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

# PROMISE - A Procedure for Multimodal Interactive System Evaluation

**Nicole Beringer**
**Ute Kartal**
**Katerina Louka**
**Florian Schiel**
**Uli Türk**

**Ludwig-Maximilians-Universität München**

Report Nr. 23
Mai 2002

.

Mai 2002

Nicole Beringer
Ute Kartal
Katerina Louka
Florian Schiel
Uli Türk

Ludwig Maximilans Universität München
Schellingstr. 3
80799 München

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

E-Mail: beringer,ukartal,kalo,schiel,tuerk@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:

# PROMISE - A Procedure for Multimodal Interactive System Evaluation

**Nicole Beringer[1], Ute Kartal[1], Katerina Louka[1], Florian Schiel[2], Uli Türk[1]**

[1] Institut für Phonetik und Sprachliche Kommunikation,
[2] Bavarian Archive for Speech Signals (BAS)
Schellingstr. 3, D-80799 München, Germany
{beringer,ukartal,kalo,schiel,tuerk}@phonetik.uni-muenchen.de

**Abstract**

This paper describes a general framework for evaluating and comparing the performance of multimodal dialogue systems: PROMISE (**Pro**cedure for **M**ultimodal **I**nteractive **S**ystem **E**valuation). PROMISE is a possible extention to multimodality of the PARADISE framework ((Walker and al., 1998; Walker et al., 1997) used for the evaluation of spoken dialogue systems), where we aimed to solve the problems of scoring multimodal inputs and outputs, weighting the different recognition modalities and of how to deal with non-directed task definitions and the resulting, potentially uncompleted tasks by the users.

PROMISE is used in the end-to-end-evaluation of the SmartKom project - in which an intelligent computer-user interface that deals with various kinds of oral and physical input is being developed. The aim of SmartKom is to allow a natural form of communication within man-machine interaction.

## 1. Introduction

The aim of this paper is to give an extended framework on dialogue system evaluation for multimodal systems in the end-to-end evaluation of SmartKom.

In the SmartKom project, an intelligent computer-user interface is being developed which deals with various kinds of oral and physical input. The system allows and reacts to gestures as well as mimic and spoken dialogue. Potential benefits of SmartKom include the ease of use and the naturalness of the man-machine interaction which are 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 allows to communicate and to operate machines at home (e.g. TV, workstation, radio),
- SmartKom Public provides public access to public services, and
- SmartKom Mobile

Because of the innovative character of the project, new methods for end-to-end evaluation had to be developed partly through transferring established criteria from the evaluation of spoken dialogue systems (PARADISE), and partly through the definition of new multimodal measures. These criteria have to deal with a fundamental property of multimodal dialogue systems, namely the high variability of the input and output modalities with which the system has to cope. As an example, the system can accept a task solution not only via three different input devices (mimic-camera, voice-recording, gesture-camera), but also many different long-term solution strategies. For example, in order to plan an evening watching TV, a subject using the system may start with a sender (or a time, or an actress, etc.), progress to give a time (or sender, or actress, etc.) and end choosing a single programme (or a series, or many programmes, or none, etc.). This inherent complexity and the tasks to be completed make it necessary to find an evaluation strategy which measures up to the possibilities of such a flexible system. Earlier evaluation strategies (PARADISE) had to deal with systems with only one input modality, used given solution strategies and thus were easily able to measure the success of a task. The advancements of SmartKom, though, cannot be adequately measured nor evaluated using an evaluation-strategy based on a monomodal system with pre-given solution strategies.

The following section gives an overview of standard problems of dialogue system evaluation which principly can be solved by the PARADISE framework (Walker and al., 1998; Walker et al., 1997). Section three describes how to define a task and extract the attribute value keys out of the description - solving a problem not uniquely belonging to multimodal dialogue evaluation. How we deal with incomplete tasks or tasks that get a very low task success measure due to incooperativity of the user, is described in section four. The scoring of multimodal inputs and outputs can be found in section four. Sections five to six give a detailed description of the status of multiple-to-one input facilities, i.e. the possibility to express the same user intention via multiple input as well as via different input modalities. Section seven defines the approach of PROMISE as a multimodal dialogue evaluation strategy which normalizes over dialogue strategy and dialogue systems. In the last section we sum up some ideas to be implemented in our framework.

## 2. Standard problems of dialogue system evaluation

Of course, multimodal dialogue evaluation has to deal with the same problems spoken dialogue system evaluation has to deal with, namely

- How can we abstract from the system itself, i.e. the different hardware and software components, in order to evaluate across dialogues and scenarios (see above)?

- How can we abstract from different dialogue strategies?

The PARADISE framework (for detailed description please refer to (Walker and al., 1998; Walker et al., 1997)) gives a useful and promising approach of how to compare different spoken dialogue strategies and different spoken dialogue systems via attribute value matrices (AVM), to compute the (normalized) task success measure (provided that a clearly defined task description is given to the user) define several (normalized) quality and quantity measures so-called cost functions , and to weigh their importance for the performance of the system via multiple linear regression dependent on the User Satisfaction value (cumulative function on the questionnaire completed by the subjects). This is not practicable, though, when dealing with a multimodal system like SmartKom**.**
Unfortunately, in dealing with multimodal systems we find a number of components which do not fit into the PARADISE approach, which are:

- The user is given a rather unprecise task definition, in order to enable a mostly natural interaction of user and system. Therefore there exist no static definitions of the "keys" (a PARADISE term) necessary to compute an AVM. Our solution is to extract different superordinate concepts depending on the task at hand. For example, when planning an evening watching TV, these superordinate concepts - we call them "information bits" - would contain movie title, genre, channel, timeslots, actors etc. Similar to a content-analysis, these "information bits" are carefully selected, categorized and weighted by hand before the tests start. This makes it possible for us to compute, normalize and compare across different tasks and scenarios.

- The number of information bits can vary within one completed task, but they must define a task unambiguously in order to finish it completely and successfully. For example, when a user asks to explicitly view a movie by name, assuming this movie is broadcasted at a set time and in only one channel, the number of information bits necessary to complete the task is one (the name of the movie). Whereas if a user doesn't know exactly what show he wants to watch, the number of information bits necessary to complete the task of planning an evening watching TV must be at least two (for example, time and channel).

In contrast to computing the task success via AVMs like PARADISE, in which case not completed tasks could implicitely influence the results, our information-bit-approach ensures that task success can only be calculated if the task has been completed.

## 3. How to deal with a bad performance due to user incooperativity?

One of the main problems of dialogue systems is an incooperative user. We consider only those users to be truly incooperative, who fall out of the role or purposely misuse the system. As an example, a user reading a book to the system or using his mobile phone will be classified as an "incooperative" user. These, of course, are not unique to multimodal system evaluation, but can occur in other situations as well. On a first cue, it is impossible to incorporate incooperative users in an evaluation without lowering task success and thus the system performance. To avoid this, there exist the following approaches:

- Only dialogues with cooperative users are evaluated using empirical methods

- Only dialogues which terminate with finished tasks are evaluated.

Both approaches will be used in conjunction, so that a clearly defined set of data can be evaluated. When deciding to follow the first idea, uncooperative users as defined above, are of course interesting for other than purely empirical reasons. Evaluating the data generated by these incooperative users, in the above sense, in order to improve the system for future development, though, is not part of our aim to judge the quality of the present state of the system SmartKom.

## 4. How to score multimodal inputs or outputs?

In contrast to interactive unimodal spoken dialogue systems, which are based on many component technologies like speech recognition, text-to-speech, natural language understanding, natural language generation and database query languages, multimodal dialogue systems consist of several such technologies which are functionally similar to each other and therefore could interfere with each other. To make this clear, just imagine the similar functions of ASR and Gesture Recognition: while interacting with a multimodal man-machine interactive system like SmartKom users have the possibility to say what information they want to have and to simultaneously give the same, an additional, or a more specific input by an "interactional gesture" (Steininger et al., 2001). There are several possible problem solving strategies for the system namely:

- First match: the information which was recognized first is taken for further system processing, regardless of the recognition method. This would of course not help in multimodal processing.

- "Mean" match: the system takes the information which is common to both of the recognition modules. This could be called multimodal verification.

- Additional match: take all the information given by several recognizers for further system processing. This would be the best solution, if we assume all recognizers to be highly accurate, which leads us to the next problem:

## 5. How to weight the several multimodal components of recognition systems?

How can we estimate the accuracy of different recognizers? For example, in talking about speech recognition, we have to deal with a very complicated pattern match, whereas gesture recognition has a limited set of recognizable gestures which can be found in a given coordinate plane.

It should be clear, that

- the gesture recognizer will be more accurate than the ASR system but
- the ASR system must get a higher weight than the gesture recognition when evaluating the system, since the complexity of the gesture recognition is much lower than the complexity of the ASR sytem.
- Apart from the problems of how to weight the different multimodal system components in an end-to-end evaluation of a multimodal system, there is also the problem of synchrony:
- Are multimodal inputs synchronous or linear within the evaluation? Are inputs from different modalities synchronous, i.e. are they describing the same user intention, although they may not be synchronous in time? Or does the system have to cope with different inputs?

## 6. PROMISE - A Procedure for Multimodal Interactive System Evaluation

In the last sections we have identified the most characteristic problems which show the need for an extended framework for multimodal dialogue system evaluation. We already gave some examples of possible problem solving strategies. Within this section we will specify these ideas and present the current version of PROMISE. Given the normalized performance function of PARADISE

$$\text{performance} = \alpha\,\mathcal{N}(\kappa) \;-\; \sum_{i=1}^{n} \omega_i\,\mathcal{N}(c_i)$$

with $\alpha$ the weight for task success[1] $\kappa$, the assumed Gaussian cost functions [2] $c_i$ weighted by $\omega_i$, and $\mathcal{N}$ z-score normalization function. Weights are defined via a linear multiple regression over $\kappa$ respectively the costs and the cumulative sum of the user satisfaction scores (see usability questionnaire (Walker and al., 1998)). PROMISE now splits this function in two parts in the way that the formula is reduced to normalized cost functions first. Instead of a multiple linear regression between the free cost variables and the dependent user satisfaction score, PROMISE searches correlations via Pearson correlation between User-Satisfaction - Cost pairs. This means that objective measurable costs will be addressed in the questionnaire to be answered by each user.

Tables 1 and 2 give an overview of the costs we defined in SmartKom, some of them equivalent to the PARADISE costs, some of them extended to deal with multimodality or to weed out user incooperativity.

[1]defined as the successful completion of a duty

[2]either mean or cumulative sum of one cost category (quantity and quality measures); differing from system to system

| Quality measures | |
|---|---|
| system-cooperativity | measure of accepting misleading input |
| semantics | no.of multiple input possible misunderstandings of input/output semantical correctness of input/output |
| helps | no. of offered help for the actual interaction situation |
| recognition | speech facial expression gestures |
| transaction success | no. of completed sub-tasks |
| diagnostic error messages | percentage of error prompts |
| dialogue complexity | task complexity (needed information bits for one task) input complexity (used information bits) dialogue manager complexity (presentation of results) |
| ways of interaction | gestures/graphics vs.speech n-way communication (several modalities possible at the same time?) |
| synchrony | graphical and speech output |
| user/system turns | mixed initiative dialogue management incremental compatibility |

Table 1: Quality measures for the SmartKom evaluation

Apart from measuring the quality of dialogue systems in general, like dialogue management (system directed, user directed or mixed initiative), elapsed time of input and output, task completion, mean response time, word count and turn count, we defined measures referring to the problems in section four and five above. Multiple inputs and outputs are scored via the "semantics" cost, to be precised the number of multiple input and the possible misunderstandings of input/output (due to multimodality).

The multimodal components of recognition systems are partly weighted via Pearson correlation using the corresponding user satisfaction scores for the recognition costs defined above. Comparing the accuracies of the different recognition systems for defining a cross-recognizer weight, we calculate "ways of interaction" and "helps". The latter defines the quality and quantity of dynamic help offered by the system in situations where the emotional status of the user changes.

[3](Oppermann et al., 2001)

| Quantity measures | |
|---|---|
| barge-in | no. of user and system overlap by means of backchanelling, negation of output, further information |
| cancels | planned system interrupts due to barge-in |
| off-talk [3] | no. of non-system directed user utterances |
| elapsed time | duration of input of the facial expression<br>duration of gestural input<br>duration of speech input<br>duration of ASR<br>duration of gesture recognition<br>mean system response time<br>mean user response time<br>task completion<br>duration of the dialogue |
| rejections | error frequency of input which require a repetition by the user |
| timeout | error rate of output<br>error rate of input |
| user/system turns | no. of turns<br>no. of spoken words<br>no. of produced gestures<br>percentage of appropriate/ inappropriate system directive diagnostic utterances<br>percentage of explicit recovery answers |

Table 2: Quantity measures for the SmartKom evaluation

The second step is to define another way to calculate the task success.

In PARADISE a set of defined static "keys" was used to measure task success via an attribute value matrix. Since "information bits" (see section two) are used, it makes no sense to calculate an AVM. As described above, these information bits can vary from situation to situation. A successful task is given, if the task was completed according to the necessary number of information bits. A task fails, if it has not been successfully completed. Therefore, we define task success in PROMISE as follows:

$\tau_j = +1$ : task success; $\tau_j = -1$ : task failure; where j is the index of the corresponding tests.

For each test the corresponding user satisfaction values are Pearson correlated with $\tau_j$.

The system performance results in the following formula:

$$\text{performance} = \alpha\bar{\tau} - \sum_{i=1}^{n} \omega_i \mathcal{N}(c_i)$$

with $\alpha$ being the Pearson correlation between $\tau_j$(task success) and the corresponding user satisfaction values,
$\bar{\tau}$ the mean value of all $\tau_j$ with j index of tests,
n the number of different cost functions,
$c_i$ the assumed Gaussian cost functions - consistently either mean or cumulative sum of one cost category i (measured over all tests)
weighted by $\omega_i$ - the Pearson correlation between cost function $c_i$ and defined associated user satisfaction values, and the z-score normalization function $\mathcal{N}(c_i) = \frac{c_i - \bar{c_i}}{\sigma_{c_i}}$, where $\sigma_{c_i}$ as variance of $c_i$.

## 7. Conclusion and future work

Our aim was to roughly define an extended evaluation framework for a multimodal dialogue system evaluation which can deal with multimodal dialogue processing. However, there are still some unresolved or solely unsatisfactorily solved problems dealing with the timing of input in multimodal systems. We are currently specifying different approaches in order to satisfactorily solve the remaining problems which we hope to present at the LREC post-conference workshop on "Multimodal Resources and Multimodal Systems Evaluation" in June.

## 8. Acknowledgements

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

## 9. References

D. Oppermann, F. Schiel, S. Steininger, and N. Beringer. 2001. Off-talk - a problem for human-machine-interaction. *Proc. of EUROSPEECH 2001, Scandinavia, Aalborg*.

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.

M.A. Walker and al. 1998. Evaluating spoken dialogue agents with paradise: Two case studies. *Computer Speech and Language*, 12.

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.*