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

SmartKom
Mensch-Technik-Interaktion

How to relate User Satisfaction and System Performance in Multimodal Dialogue Situations? - a Graphical Approach

Nicole Beringer
Sebastian Hans
Katerina Louka
Jie Tang

Ludwig-Maximilians-Universitt
Mnchen

Report Nr. 24
Juni 2002

Juni 2002

Nicole Beringer
Sebastian Hans
Katerina Louka
Jie Tang

Ludwig Maximilans Universitt Mnchen
Schellingstr. 3
80799 Mnchen

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

E-Mail: beringer,hanss,kalo,tang@phonetik.uni-muenchen.de

Dieses Technische Dokument gehrt zu Teilprojekt 1: Modalittsspezifische Analysatoren

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

Inhaltsverzeichnis:

How to relate User Satisfaction and System Performance in Multimodal Dialogue Situations? - a Graphical Approach

Nicole Beringer, Sebastian Hans, Katerina Louka, Jie Tang
Institut für Phonetik und Sprachliche Kommunikation
Ludwig-Maximilians-Universität München
Schellingstr. 3, 80799 München
{beringer, hanss, kalo, tang}@phonetik.uni-muenchen.de

Abstract

The goal of this paper is to present in detail a graphical evaluation tool for multimodal dialogue systems which is used to compare the users' satisfaction with the system's technical performance quasi objectively. It is also used to define weights for the calculation of overall system performance (user satisfaction & technical performance) of multimodal dialogue systems.

1 Keywords

Multimodal dialogue systems, end-to-end evaluation, graphical evaluation tool, evaluation framework, SmartKom

2 Introduction

When evaluating multimodal dialogue systems the evaluators have many problems to solve and only partly can transfer established methods from spoken dialogue system evaluations (see (Beringer et al., 2002b) for further details) such as the PARADISE framework (Walker et al., 1997). In the end-to-end evaluation of the multimodal dialogue system SmartKom the evaluators have to deal with the innovative character of multimodality. Therefore, we developed a new evaluation framework for multimodal dialogue systems, PROMISE (**Pro**cedure for **M**ultimodal **I**nteractive **S**ystem **E**valuation) (Beringer et al., 2002a) since established methods cannot be transferred unambiguously from monomodal frameworks like PARADISE.

This new framework combines established methods from spoken dialogue evaluations and takes into account new methods to handle multimodal characteristics like gestural input combined with speech input, graphical vs. speech output or userstate information via facial expression of the user.

One of the features of spoken dialogue evaluation frameworks, namely the independence of systems and tasks by weighting objectively measured qualities and quantities of the system by subjective user satisfaction has been implemented in PROMISE as well.

To obtain weights, user satisfaction can be directly correlated with the objective quality and quantity measures (further referred to as costs) via Pearson correlation (see (Beringer et al., 2002a) for further details). But not all costs can be given corresponding questions in usability questionnaires. However, to obtain a normalization over differing systems, scenarios and tasks we have to weight them somehow.

To handle this problem, the evaluators have to objectively compare with costs the recorded passages of the dialogues in question.

This was the motivation to develop a graphical approach to check and relate user satisfaction and system performance objectively.

The paper is structured as follows: section 3 describes briefly the function of the multimodal SmartKom dialogue system which has to be evaluated. In section 4 we give a general outline of PROMISE. Section 5 describes the possibility to define weights to normalize over systems, scenarios and tasks. The requirements and characteristics of the graphical evaluation tool as well as some positive side effects will be presented in section 5. The paper finishes with a short summary and outlines our future work.

3 The SmartKom project

In the SmartKom project, an intelligent computer-user interface is being developed which deals with various kinds of oral or physical input. Potential benefits of SmartKom include the ease of use and the naturalness of the man-machine interaction due to multimodal input and output. However, a very critical obstacle to progress in this area is the lack of a general methodology for evaluating and comparing the performance of the three possible scenarios provided by SmartKom:

- SmartKom Home/Office to communicate and operate machines at home (e.g. TV, workstation, radio),
- SmartKom Public to have a public access to public services, and
- 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. One of the requirements of the project is to develop new modalities and new techniques.

4 General Outline of the PROMISE Framework

PROMISE (**Pro**cedure for **M**ultimodal **I**nteractive **S**ystem **E**valuation) is an extended evaluation framework for multimodal dialogue systems (Beringer et al., 2002a), where we aimed to solve the problems of scoring multimodal inputs and outputs, weighting the different recognition modalities and how to deal with non-directed task definitions and the resulting, potentially uncompleted tasks by the users.

Advantages of established methods like abstracting from systems, scenarios and tasks are included in PROMISE as well as modality specific measures. The latter, of course, is the basis of a number of problems. The most challenging is described in the following subsection.

4.1 Scoring multimodal inputs and outputs

In contrast to interactive monomodal spoken dialogue systems, multimodal dialogue systems consist of several equivalent technologies which are functionally similar to each other. In other words, multimodal dialogue systems are based on many component technologies like speech recognition, gesture recognition, recognition of emotional states, text-to-speech, natural language understanding, natural language generation, generation of graphical presentation, synchronization of speech and graphics and database query languages. Taking the example of recognition, the different modalities can and will interfere with each other.

To evaluate interfering functionalities multimodal inputs have to be identified and further processed on according to the "total or nothing" principle[1].

Another problem in scoring multimodal input is how to estimate the accuracy of different recognizers. I.e., in talking about speech recognition, we have to deal with a very complicated pattern match, whereas gesture recognition has a limited set of recognizible gestures which can be found in a given coordinate plane.

4.2 Dealing with non-directed task definitions

Apart from the scoring problems PROMISE also offers a solution for handling non-directed task definitions. In contrast to task requirements, where the user has to check several functions of a system in a defined order, SmartKom offers a variety of different functions which can be combined in any order to get the wanted information. Therefore we had to define more dynamic "keys" (a PARADISE

[1] Only those combined inputs get access to the evaluation which are not ambiguous in their content.

term) to extract different superordinate concepts depending on the task at hand referred to below as "information chunks" or "bits" . These "information chunks" are carefully selected, categorized and weighted by hand before the tests start to compute, normalize and compare across different tasks and scenarios.

The number of information chunks can vary within one completed task, but it must define a task unambiguously in order to complete it successfully.

4.3 Abstraction from systems, scenarios and tasks

The requirement of abstracting from systems, scenarios and tasks is assumed to be necessary not only for spoken dialogue systems evaluation but also for the evaluation of multimodal dialogue systems. This can be done by weighting successfully completed tasks with correlation coefficient of the Pearson correlation between user satisfaction values and task completion. Due to the more dynamic task definition in SmartKom, PROMISE allows only two values for task success:
$\tau_j = +\ 1$: task success;
$\tau_j = -\ 1$: task failure;
where j is the index of the corresponding tests.

To abstract from dialogues, PROMISE uses the mean value $\bar{\tau}$.

To compute the system performance we have to normalize over the cost functions via a z-scored normalization function: $\mathcal{N}(c_i) = \frac{c_i - \bar{c_i}}{\sigma_{c_i}}$, where c_i i-th cost, σ_{c_i} variance of c_i, $\bar{c_i}$ the mean of c_i.

5 The SmartKom Graphical Evaluation Tool

5.1 Requirements

For the SmartKom evaluation we designed a graphical evaluation tool which gives the possibility to compare user satisfaction values (out of a usability questionnaire) about a given functionality with the corresponding quality and quantity measures (objectively measurable technical evaluation) of the respective dialogue. A human evaluator checks both parts and decides which of the two is more likely or if both are equally likely. To get a better idea of the dialogue (s)he has to evaluate, the video or parts of it can be displayed as well - either by choosing the start and endpoint by hand or by clicking offered timestamps (see section 5.4 below). For evaluators' comments we must also provide a text field. The quasi objective scoring is written in an SQL database described below in section 5.3.

Table 1 gives an overview of the cost and usability pairs we defined for SmartKom.

The tool offers a defined course of evaluation results. This balances individual differences both of the evaluators and of the users. Making the data available to a number of evaluators by offering a platform independent tool the evaluation can be done highly objectively.

5.2 Video display

For the linux version of our Graphical Evaluation Tool we had to provide a video player which allows starting, stopping and repeating parts of the filmed dialogues. Video playback is implemented using a specially hacked up copy of XAnim[2]. XAnim is embedded into the interface by means of two wrappers:
embedded_xanim and XAnimRemoteControl.

5.2.1 Modifiactions to XAnim

XAnim had to be modified to allow viewing of single parts of a video and to allow rapid stopping of replay. It now sports four new command line options, to specify start and end times in frames or milliseconds.

5.2.2 embedded_xanim

embedded_xanim is an interactive interface for XAnim written in C.
It displays a window with the specified size at the specified position. It then goes into a loop reading lines from standard input which are interpreted as XAnim command line arguments and an optional window title. XAnim is run with these arguments, thus playing the desired video. Furthermore, the video can be stopped at any point during replay using the 'stop' command.

[2]XAnim 2.80.1, see
http://xanim.va.pubnix.com/home.html

Quality and quantity measures	usability question
Transaction success Task complexity	The task was easy to solve
Misunderstanding of input Offtalk	SmartKom has understood my input
Misunderstanding of output	SmartKom can easily be understood
Semantical, syntactical correctness Incremental compatibility	SmartKom has answered properly in most cases
Mean system response time Mean user response time Timeout	The speed of the system was acceptable for each situation I always knew what to say
Acc. gesture recognition	The gestural input was successful
Acc. ASR	The speech input was successful
Dialogue complexity	SmartKom worked as assumed SmartKom reacted quickly to my input SmartKom is easy to handle
Percentage of appropriate/inappropriate system directive diagnostic utterances Percentage of explicit recovery answers repetitions No. of ambiguities	SmartKom offered an adequate amount of high quality information SmartKom is easy to handle
Diagnostic error messages Rejections Timeout	SmartKom needs input only once to successfully complete a task
Help-analyzer	SmartKom offers adequate help
Output complexity (display)	The display is clearly designed
Mean elapsed time Task completion time Dialogue elapsed time	SmartKom reacted fast to my input
Duration of speech input Duration of ASR	SmartKom reacted fast to speech input
Duration of gestural input Duration of gesture recognition	SmartKom reacted fast to gestural input
BargeIn Cancel	SmartKom allows interrupts
Dialogue complexity	Was the task difficult?
Gesture turns Ways of interaction	input via graphical display
Display turns	output via graphical display
Speech input	speech input
Speech synthesis (synchronicity)	speech output
N-way communication Ways of interaction Error rate of questions Input complexity	Possibility to interact in a quasi-human way with SmartKom

Table continues next page

Quality and quantity measures	usability question
Recognition/duration of facial expression Prosodic features	SmartKom reacted towards my emotional state
Synchronicity Graphical output (turns) Cooperativity	How do you score the competence of the agent? Were actions of the persona natural?
Gestural input	Gestural input

Table 1: Quality and quantity measures for the SmartKom evaluation compared with user satisfaction values

5.2.3 XAnimRemoteControl

XAnimRemoteControl is a high-level Java wrapper-class for embedded_xanim.
During initialization, an instance of embedded_xanim is started. Then, arbitrary video-files can be played with the play()-method, and replay can be stopped with the stop()-method. The quit()-method cleans up and kills embedded_xanim.

5.3 The Back-end - A MySQL Database

The data processed by the tool comes from and goes into a MySQL database.
The questions and answers of the technical evaluation (mainly extracted from the logfiles) are read in from the TE_Frage & TE_Antwort table pair[3]. The results of the ergonomic evaluation (from the questionnaire) are read from a slightly more complicated arrangement of tables, where the questions are split into one or more subquestions. The technical and ergonomic questions are connected by a link table. The result is written into a separate synthesis table.

5.4 The Evaluation GUI

The evaluation process is based on a GUI programm written in Java. It is deliberately selected for its character of platform independence, so the program can be easily executed on the different computer types in the institute.

[3]Name of the cost and the value

5.4.1 Functional description

The graphical interface is composed of 3 sections: a Java window for the evaluation, a XAnim window[4] for the video display and an emacs window for the label display. Figure 1 shows a screenshot of the GUI.

The Java window is divided into 3 boxes of the same size. The left box is displayed for the result of technical evaluation, and the data is imported from the related MySQL database, in which the results of different sessions are located. The right one shows a display for questions and answers of ergonomic evalution related to the technical part. The middle section offers pre-defined decision categories. Comments can be written in plain text in the comment field.
When the GUI is started with a session, an aligned XAnim window with the video file in questions as well as an aligned emacs window including the corresponding annotations is initialized.

5.4.2 XAnim control

During the evaluation process, human evaluators can navigate through the video file by clicking the timestamps in order to see a sepecified part of the video sequence. It is integrated into the Java program, so the evaluator can mark the start and end point in milliseconds to fire an event, which is handled correspondingly in the programm to run the embedded_xanim with the specified segment.

5.4.3 Database connection

Another main component of this tool, which runs in the background, is the database

[4]see the section 5.2 Video display

connection. The database used in the evaluation process is a MySQL database, which includes the essential tables for the whole evaluation(See the section 5.3 for details). The program communicates with the database in two ways:

- In the initialization process of the GUI, the program makes queries to the database to extract the related results of technical and ergonomic evaluation in order to graphically represent them in the provided boxes.
- During the evalution process the evaluator makes decisions and comments. The outgoing results are then immediately inserted into the corresponding table of the database.

5.5 Side Effects

Apart from allocating user satisfaction and system performance via the evaluation GUI the tool offers some positive side effects, namely

- the possibility of annotating userstate and gestural input (Steininger et al., 2001; Steininger et al., 2002)
- the possibility to score multimodal recognition facilities
- the repetition of video sequences

Using a graphical interface not only for evaluation purposes but also for presentation of detailled evaluation results we make it possible for every scenario that the developers get access to all content-related problems, results and protocols within the corresponding evaluation phase. While using the display, they can easily find out by checking the related objectively measured and user satisfaction values on which evaluation part to concentrate (user satisfaction scores or objective costs) to improve the system.

6 Conclusion

The Graphical Evaluation Tool allows a visual representation of dialogues that can be used both for evaluating the system and comparing user satisfaction with objectively measured costs. Human evaluators have to follow the same course of evaluation by the Graphical Evaluation Tool. This balances individual differences both between the evaluators as well as the users.

Apart from the primary functionality of evaluation, our tool offers the possibility of annotating user state and gestural input.

Via the GUI, it is possible to score the different recognition modalities as well.

Finally, the controlled playing of video sequences can be done platform independently due to the Java implementation.

7 Acknowledgements

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

References

Nicole Beringer, Ute Kartal, Katerina Louka, Florian Schiel, and Uli Tuerk. 2002a. Promise - a procedure for multimodal interactive system evaluation. *Proc. of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation (to appear).*

Nicole Beringer, Katerina Louka, Victoria Penide-Lopez, and Uli Tuerk. 2002b. End-to-end evaluation of multimodal dialogue systems - can we transfer established methods? *Proc. of the LREC 2002, Gran Canaria, Spain (to appear).*

S. Steininger, B. Lindemann, and T. Paetzold. 2001. Labeling of gestures in smartkom - the coding system. *Springer Gesture Workshop 2001, London (to appear)*, LNAI 2298.

S. Steininger, S. Rabold, O. Dioubina, and F. Schiel. 2002. Development of the user-state conventions for the multimodal corpus in smartkom. *Proc. of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation (to appear).*

M.A. Walker, D.J. Litman, C.A. Kamm, and A. Abella. 1997. Paradise: A framework for evaluation spoken dialogue agents. *Annnual Meeting of the Association of Computational Linguistics. ACL.*

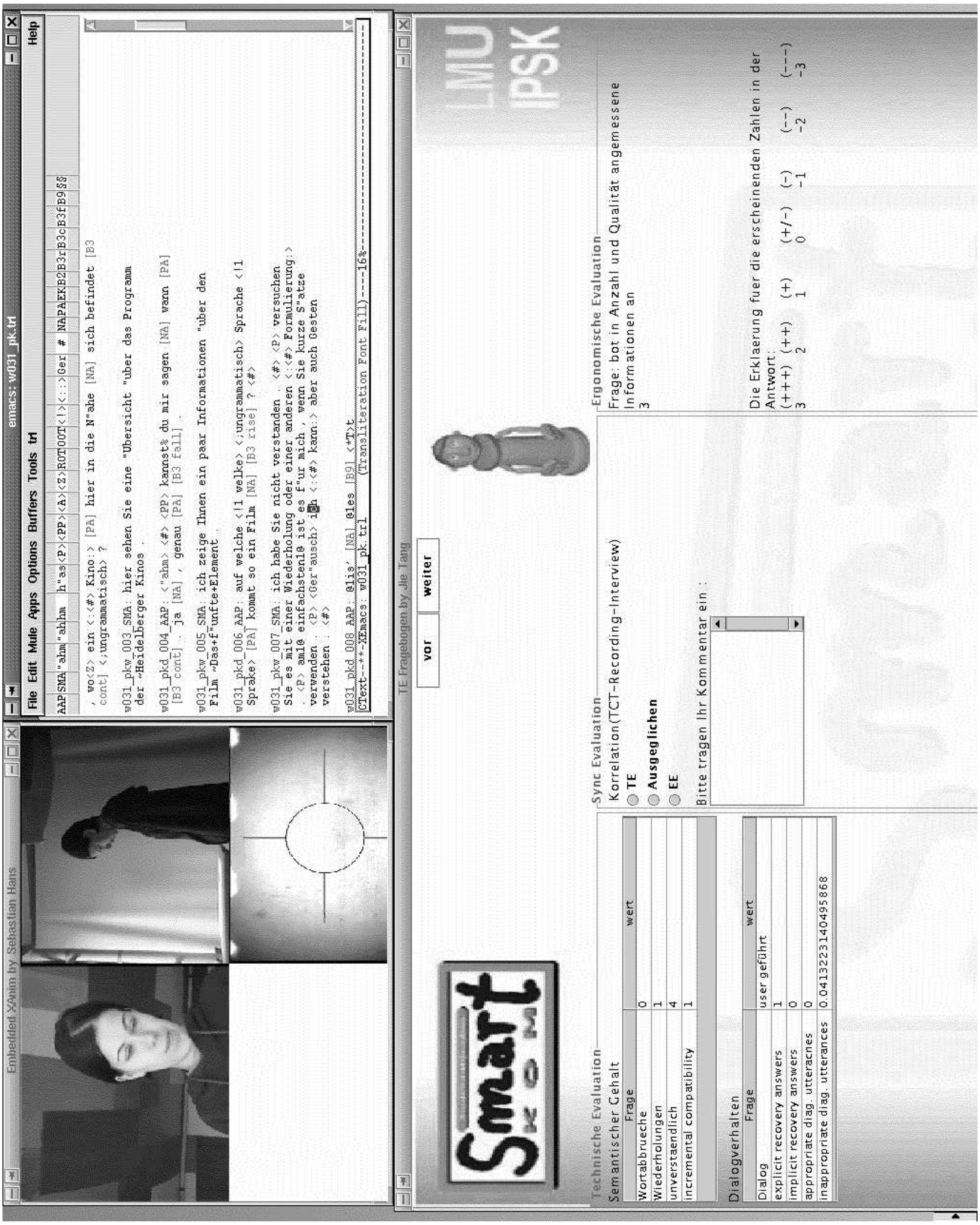

Figure 1: Graphical evaluation tool for the SmartKom evaluation