Challenges and Important Aspects in Planning and Performing Evaluation Studies for Multimodal Dialogue Systems

Susanne Höllerer

ftw. Telecommunications Research Center Vienna
Tech Gate Vienna
Donau-City-Straße 1/ 2nd floor
1220 Vienna
Austria
hoellerer@ftw.at

Abstract

In this paper I want to discuss the problems researchers face in trying to plan and carry out an evaluation study for multimodal systems – particularly in qualifying the purpose of the testing, defining the intended user group for their application, arranging the testing setting and aligning the evaluation plan. It is my intention to show which aspects should be taken into account and which basic standards should be fulfilled. Furthermore, I provide two sections about points to consider in performing the study as well as in the analysis of the received data. Then I describe possible difficulties concerning the evaluation of (multimodal) systems and try to sketch longer term solutions. Finally, I list possible options on how to utilize the results of evaluations studies in further research.

1. Introduction

Multimodal dialog systems should be efficient, easy to handle and comprehensible for intended users – so how should the evaluation of such dialogue systems be designed and carried out in order to accomplish these goals and how can the outcome and the conclusions of the concerned studies be used for further research and development?

How can researchers and developers of dialogue systems answer the needs and preferences of the users, how can they accommodate their special interests and characteristics?

And how can problematic issues researchers and developers face today be tackled and solved? To which extent can experiences made until now help to find solutions for the future?

These questions need to be answered already in the very beginning of the whole development process – before the start of the planning and mental development of the system one has to designate the goals of the system and which functions it serves. At the same time the target group has to be defined – this can on the one hand be a small group of experts and for a special field of application or on the other hand the entire population, depending on the system or object. During the development process these facts must be taken into account in order to produce the most efficient system for the special target group. To this end, it is useful to perform an iterative mode of evaluation which means that for every important phase of development an evaluation study is provided so as to find out about the direction the development of the systems leads to and to make sure that the intended users are able to handle it. Especially as far as multimodal dialogue systems are concerned, evaluation studies are a relevant part of the development and – at the same time – a challenging task. The particular difficulty is to provide methods for logging and analysing two or more different modalities and to test each of them separately as well as combined with the other(s). This means that the developers and researchers receive much data, which require experience and reliable methods to be analysed.

Because of the innovative design and handling of those systems a careful evaluation planning has to be provided.

I wrote this paper from a social scientific point of view – as a different perspective concerning the preparation and the procedure of evaluating a system. Social and empiric science can provide information on methodological issues, questions concerning the analysis of the data, the selection of test subjects, the arrangement of the setting and the formulation of the specific tasks for the test persons.

2. Goal of the paper

The main intention of this paper is to show how important a careful planning and performance of an evaluation study – concerning especially multimodal dialogue systems – is. It should be made clear which features of the testing process are particularly relevant and which problems may appear and which challenges the evaluation of a system, possessing more than one modality for handling, may involve. Furthermore, in this paper important aspects which may appear negligible at the first sight should be mentioned, for example the range of persons who are going to use this system in the future, ergo the intended user group: what are their characteristics, their needs and how can the system serve them? For the designer and the researcher, this means also knowing exactly the functions of the system. Another aspect would be the setting in which the testing should take place: how should it be arranged and which role plays the tester?

A very important part of this paper is the one about how the results of the evaluation research in general and the experiences of each researcher can contribute to the further research done in these fields and consequently to establishing standards for the design and the performance of evaluation studies of multimodal systems.

I also like to state my point of view concerning the present as well as the longer term problems researcher might face in developing and evaluating multimodal dialogue systems, and also how they might be avoided.

3. What is the image of the intended (average) user and how does it affect the development of the evaluation of a system?

As I mentioned before, theoretically the entire population can be the target group of a certain system or object, for instance of information extracting systems like automatic telephone enquiry for train schedules. As far as IT systems are concerned, in the last years it was often assumed implicitly that the circle of intended users is a rather small one (compared to the one of objects of everyday life) and is composed of experts of fields like computer technology and science, managers or other academic job-holders in hierarchical higher positions.

But today such systems should serve everybody. It is the developers´ duty to design the system in such a way that it can also be conceived and used by non-experts. That concerns especially the presentation of the graphical user interface, which the user gets the first impression of before even having tried out how to handle the application.

So if one has in mind that the target group may be as heterogeneous as the general population there is no possibility to postulate any specific knowledge or experience concerning multimodal dialogue systems among all persons. This means that the researcher has to begin at the very start and make the use of the system as easy as possible. That is for sure a very challenging task – and an important one, because the design and the usability of the system are important factors for its acceptance among the intended users. One has to consider that persons of every age group, sex, society position and socialisation background may use this system. The sample the researcher assorts should be representative in so far that each of these parameters is taken into account. One possibility to find out about those features is to provide a user questionnaire.

One option is to search test persons of a certain age, sex and education level. The last parameter is useful to get some information about the position in society they bear. Another way would be to consider income or field of profession, or respectively the job they are working in. It is hard to find out much about the economic or social background of the test persons without violating their privacy. And one must not believe that one statistical feature gives information about a person's standard of living. So this parameter is a rather hard one to obtain. Nevertheless it should be included in the evaluation.

The aspect of age is also an important one, because one may find big differences between younger and older people concerning their competence as well as their experience with modern technological instruments and systems – a phenomenon which does not apply universally. But there may be the tendency that older persons are more sceptical and reserved if not afraid to serve as test persons for evaluations of such systems. They often argue that they need not be taken into account, because they are too old – which is of course a misbelief.

It is common to consider sex as a variable, too, because it is interesting to view possible differences between women and men in handling technological systems and to react to them in the further development of the system.

4. Recommended standards for multimodal dialogue systems

It seems to be of use to establish basic standards that need to be fulfilled in order to provide a system appropriate for a great range of users. This is particularly relevant for multimodal dialogue systems, which provide several ways of handling and therefore require extraordinary user-friendliness. The standards described in this chapter can be seen as provisional and extensible – they should serve as basic points of orientation.

These standards make clear which direction further research should take and on which aspects it should focus, but also which main issues any evaluations study should focus on.

4.1. Easy intelligibility of the functions and applicability of the system

In order to be able to use the system in an efficient way the users have to understand which actions one can perform and which goals one can accomplish with it. That is to say that the instruction manual must be clear and specific. But also the design of the user interface should give a clue to how to use the system.

4.2. Distinct visual design of the graphical user interface

The user interface, i.e. the part of the system the user sees and interacts with, should not be complex, but the various elements should be arranged clearly and distinguishably. Concerning this feature, knowledge from fields like psychology or the specific domain of advertisement could be advantageous, but the cooperation between these fields and the one of IT is not that strong yet.

4.3. Good intelligibility of the commands

The language in which the user communicates with the system is usually a set of commands – either given via speech or via GUI. And vice versa the systems gives commands or poses questions to the user – often in the form of spoken prompts. It is necessary to formulate these in a simple and intelligible way so that the user is able to catch it.

4.4. Good speech recognition

A dialogue system that provides the modality of handling via speech needs to have an excellent speech recognizer. That is a prerequisite for efficiency, which is an overall goal of such systems. This means that it should also work in noisy environment, as the system should be adaptable in awkward situations where the user cannot have regard of a clear articulation. Unfortunately – although there has been much research done in this area – it takes a long time to develop a good recognizer respectively it is hard to find a recognizer appropriate for the functions a system should fulfil

4.5. Efficiency as well as smooth performance of the actions

Multimodal dialogue systems have the special aim to work smoothly even in difficult or stressful situations, for instance if the user needs both hands for other actions. It would be very exhausting for the user to be forced to

repeat the commands or questions a several times because of the slow processing or the long upload time of the system.

4.6. Clear, intelligible output (speech-output as well as output via the GUI)

In order to provide a smooth process and a good information extraction respectively an optimum support the output the system delivers should be correct as well as intelligible.

5. Advantages of multimodal dialogue systems

The advantages listed in this chapter are supposed to supplement or – in part – condition each other. This list should – on the one hand – emphasize the differences between single- and multimodal dialogue systems and – on the other hand – show which anticipations researcher have concerning these kind of systems.

To reach the intended users as well as to make the system interesting for them, one has to emphasize its advantages in achieving a certain goal, possibly by comparing it to other kinds of systems or – in general – ways to reach this goal (for instance using a multimodal dialogue device to extract information about the surroundings instead of a simple map).

One big use of such a multimodal dialogue system is for sure the *flexibility*. The overall goal of the development of multimodal applications is for the users to interact with the system the way they like to – depending on the situation they are in. For example when driving in his car, the user cannot use his hands to operate the system – so there has to be one or more other ways to handle it in order to fulfil the claims of efficiency and usability. In this case, the modality of handling via speech input is an optimum alternative.

The optimum situation would be that every user was free to interact with the system the way the situation requires it – and to alternate the one modality with the other(s) in a spontaneous way. The system should therefore be designed to react and adapt to this user-specific behaviour. This demands – in case of a multimodal dialogue system – an excellent speech recognition as well as a synoptic user interface quick to apprehend.

Beside flexibility higher *efficiency* is another advantage of multimodal dialogue systems – provided that sufficient evaluation studies has been performed in order to find out about how a system needs to be designed to serve the users well. It´s clear that efficiency – at least in part – grows proportionally with flexibility (and the other way round), so these two aspects are connected tightly.

A third advantage which may be of great importance for "everyday users" is the *individuality and personality* a system gets when becoming multimodal, hence being able to be integrated smoothly in ones everyday life and supporting the performance of certain actions.

The great use of multimodal dialogue systems in comparison to other systems is the fact that they combine the advantages of the single modalities they include, this means that the user can profit from the advantages of handling via the GUI as well as via speech. In detail, this would be promptness as far as the modality of speech is concerned – action can be executed far more faster by speaking the commands that by typing them. The other advantage which is already known is the possibility to keep ones hands free for other actions which is, for instance, very important while driving the car. Regarding the modality of handling via the GUI the main advantage lies in the privacy of the commands the user is giving and of the actions the system is executing. While speech can be received by persons around the user, actions like typing are not audible.

Disadvantages of one modality might be eluded by using the other modality – for instance if the speech recognizer does not work properly.

6. Important items in planning and carrying out an evaluation

To evaluate a system one has to know exactly which functions it possesses and who the intended users are (cf. Nielsen 1993: 170). The evaluation study is performed to serve the purpose of finding out more about how the system should be designed in order to answer the needs and interests of the users. As I mentioned in the introduction the best way to carry out an exhaustive evaluation study is to perform several smaller "steps of evaluation". This means that – depending on the development stage of the system – the respective properties, the design and the effect on the users need to be measured.

And for each of these steps some important points must be considered. To receive sufficient and eligible data for the analysis afterwards, the evaluation study needs to be planned carefully, tasks for the test persons to perform must be formulated – which are supposed to accomplish the intentions the researchers have –, methods to log the process of evaluation need to be found as well as methods to capture the impressions and experiences of the test persons. The choice of these instruments depends on which aspects of the tests are important for the developers on the one hand, and – on the other hand – how easily the requested information can be extracted. A good way to find out which methods are appropriate for the evaluation study is to evaluate the logging methods themselves. That is also useful to assure that the methods one uses really measure what they are pretending to measure – hence if they are suitable for what the respective developer wants to find out. A good method for logging the evaluation process is to use instruments like audio recorder, video camera, mouse tracker, screen logger or eye tracker. But one must be aware that receiving too much data out of an evaluation study can be as well a problem as receiving too little.

Not only the methods and the technological equipment are to be considered – the whole setting of the testing process needs to be planned. The role of the tester who stays with the test persons must be defined – mostly he is the one who explains the aim of the testing as well as the specific tasks and who observes the test person during the performance. This raises some questions: How much information should the test person be given in order to not affect the authenticity of the situation and the (possible) impartiality of the user? Should the tester answer questions during the testing? Where should he place himself? To which extent should he adapt himself to the test person (concerning behaviour, speech, …) to provide a more informal setting or how can he prevent himself

from doing so? Are there differences in the performance of the test persons depending on the sex, the age or the credibility of the tester?

As far as the test persons are concerned, should they be given some time to get to know the system better (some minutes without logging or even observing) or should they be tested from the very beginning?

And how should the testing setting look like to provide as much authenticity as possible?

All these questions can become problematic when too little time and know-how is spent on the preparations of the evaluation studies – the difficult issues are explained in detail in chapter 7.

7. Analysis methods

It is important to find appropriate analysis methods as well, for instance annotating schemes to analyse spoken language and synchronize it with actions like mouse movements or clicks. That is a good and rather objective possibility of spotting the problems the users had performing the tasks, but also the points where the test persons apparently used the system in an efficient way, for example – concerning multimodal dialogue systems – combined speech and handling via the GUI. Especially for large numbers of test persons and hence a lot of data such standardized analysing methods are useful. However, the range of good and reliable annotating schemes is not that great. The few that are made use of in empirical studies fall short of easy applicability and efficient programming. Much remains to be done in this field of research. Also the methods themselves need to be tested to find out if they work the way the researcher wants them to. If the methods fail, the whole study needs to be repeated.

However, also the subjective impressions of the test persons are important for the analysis, so one should not surrender a questionnaire, an individual interview or informal talk with the test persons after the testing. These data need to be analysed either quantitatively – in the case of a questionnaire – and presented in statistics or analysed in a qualitative way, that is to collect the test persons' impressions and statements and to detect positive or negative tendencies.

But not in every case all the errors of a system can be detected: one cannot be sure that all the problems could actually be recovered – "one troubling aspect of testing is the uncertainty that remains even after exhaustive testing by multiple methods". [Shneiderman 1998: 125]

8. Problems to be solved concerning the process of developing and testing a new system

There is a range of problems researchers of multimodal dialogue systems have to face during the process of developing and optimizing the system. In some ways, preparing the evaluation study for multimodal dialogue systems does not differ from preparing one for "singlemodal" systems. Just a few aspects are more challenging as far as multimodal systems are concerned.

8.1. Defining the user group and test subjects

First of all, it is difficult to find out about the intended user group: how should the researchers know which persons the system will be used by? And how can they be sure that the users they design the system for are really the ones who will use the system in the end? A step towards finding a solution to this problem would be to carry out a survey among the supposed target group or among the whole population to get a clue about who is interested in the product and may benefit from it.

The range of test persons should be representative for the group of intended users, that is to say that the test subjects should represent the properties of the target group. If the system was designed primarily for elder persons, it is recommended to choose such persons for the evaluation study. In this regard the question must be raised where one should find appropriate test subjects. There are several possibilities:

One may look for persons in public institutions or buildings like schools, universities or on the street. An alternative would be the search for people by an advertisement. Or one may get access to a range of test persons by buying (or exchanging) subjects databases.

8.2. The discrepancy between researcher and user

Another difficulty in the process of developing a (multimodal dialogue) system is the discrepancy between researcher/developer and "normal" user or test subject. The researcher who designs the system is an expert in this field, he/she possesses knowledge and experiences concerning this specific system and knows how to handle it – so one can assume that he/she is the person appropriate for testing the system. That is true – to some extent. The persons, who understand the functions and operations of the system best, may also know how to measure and optimize them. The problem, which may occur, is that the researcher knows the system to well. This means that he/she is not able to put him/herself in the situation of the non-expert user and, therefore, blind out all his/her knowledge. One may argue that just because of these problems evaluation studies are carried out. That is correct. But it is not enough to perform one or more evaluation studies, it has to be guaranteed that the study is performed in a correct way, this means to really find out about the user group and its needs and expectations. The researchers are – in some way – preoccupied. So they do not seem to suit for planning such a study. A possibility to avoid this problem would be to separate the role of the researcher and the one of the evaluation designer strictly. But here another difficulty appears: how can the evaluation designer know enough about the system to understand its functions and features and at the same time know not too much about it in order to stay as objective as possible?

8.3. The evaluation setting

In order to get valid testing results, not only the tasks to fulfill need to be chosen carefully, also the setting where the testing should take place has to be planned regardfully.

The easiest possibility is to perform the tests within an isolated laboratory or at least in the rooms of the company which developed the system. This would mean that the testing situation could be controlled rather easily and that no unexpected disturbances would happen. These apparent advantages entail one negative aspect. Choosing such a testing setting would mean that the authenticity of the situation would be in peril. Especially as far as

applications are concerned which are not designed to be used at home or in a quiet and private place, testing within the circumstances mentioned above would not represent the conditions which the user has to face when using the application in reality. The researcher cannot foresee all the different situations in which the system may be applied but he knows the intended user group and the functions of the application and therefore can assume how it is going to be used.

Portable devices for example are supposed to be applied on the way, for instance on the streets, in public buildings and institutions, while walking or traveling by car, at different events or in likewise noisy environment. The noise must not be underestimated – as well as other factors, for instance when information is required as quick as possible (train departure times for example). A system, which works perfectly within the laboratory setting, might turn out to fail when being used in real surroundings. How should these settings be imitated in the laboratory to gain valid results?

As a matter of course, one must in this case consider the development phase of the system. If there is not an application to be carried around yet it can hardly be tested like if there was. An iterative kind of evaluation study requires several different testing settings.

8.4. Methods

Finding appropriate methods for logging the testing process might be a problem as well. The choice depends on which modalities the system has, as there are several options for each of them to be logged and measured. Most multimodal dialogue systems offer at least the two following modalities: the speech-modality and the handling via the GUI.

In order to log spoken user-output, one could for instance use a simple recorder with a microphone or a camera which could also tape visual impressions like the gesture and the face of the test subject as well as the monitor of the computer or the display of the application device (if it is not too small) – depending on which kind of system is tested. At the same time, the output of the system should also be taped for to liaise the both kinds of output in order to get information about the quality of the speech recognition and the smoothness of the whole process.

It is not as simple to find methods – beside the camera – to trace the operations on the monitor or the display, ergo the handling via the GUI. There exist some software tools like screen logger or key tracker which log the mouse movements or clicks as well as the input via the keyboard or the selection via the menu. Unfortunately, the existing software is either very expensive or only available for companies of specific fields.

In addition, there are other tools to log the process in order to gain more information about the handling of the system – for instance a so-called eye-tracker that logs the eye movements of the user. Through its analysis one may find out about which elements of the GUI are bold and how easy or difficult it is for the user to understand how to operate the system.

The challenging aspect concerning multimodal systems is to connect the methods used for logging the handling of different modalities. One kind of information needs to be related with another. The speech signals must be synchronized with the manual actions, for instance. This intention requires another software or program like an annotation scheme.

8.5. Possible solutions and recommendations for the future

As I said before, it is necessary to involve several persons in the developing and the testing process of the system, as more perspectives are required for an effective evaluation study. Concretely, this means that persons from several fields of research should work together, the tasks should be distributed and the roles the persons occupy within this process should be defined well. The researcher, the developer, the designer, the market research institute, the tester, several university institutes like psychology, sociology and computer science – all of these persons and institutions have competences in their specific fields and can contribute to producing a good working system. Through exchanging experience and know-how, as many difficulties as possible might be avoided.

In my view, this strategy will play an important role in the future, for aspects like user friendliness and acceptance of the system by the users are more and more coming into prominence. It is not any longer the group of IT experts and business people only who need applications of new technologies, but “average persons” from every part of the society.

Nowadays the number of those companies increases which specialize on evaluation studies and tests on usability – a fact that indicates the prominence of these aspects.

While IT companies spent most time on producing new systems and optimising new technologies, the aspect of user friendliness was rather neglected. The big chance to catch up on these experiences is the cooperation with persons of other fields or companies; to get support at finding the right test subjects, equipment and methods.

9. Evaluation outcomes as resources for further research

First of all, the outcome of the evaluation studies serves the improvement and the development of the evaluated system. But the received data are not useless after completing the evaluation process. The lessons one draws out of this testing can be used for other – similar – studies. On the one hand developers get to know the logging and analysing methods better, on the other hand they learn more about this kind of systems in general and how the intended users manage them – this knowledge can be made use of in further research.

To mention the economical aspect, the results of an evaluation study can of course also be used in cooperation with other technological enterprises or research centers with commercial as well as scientific interests; they can be exchanged or sold.

This procedure does not need to be restricted to similar, ergo technological, fields of research – instead the knowledge can also be connected to different fields like psychological or sociological research or the particular field of advertisement, where methods of usability and analyses of effects on consumers and users have long tradition. Knowledge from these disciplines can be used for evaluation studies and – vice versa – the results of these kinds of studies can be made use of in other fields.

10. Conclusions

The challenges in designing evaluation studies for multimodal dialogue systems are plain to perceive: in contrast to dialogue systems using one modality, evaluating multimodal systems demands more than one perspective of testing and hence just as many methods of logging. To use the received data for the improvement of the system and for further research it has to be processed with the support of suitable analysis methods – the particular challenge at this is to find an appropriate method for each modality and each kind of data.

Another aspect is the definition of the purpose as well as the intended users, and as a main task the designing of the user interface and the systems functions in order to meet the interests and needs of the target group.

The field of research of multimodal dialogue systems and applications is relatively new and few standards concerning the design of the user interface or the ways of testing and analysing have been established. But today usability studies are attached more importance than ever – for every kind of system or object, not only in fields of technology. There are – on the contrary – branches that deal frequently with aspects of usability and already gained precious information, for instance (cognitive) psychology. This knowledge can be useful for enterprises or persons who develop such multimodal systems. In my opinion, the cooperation with other companies or even other fields of research and hence the exchange of experiences and know-how is one big chance to improve usability testing, even for very specific applications.

Acknowledgements:
This work was supported within the Austrian competence center program ***K****plu*s, and by the companies Kapsch, Mobilkom Austria and Siemens.

11. References

Nielsen, Jakob (1993): Usability Engineering. Academic Press. San Diego.

Shneiderman, Ben (1998): Designing the User Interface. Strategies for Effective Human-Computer Interaction. Addison-Wesley. Reading, Massachusetts.