this needs further investigation. It is important to note that an affirmation need not be expressed in a positive mood, the user can be also be quite annoyed ("yes, yes, I know, go on"). This especially happens if the users interrupts synthesis output from the system that is no longer needed.
2. **Dissatisfaction/helplessness:** We included this category after the analysis of the first nine sessions. Only about 15% of all cases belong to this category. Here the subject is dissatisfied with the output of the system and expresses this verbally. It is desirable that the system is able to react to this kind of »helplessness« barge−in, perhaps by asking what is wrong or switching to a more guided dialogue behavior to solve the confusion. The gestures we found were diffuse and fast and occurred in an emotional context. We have too few examples to conclude much − but the assumption that gestures during dissatisfaction episodes tend to be more dynamic than during pondering or backchanneling episodes seems worth investigating further. Since interactional gestures are executed fast also, they would have to be distinguished from the »dissatisfaction indicators«. Actually, this could be possible, because the interactional gestures (at least the pointing gestures) show a distinctive ballistic acceleration curve. Supporting gestures seem to be characterized by a more variable acceleration curve. The emotional gestures of these episodes were mostly non−display−touching gestures, e.g. rubbing of the chin, rubbing of the nose.
3. **Abort, premature request, successive request:** The subject begins a request but changes it, aborts it or adds additional information. About 30% of all cases belong to this category. We put these three cases together because they all warrant the same reaction from the system: To include more information before presenting a result[8]. The gestures during these episodes seemed diverse on the first glance. But on a closer look they mimic the intent of the user: For example a pointing gesture was aborted if a request was reconsidered or fast unspecific movement was made at the point the user interrupted the system with another request. Beginning, end and changes of the gestures were closely tied to the semantic units of the speech. So, a sudden change of gesturing could be an indication of barge−in situations where the system should listen to further input before presenting its output.

5 Discussion

Before discussing the results a short note about the relevance of our analysis is in order. The data we analyzed was won with a WOZ−experiment. That means that the

[8] We include the abort here, because we think it is more useful not to stop the processing completely, but to ask for information.

response times of the simulated system will not be the same as the real ones, which, of course, can influence the occurrence of overlap situations. The output by the Wizards was not timed exactly, they were simply instructed to react not too fast to the user requests. This resulted in answering times between one and several seconds. The answering times were highly variable. From existing dialogue systems and the high complexity of the tasks, it can be assumed that the real system will be slower. A slow paced system with predictable answering times could lead to much less overlap than we found in the simulated data. Nevertheless, SmartKom aims to allow the user a natural dialogue. If this goal can be reached the amount and quality of overlap will probably be comparable to the amount and quality we found: If a dialogue is not turn based as in existing systems but is similar to a natural dialogue where dialogue partners compete for the turn chances are high that speech overlaps will be quite frequent. We therefore think that our results are especially relevant for systems that try to achieve an almost natural dialogue with the user.

To summarize our analysis and results: Realistic multimodal human−machine dialogues were analyzed in depth with regard to overlapping speech and synthesis. The question was, which (if any) gestures occur during barge−in situations[9]. The clearest result is that many barge−in situations are not accompanied by gestures. This probably does not result from an overall lack of use of gestures from the subjects because there were both subjects in the group that used many and that used few gestures. The composition of gestures that show up during a barge−in situation seemed to indicate that the kind of these gestures is not random: Only very few interactional gestures occurred. Supporting gestures were more frequent. On the other hand unidentifiable and emotional gestures seem to show up during an overlap situation more often. This makes sense: Interactional and supporting gestures probably are used in situations when the dialogue is productive and running smoothly. They are specific, goal−directed gestures whereas unidentifiable and emotional gestures are more diffuse gestures − possibly indicative of a more diffuse situation. Being more diffuse means also that there is a greater variation in the kind of gestures that are subsumed under the category of »unidentifiable« and »emotional«.

The small number of analyzed cases does not allow reliable generalizations. Nonetheless from the in depth study of barge−in situations interesting clues can be derived with regard to the features that seem promising to study further. We think three points can be made: First, mostly interactive gestures show up during a dialogue running smoothly. Second, diffuse gestures can be indicative of a barge−in situation. A multimodal system may use them as a warning to expect new or corrected input. Third, since the more diffuse gestures are the ones that show up during barge−in situations and because it is extremely difficult to describe the many variations of these gestures let alone to recognize them, it seems more practical to look into the dynamic features of the gestures, namely sudden changes in the use of gestures, their pacing, direction, velocity or the acceleration curves. Speech and nonverbal behavior are closely aligned, especially with regard to the dynamic features of the gestures [8], [9]. Therefore it can be hoped to exploit this fact for enhancing the performance of multimodal human−computer−interaction systems. There are many questions that are still open and seem worth studying further. Will the general trends of possible gestural indicators for barge−in remain the same after looking into more data? Are the barge−in situations without gestures similar or different to the ones with gestures? Can the interactional gestures really be used as an indicator for a smoothly running dialogue − and how can this be exploited for practical needs? During speech and synthesis overlap gestures seem to be rare. It would be interesting to see if they show up more often in the case of barge−in to display output.

•REFERENCES

1. Schegloff, E. A., "Overlapping talk and the organization of turn−taking for conversation", Language in Society, 29 (1), 2000.
2. Heins, R., Franzke, M., Durian, M., and Bayya, A., "Turn−taking as a design principle for barge−in in spoken language systems", International Journal of Speech Technology, 2 (2), 155−164, 1997.
3. Dahlbäck, N., Jönsson A., and Ahrenberg, L., "Wizard of Oz Studies − Why and How", Knowledge−Based Systems, 6 (4), p. 258−266, 1993.
4. Jönsson, A., and Dahlbäck, N., "Talking to a computer is not like talking to your best friend", Proc. of the First Scandinavian Conference on Artificial Intelligence, Tromso, Norway, p. 297− 307, 1988.
5. Gramkow, K., "Overlap management and interactional competence", Odense Working Papers in Language and Communication, 19 (2), 2000.
6. Steininger, S., Lindemann, B., and Paetzold, T., "Labeling of Gestures in SmartKom − The Coding System". To appear in Proc. of the Gesture Workshop, London, 2001.
7. Oppermann, D., Schiel, F., Steininger, S., Beringer, N., "Off−Talk − A Problem for Human−Machine−Interaction?". Proc. of Eurospeech , Scandinavia, Aalborg, 2001.
8. McNeill, D., Hand and Mind: What Gestures Reveal about Thought, University of Chicago Press, Chicago, 1992.
9. Wachsmuth, I., "Kommunikative Rhythmen in Gestik und Sprache", Kognitionswissenschaft, 8 (4), 151−159, 2000.

[9] Related to the barge−in problem is the general problem of user comments that are no requests (called "off−talk"). See [7]