Project ref. no.	FP6-507609
Project acronym	SIMILAR
Deliverable status	R
Contractual date of delivery	30 November 2005
Actual date of delivery	December 2005
Deliverable number	D69
Deliverable title	An exploration of selected challenges in usability evaluation of multimodal and natural interactive systems
Nature	Report
Status & version	Final
Number of pages	88
WP contributing to the deliverable	SIG7
WP / Task responsible	NISLab
Editor	Laila Dybkjær
Author(s) (alphabetic order)	Niels Ole Bernsen, Pedro Correa, Hans Dybkjær, Laila Dybkjær, Enrique J. Gómez, Thomas Hansen, Svend Kiilerich, Pablo Lamata, Benoit Macq, Torben Kruchov Madsen, Luciana P. Nedel, Giorgos Nikolakis, Fabio Paternò, Angela Piruzza, Samuel Rodríguez, Carmen Santoro, Daniela Trevisan, Dimitrios Tzovaras, Jean Vanderdonckt
EC Project Officer	Mats Ljungqvist
Keywords	Usability, evaluation, challenges, tools, methods, criteria
Abstract (for dissemination)	This deliverable provides a collection of chapters describing various challenges and ongoing work on usability and evaluation within the areas of spoken multimodal dialogue systems, vision-based systems, haptics-based systems, mixed-reality systems in surgery, and tools for remote usability evaluation.

Deliverable D69

An Exploration of Selected Challenges in Usability Evaluation of Multimodal and Natural Interactive Systems

Alterface SA, Belgium
GBT, Polytechnic University of Madrid, Spain
ISTI-CNR, HIIS Laboratory, Pisa, Italy
ITI-CERTH, Greece
NISLab, University of Southern Denmark
Tele, Université catholique de Louvain, Belgium

December 2005

Contents

1 Introduction

This deliverable takes up the thread from the previous SIG7 deliverable D18 entitled "Current practice description and evaluation: Towards a framework for usability evaluation" and further explores some of the challenges in usability evaluation discussed in D18.

The present deliverable deviates from the original plan in that it does not present "results of iterative framework testing" which was the planned title and contents of the report. It has eventually become clear that it is not realistic to achieve the original goal of SIG7, i.e. to come up with a draft usability evaluation best practice for multimodal and natural interactive systems, given the resources available in terms of manpower and expertise. Already D18 contained the first signs showing that this was not realistic. D18 provided, as planned, an overall description and evaluation of current practice in usability evaluation of multimodal and natural interactive systems within the areas of expertise of the partners, i.e. spoken multimodal dialogue systems, vision-based systems, haptics-based systems, mixed-reality systems in surgery, and tools for remote usability evaluation. However, the work towards a framework was not so much done in terms of establishing a very first framework, as it was done in terms of pointing out challenges in usability evaluation which need further investigation within the particular areas. With no first framework established, the achievement of the original goal of D69 became impossible since there was no framework to test. Moreover, a general framework – even just one spanning the areas of expertise of the SIG7 partners – would be difficult to establish, if not impossible. While usability evaluation methods are general and applicable across areas, the criteria to use and the need for tools are much more specialised.

In relation to the above, it has become clear that a significant number of the SIG7 partners are mainly in SIG7 to learn something about usability evaluation rather than because they are usability experts. It has therefore been decided to revise the original plans and work on a usability guide for dummies as the next goal. This guide could be established and tested as a collaborative effort among the SIG7 sites and would also allow for participation by other parties.

The present report does not describe collaborative SIG7 work but rather reflects what is going on concerning usability and evaluation at each of the SIG7 sites. It gives an impression of some of the challenges people are addressing within the partners' areas of expertise. In the following, we provide an overview of each chapter of the report.

Chapter 2 describes results of the first field test of the pronunciation trainer, carried out at a number of schools and other institutions teaching Danish for immigrants. The chapter focuses on results of the students' use of the system for training Danish single word/phrase pronunciation. The pronunciation trainer system has been developed by NISLab who has also advised the institutions on how to install and use the system, dispatch to NISLab of logfiles from the students' use of the system, etc.

88 students from 9 institutions trained with the system for periods ranging from a couple of hours to 4 months, producing a total of 821 logfiles. Logfile filtering resulted in identification of 22 students from 4 institutions whose learning progress could be meaningfully analysed based on the available data. The data analysis shows promising results on student progress.

Chapter 3 pinpoints the need for tools in support of usability and evaluation. Not least for industry this is crucial to keep costs down while adding more usability. Focus is on spoken interaction. The chapter reports on a tool called DialogDesigner that is being used for commercial design, specification and evaluation, and that is in the process of being further

expanded. The tool supports rapid development of an electronic system model. Once an electronic system model is available, it is possible to support the further development and evaluation process in various ways as the chapter shows. In the described version of the tool one option is to view a graphical presentation of the dialogue model which can be made more or less. A second option is to run a Wizard of Oz simulation. This part of the tool can also be used for walkthrough evaluation and it can be used as part of presentations to and discussions with customers. Simulations or walkthroughs are logged and can be analysed later. The simulation log can also be used normatively to generate test scripts. A third option is to extract HTML versions, e.g. of the entire interaction model, which may be used for presentations to customers.

The focus of Chapter 4 is on vision-based systems. The chapter points out that standard usability evaluation methods still apply also when the evaluation concerns novel systems. It stresses that the involvement of users in the test process is crucial but also that it requires a great deal of care. Three methods involving users are described as being particularly useful for the evaluation of vision-based systems, i.e. think-aloud, contextual inquiry, and interviews. As examples of useful evaluation criteria are mentioned learning curve, fatigue, friendliness, effectiveness and efficiency.

Chapter 5 takes its point of departure in deliverable D18 in which a number of weaknesses in usability evaluation criteria used for haptic-based systems were identified, and an evaluation procedure was proposed which involves iterative usability testing with users at different stages of system development. The chapter describes the usability evaluation of a cane simulation application following the procedure described in D18. A haptic glove providing force feedback was involved in the application. Gloves were evaluated already before application development. During development the application was evaluated with representative users in order to discover weaknesses and errors and to obtain input on how to improve the application. The test is described and so is the analysis of results. As expected the test provided important input to the developers.

Chapter 6 deals with usability evaluation of surgical simulators. Specifically focus is on a comparison of a set of laparoscopic simulators that use several resources of virtual reality technology to meet different training needs. Basically these resources are classified into the three categories of fidelity (levels of realism), virtual resources (e.g. instructions to guide a task) and evaluation resources (e.g. performance and progress). The simulators have been compared along these three dimensions but results have not yet been compared with respect to training outcome and thus no estimate of the efficiency of the systems is available yet.

Chapter 7 also studies surgical simulators. The main objective of the presented study is to identify a theoretical and practical basis that explains how mixed reality interfaces may support and add value to interactive applications. Focus is on a method for evaluating virtual reality applications. The method has been tested on a mixed reality maxillofacial surgery system. The idea is to use the test results to find a model that will enable the identification of the contributing factor of each of the analysed variables in the user interaction. The maxillofacial surgery system used for the test is presented. Then it is explained how errors inherent from the computer system and devices used were calculated (accuracy measure). Finally, the usability testing (including scenarios considered, hypotheses and variables) and the actual experiment (including task descriptions, pre-tests, subjects, and logging) are described. The method that will be used for evaluating the interaction is briefly mentioned but no results from carrying out the test are presented.

Chapter 8 describes a tool for remote usability evaluation of web sites and illustrates its use. The tool uses information from logfiles, videos, and eye-tracker data. Based on a model of

how optimal completion of various tasks would be carried out, the tool automatically evaluates the actual interaction. The model may be seen as the expected user behaviour and if the actual behaviour deviates, the information provided by the analysis tool will make aware of it. The analysis results include among other things a record of tasks not completed, a list of errors occurring during the performance of tasks, time for completing a task, and information about the user's behaviour during task performance. Based on this data evaluators may identify problematic parts of the web site and try to improve them.

2 Evaluation of Danish Pronunciation Trainer

First Evaluation based on Field Data collected at Training Sites December 2004 – primo May 2005

Niels Ole Bernsen, Torben K. Madsen, Thomas K. Hansen and Svend Kiilerich
Natural Interactive Systems Laboratory
University of Southern Denmark
Campusvej 55, 5230 Odense M

2.1 Summary and Overview

This report describes results of the first field test of the pronunciation trainer, carried out at a number of schools and other institutions teaching Danish for immigrants. The report focuses on results of the students' use of the system for training Danish single word/phrase pronunciation. The pronunciation trainer system has been developed by NISLab who has also advised the institutions on how to install and use the system, dispatch to NISLab of log files from the students' use of the system, etc.

2.1.1 Results

The first field test of the pronunciation trainer has generated a large (+800 log files) and valuable data corpus on the practical use of the system at the training sites. Based on this corpus, it is possible to make a first assessment of the system's usefulness for improving Danish pronunciation of single words and phrases in students learning Danish as a second language. Four schools in the Odense area have contributed the data analyzed in the present report. Another five schools, outside of the Odense area, have received the system for installation but have either not contributed data at all or have contributed data, which is too sparse for analytic purposes.

We can observe that, in general, all students whose progress is analyzed in this report have done serious and dedicated pronunciation training. This shows that the system actually works in practice in the field, which makes it probable that the pronunciation trainer actually supports the students' motivation to train with the system.

We also observe that, as a whole, the student body has done insufficient training for providing us with the ideal outcome of an evaluation of the system's use and usefulness at the training sites. On average, the students have done 1/4 of the full training programme of 4500 word pronunciations. By comparison, the ideal outcome just mentioned would be a sufficient number of students who had completed the full training programme. Since, generally speaking, all training sites were up and running by 1 February 2005, and since log file collection was terminated around 1 May 2005, our training data represent at least 12 weeks of data from the training sites that have delivered data that can actually be used for evaluation purposes. However, many students stopped their training after considerably less than 12 weeks. Since we do not have direct access to the students, we must find other means of intensifying the students' training efforts.

Until now, the use, usability and usefulness of the pronunciation trainer have only been analyzed in the NISLab laboratory environment with, among others, Chinese and Finnish students, balelining with native Danish speakers, etc. However, laboratory testing is quite

different from field-testing of the system, and it is only the latter which can provide us with realistic information about the real applicability of the system.

The completed field test demonstrates that the students make good progress with the system. On average, the 16 students in the main group who are not already good Danish speakers, have achieved a minimum of 16.4% improvement on their start score as measured in absolute percentage points, and with an average training effort of about 20% of the full recommended effort. These averages cover very large individual differences in progress, implying that many students have made considerably larger progress than shown by the average of 16.4%. We regard it as a conservative estimate that the average student progress, based on a considerably larger training effort than has been the case so far and one which approximated the full recommended training programme – will attain 40% in absolute percentage points above their start scores. Thus, for instance, a student who scores 30% pronunciation correctness at the start will be able to score 70% by the end of the pronunciation training and under the same test conditions that obtained at training start..

The phenomenon of *test conditions* represents a discovery done during the field test of the pronunciation trainer. The phenomenon is discussed extensively below, including the need for proactive measures, such as user manual and system revisions, which we believe follows from the phenomenon.

It is our opinion that the results to be presented provide a somewhat convincing basis for claiming that the pronunciation trainer is a useful tool for self-training of Danish pronunciation by people wanting to learn Danish as a second language. However, we have to add the qualification that we would like to have results from a suitable amount of students who have completed the full training programme. Moreover, we would like to analyse relationships between the students' progress during training and their personal data (first language, age, duration of residence in Denmark, duration of language training at a language school, etc.), as well as data on how the training were organised at the training sites, including data about the students' use of the listening trainer which will not be described in this report. Our data shows, for instance, that the student who trained most of all hardly made any progress in Danish pronunciation. It would be very useful to understand why, and such understanding might require access to personal data about the student as well as data about that student's training conditions.

2.1.2 Plan for the report

In the following, the field test phase is described, including general data about training sites, students, and log files received, as well as some general experiences made and what we, based on these, intend to do differently in future (Section 2.2). We then present the methodology we used to sort and analyze the log files from the students' training and self-tests, and we list some of the many independent variables that are involved in analyzing the field data, many of which we know far too little about at the moment (Section 2.3). Section 2.4 illustrates a goodly part of the measurements that have been made per student based on the log files, and we show progress graphs for all 22 students as organized per training site. Section 2.5 presents the log file analysis, including several measurements per student, and discusses the data corpus as a whole from various perspectives. Section 2.6 concludes the report.

2.2 The Training Phase

This report from the Danish pronunciation trainer project describes the results of the first field test carried out at schools and other training sites where immigrants are taught Danish. The pronunciation trainer has two parts. The first part is a listening component with which the students can test their ability to distinguish between various only slightly different Danish word pairs. The second part is a pronunciation trainer component to which the students can pronounce single words and get immediate feedback to which extent the words have been recognized by the speech recognizer of the pronunciation trainer. In the following the use of the pronunciation trainer in the performed field tests is described as well as the results achieved.

2.2.1 Training sites

The interest for the pronunciation trainer has been big. Table 2.1 shows that the pronunciation trainer has been installed at nine training sites. The installation at the first training site took place in November 2004. Since then installations have been made in December 2004 and January 2005. By this time the installations we know of were by and large made. We do not know for sure whether all training sites stated have actually installed the pronunciation trainer upon reception, or what did not work as planned in case the system was not installed. All training sites have been instructed to contact NISLab in case of problems with installation, use or forwarding of log files.

Training site	**Geography**	**Number of log files from training site**	**Number of students in training**	**Number of students with measurable results**
Aalborg Sprogcenter	Aalborg	0	0	0
AOF	Svendborg	19	7	0
EUC Nord	Hjørring	0	0	0
FO	Odense	228	25	4
Monitor	Odense	252	15	10
Nordjyllands Idrætshøjskole	Aalborg	0	0	0
Riisingprojektet	Odense	91	7	4
Studieskolen	Odense	203	19	4
S&S	Svendborg	28	15	0
In total	**9**	**821**	**88**	**22**

Table 2.1. Overview of training sites, log files and students.

The project purpose of installation at a number of training sites has been twofold. Partly it has been important (1) to gather experience as to installation and use of the system at various training sites. An analysis of these experiences is essential in order to be able to optimize future installation and use of the system. This is due to the fact that the training sites are very different when it comes to, among other things:

- Student population

- Technical equipment
- Technical expertise
- Leadership and purpose
- The possibility and motivation of the staff to spend the time needed for installation, communication with NISLab, and motivation of the students of the site to use the system

Therefore it is important regarding a later comprehensive installation of the system to gather experience in using the system under the extremely different conditions proved in the actual field experiment. The gathering of experience should also include the efficiency of the technical support given to the training sites, possible flaws in the first edition of the user manual for the system etc.

Furthermore (2) the purpose of the installation of the system at the training sites has of course also been to gather results showing to which extent the students are actually making progress through pronunciation training with the system.

During the field test period NISLab has supported the installation of the system at the training sites, partly through being present at the training sites in the local area, partly through hotline support (email and telephone) for all training sites. Apart from this instruction the training sites have received the first edition of the system user manual.

As appears from Table 2.1 three training sites have not delivered any log files at all. Two other training sites have indeed delivered log files, but none of these fulfil the conditions stated in Section 2.3 below, in order to be usable for evaluation of the students' progress. Some of the possible reasons for the lack of producing usable test results could be:

- Loss of spirits upon reception of the system, maybe due to lack of time for the staff at site
- Technical and insuperable difficulties making the system run at site, maybe even in spite of request for NISLab support
- Technical and insuperable difficulties forwarding log files, maybe even in spite of request for NISLab support
- That the amount of produced log files enabling NISLab to evaluate the progress of the students at the training site was insufficient, that is insufficient training and test activity at the training site

On the other hand four training sites have delivered a large number of log files, which in total have made us able to analyze progress when using the pronunciation trainer for 22 students out of a total of 88 students who have tried the system.

It seems clear that a big number of students have only briefly been using the pronunciation trainer without using it more intensively to improve their Danish pronunciation. The reasons for this can be manifold. For instance it appears from Table 2.1 that FO, Monitor and Studieskolen have delivered almost equally big amounts of log files to NISLab. Nevertheless only four out of 25 students at FO, and four out of 19 students at Studieskolen have delivered sufficient amounts of data for analyzing purposes, while as many as 10 out of 15 students at Monitor have delivered sufficient amounts of data. As these differences are very significant it cannot be excluded that they are hiding some circumstances, which could be important to be aware of in order to optimize the success rate of the future use of the pronunciation trainer in field tests. A possible explanation could be that the staff at Monitor has to a big extent taken care of the students, encouraged them to use the pronunciation trainer, assisted them in connection with the training, and generally been more enthusiastic towards the students. Riisingprojektet has delivered far less log files than the three afore-mentioned schools.

Nevertheless we have analysable data for four out of seven students from Riisingprojektet, comparable with the results from Monitor. The opposite extreme so to say, is seen for instance at S&S where as many as 15 students have only delivered 28 log files in total. If these significant differences are not primarily due to the attitude and support of the staff, another explanation could be that the student populations are very different at training sites which show big differences in the ratio between the number of students who have only tried the pronunciation trainer and the number of those who have carried through substantial training with the system.

At least at three of four of the active training sites the pronunciation training is continuing with a considerable number of students at the time of writing. Table 2.1 shows that in fact only the four training sites in Odense have delivered usable log file sets. This may be due to the fact that these training sites received support from NISLab whose staff was present at the training sites, in some cases more than 10 times. Still, we have no good explanation why the number of received log files is decreasing proportionally with the geographical distance of the training site from NISLab.

For all training sites the procedure is that all training and test sessions of the students with the systems are logged, and that the log files are sent to NISLab for selection and subsequent analysis, given that the log files for a student comply with the criteria of the project for measurability.

2.2.2 General experiences

As far as we know the training sites have generally received the pronunciation trainer with great enthusiasm - at least as a first reaction. At the same time it has shown that several training sites for many different reasons have not had or been able to spend the time and expertise needed for creating results with the pronunciation trainer which could be evaluated. This was expected from the beginning, and was anticipated by recruiting a bigger number of training sites than necessary from a point of view, which solely focuses on test results and their analysis.

A large number of log files have been delivered from the training sites, 821 in total, and they show generally that some aspects of installation and use of the pronunciation trainer should be improved in the months to come. The system itself is fine, both technically and with regard to usability for practical self-training, meaning that the difficulties of some training sites are due to other factors. The same counts for the forwarding of log files to NISLab.

The main problem remaining to be solved concerns the training of the students with the system. The students can train as well as test themselves with immediate feedback regarding their performance. However, (1) at the present time the system does not distinguish between training and test situations. This means that the log files have to be separated manually at NISLab before evaluation can be done. Furthermore (2) the instructions for the users and the teachers at the training sites in the user manual are not sufficiently explicit concerning recommendations about how to train and test. This means that comparable test results from a particular student are not generated with the desired level of regularity.

These two problems are being solved, (1) by adding functionality to the system so that the students themselves can choose whether they want to train or test. The test files are then marked as test files, and these are the files that the subsequent evaluation will focus on, while the training log files are only considered as data when the total training effort of a particular student is being looked upon. Furthermore this means that the student immediately after a test can get feedback regarding the total test result. (2) is solved with a new version of the user

manual which explicitly recommends, or maybe rather dictates, a certain procedure of training and testing in order to facilitating the subsequent evaluation. It has complicated the result analysis that the students not with a suitable regularity try to pronounce the same sequences of the 450 Danish words of the system.

In this connection it should be stressed that we face the task of creating an entire new ”culture” in using new information technology for self-training purposes. Nobody is familiar with, let alone used to, this culture, and the field test character of the pronunciation trainer evaluation described here has meant, that we have not been able to communicate directly with the students in order to change their test behaviour. The access to the students goes through the training site, persons at the training site who might themselves, or might not, instruct the students, and a new version of the user manual which it was considered not to send out in the middle of the field test period. Therefore we have to set the improvements in progress on hold till a new field test period can be carried through.

Another important point is that the analysis of the test results below so far have revealed an entire new phenomenon for us which has to be considered by the next version of the user manual, and maybe also by the system, namely the more precise *test conditions* under which the students are testing. This discovery is discussed at length in this report.

An important observation from the test files received is that almost all test persons have not yet trained and tested long enough. The first training sites received the pronunciation trainer in November-December 2004 and most of them had the pronunciation trainer up and running at the end of January 2005. The user manual stresses clearly enough that training is a decisive factor for learning Danish pronunciation, and that one can easily argue that the training effort which is in this report stated as recommended, that is pronunciation of 4500 words during 10 weeks, is too small and should be increased. However, partly because of the necessity of reporting results at the present time, and partly because of the de facto too small average training effort per student which we have noted from the results received, it is difficult in this report to draw the conclusion we would mostly like to present. This is namely to be able to *prove* from the analysis of the data received that with a ”full” training effort a considerable number of students would be able to pronounce single words in Danish almost as the Danes themselves can do it, that is with a test score of +70% and under the most difficult test conditions.

What we can demonstrate below is that the test persons generally have achieved considerable progress with the pronunciation trainer.

Table 2.2 shows an example of a log file. The first column states the number of the word; the second column states the written form of the word, including, in irregular cases, the simplified phonetic transcription of the word; third column shows how often test person has seen the orthography of the word; fourth column shows how often the test person has heard the word pronounced before the person has himself tried to pronounce it; fifth column shows how often the test person has seen and heard the word pronounced on video before the person himself has tried to pronounce it; and the sixth column shows how the system scored the pronunciation of the test person. Please note that far from all log files are as regular as the one shown, which only has one single ”hole” as word number 10 has not been pronounced. For instance the test person can leave out the pronunciation of many or all words, repeatedly pronounce one particular word etc. This stresses the importance of the test persons’ self-generation of ”clean” test log files through a clear separation between training and test in the interface of the system.

ID	Word	Seen	Audio	Video	Score
1	Objektiv [objægtiu]	1	3	1	2
2	Fortælle	1	1	1	2
3	Begynde [begøne]	1	1	1	2
4	Føtex [føtæks]	1	1	1	0
5	Arbejde [Abaide]	1	2	1	2
6	Betyde	1	1	2	0
7	Lunge [långe]	1	2	1	2
8	Betale	1	3	1	0
9	Forsøge	1	2	1	0
10	Pile	0	0	0	
11	Fortsætte	1	3	2	0
12	Gade	1	2	3	2
13	Dele	1	1	1	0
14	Opleve	1	2	1	1
15	Billigst [bilist]	1	2	2	0
16	Udvikle	1	3	1	2
17	Bakke	1	2	1	1
18	Cykel	1	2	1	2
19	Nyheder	1	2	1	2
20	Respons [ræspons]	1	2	1	0
21	Brugsen	1	3	2	2
22	Rideklub	1	3	1	0
23	Omhyggelig	1	1	1	0
24	Gymnastik	1	2	2	0
25	Fortryllende	1	2	1	0

Table 2.2. Example of log file.

2.3 Methodology

2.3.1 Collection of test files

Since November 2004 821 log files have been collected from six training sites, cf. Table 2.1. We have received usable test files meeting the demands described in the following Sections from:

- Studieskolen, Odense, 4 students
- Monitor, Odense, 10 students
- FO, Odense, 4 students
- Riisingprojektet, Odense, 4 students

The data material making the basis of the following analysis is made from test results of 22 students from four different training sites.

2.3.2 Sorting of files from the training sites

The current version of the pronunciation trainer + user manual encourages the users to train pronunciation with the 450 (numbered) Danish vocabulary of the pronunciation trainer and test continuous series of words whenever convenient. We have noted that this combination or

mixture of training and test makes unnecessary demands on the subsequent sorting of log files from the training sites. The next version of the pronunciation trainer will solve this sorting problem. It will offer the students to choose between training and testing and mark all log files as either training or test files. The training files will also be used in evaluating the progress of the users, that is to find out how much and how long they have trained. However, the training files will not be used for evaluating the progress of the users. Only the test files are used for this purpose. These changes will of course appear from the next version of the pronunciation trainer user manual which at the same time will recommend very explicitly to train and test in a systematic and thorough way. In this way we can hopefully totally avoid cases in which for instance student E starts testing the words 205-267, and then tests number 55-96, and never repeats a test sequence, but at the most for instance, ”re-tests” later with the sequence 228-301, by which we only have the intersection 228-267 with the afore-mentioned sequence at our disposal for progress evaluation.

For the time being in the log files we have to sort between training files or training file fragments, and test files or test file fragments. The user will often have trained and tested the same file, and the training does not have to include pronunciation of Danish words. Sometimes the training simply consists in looking at the written words having to be pronounced, listen to their Danish pronunciation, or look at and listen to their Danish pronunciation on video. Sometimes a training file consists in pronouncing a few words several times in order to pronounce them correctly etc. All of this is of course fine for training purposes, but these files cannot be used for systematic evaluation of the users’ progress.

Therefore we sort out the log files, which only or mostly consist in training. What is left is a subset of all log files from a particular student containing test pronunciation of words in succession. We can call these files ’test files’.

2.3.3 Selection of test files for evaluation

The test files are organized in maps per training site and per student at the training site. Then they are organized by test date and according to sequences of words pronounced at the particular date. This is entered in the log file extension. The next step is selection of test files for evaluation. This is done through a selection of test files for each student in which the user (1) tests with the *same* fragment of the 450 test words in numbered order and (2) does it with a considerable interval of time. (1) meaning for instance as illustrated above, that the fragments 375-431 and 400-450 only have the words 400-431 in common. (2) meaning for instance that student E’s test with the words 150-200 on 20.12.2004 and with the same words on 2.1.2005 is worthless for evaluation purposes because the student did not train at all between the two tests.

In the approach to the analysis of the test files it has for us at NISLab been an empiric question to which extent the conditions (1) and (2) above have been met in the log files. Given the relatively small training effort on average per student in the data material, it has mostly only been possible to find fragments of +25 words repeated twice in the data material. Sometimes the fragment is repeated thrice. In these cases we have also measured the middle fragment in time. On average for all students the time interval between the two tests used for evaluation = 5.3 weeks. In two cases we have gone down to one week of time interval that is up to 10 days between first and last test. The results of these short training and test intervals are naturally subject to an increased unreliability compared to results based on 5-10 weeks of training and including a proportionally bigger training effort. The number of words in sequentially ordered fragments are compared with at least 25 identical words. We would have preferred at least 50 words, but this has not been feasible.

2.3.4 Primary evaluation algorithm

As the most users are adhering to the guidance of the user manual about trying to pronounce all 450 words in the numbered order over a period, we can define the notion of a 'test cycle'. A test cycle is a time interval in which the user pronounces all 450 words at least once. This takes usually quite some time, such as six weeks, also because many users are progressing slowly and furthermore train with the same, e.g. 100 words several times, before they proceed to the next sequence of words. As noted in Section 2.3.3 we cannot use repeated test sequences for evaluation purposes if the time frame between them is too short. When the first training site started to deliver data in November 2004 and many of the other ones did not start till late January 2005 or even February 2005, it can be implied that only a relatively small number of users have generated more than two test cycles within which identical test sequences are adequately separated in time to make an evaluation meaningful.

The method for selection of test files for evaluation can therefore be summarized as follows:

- Find user X at training site Y who has at least carried through +1 test cycle. It is sufficient that the user has tested all 450 words + started the next test cycle;
- Find a sequence S of at least 25 words (and preferably more) which X has pronounced at least twice;
- Prefer that S has been pronounced with at least an average test cycle time frame, that is approx. six weeks;
- Clean up the selected log files by removing repetitions of the word pronunciations, note "holes" if any in which words in the word sequence have not been pronounced etc. In this way defective statistics can be avoided;
- Calculate in % the correctness of X's pronunciation of S first time, n'th time and last time it occurs in the material available;
- Correctness is calculated by adding the student's test sequence score of words pronounced and calculate the score in % of the maximum score, that is a score of two per test word;
- Plot X's increase or decrease in pronunciation correctness in a graph for the training site in question;
- Do this for all test persons at training site Y and continue with training site Z;
- Repeat the algorithm for training site Z.

The result of this procedure is for X, Y, and S, a graph for early, possibly intermediate, and late testing.

2.3.5 Training effort

Use of the algorithm in 2.3.4 as plotted in the test graph does indeed take into account the *time* between the start result Rst, the middle result Rml, and the end result Rsl, but the algorithm does not take into account the *training effort* which student X has invested in getting from the start result Rst to the end result Rsl. We are also interested in knowing about the effect of using the pronunciation trainer for different students. Maybe "it has no effect" for some students, whereas it "has great effect" for others. This cannot be seen from the fact that, e.g., both student E1 and student E2 have progressed from the start result 50% to the end result 60%. If E1 has achieved this in a few weeks with modest training, whereas E2 has made it in 13 weeks with massive training, then it tells us something about the differences between students, such as their individual aptness, the influence of their mother tongue on

their aptness etc., and the efficiency of the pronunciation trainer in this connection. Exactly what it tells us can, however, only be decided if we have personal data about every single student being evaluated, which is not the case at the moment as already mentioned.

In order to take the training effort into account we have annotated all progress results in Table 5.1 with parameters illustrating the training effort involved and also the student's total training effort. These annotations make it possible to estimate the efficiency of the pronunciation trainer for each student and make it easier to compare the students' results. In this connection we found that the training effort can be calculated in various ways of which some are more misleading than informative regarding the actual training effort of the students. One of the most central benchmarks of this report for the students' training effort is the number of words they have pronounced during training and testing.

Concerning establishing a main and more meaningful goal for the training effort than the number of trained words, we have defined the notion of *a full amount of training* which = 10 x 450 pronounced words in 10 weeks. In this way the student's actual training effort can be stated in % of a full amount of training. It is an important discussion whether a full amount of training actually should be exactly the amount stated. We consider this a partly empiric question, whose answer can be found by looking at a training effort carried through by an average number of students in order for them to reach an appropriate level in Danish pronunciation. For the time being it is too early to decide the answer on this, or these, questions, as the amount of training for the student population analyzed in this report is not sufficiently big. For related reasons we are neither able to evaluate the long term effects of progress with the pronunciation trainer. Such an evaluation requires both longer training time by a number of students, a subsequent training pause, a new test, as well as supposedly also knowledge of personal student data.

2.3.6 Secondary evaluation

It is never possible to control students to a full degree when they perform self-training and self-testing as opposed to training and testing in the classroom. Therefore the following problem is a necessary consequence of the students' independent work with the pronunciation trainer. As described above the progress of the students is measured by a comparison of their pronunciation of the *same* fragment of words over time. The same fragment of words will only in unlikely cases be found the first or the last day the student is testing with the pronunciation trainer. Therefore we have also considered another goal of training progress, that is a comparison with three test fragments which do *not* all necessarily contain the same words, but are characterised by the (1) differences from the test fragments described in Section 2.3.4, and (2) that they have been tested from the start, a randomly chosen place somewhere in the middle, and at the end of the log file material from the student. This gives a useful control measure for the goal described in Section 2.3.4 for the student's progress.

In total the primary and secondary evaluation of the students mean, that all student progress in the data material (the log files) has been measured at least five times for each student.

2.3.7 Training and test conditions

In the analysis of the progress results in the received log files we sometimes found (by means of the algorithm in Section 2.3.4) an inexplicable big *drop* in the test performance of the student over time. Such drops can occur two places in the quantitative test result room: the student has a high start result, often comparable with the result for native Danes, lying on 78% and upwards according to the project baseline measurements with native Danes – and

next test result shows a drop of 5-15%. Such a *high-drop* does not mean anything. The student can, or is close to being able to, pronounce single words in Danish correctly, and the ”fitness level” of the student decides the measured variation, which will also be found with native Danes.

We define in connection with this discussion *high* as >70% correct, *middle* as 50-70% correct, and *low* as <50% correct.

On the other hand it was surprising to find *middle-to-low* drops or even *high-to-low* drops with some students. The analysis showed here a, for us, unexpected connexion with the fact that these students had changed the *test conditions* in the start and the end test. The test conditions are for the student in the test before the words in the word sequence were pronounced, that he had only *read* the word to be pronounced, what all did; *read* it and *listen* to it on tape; *read* it and *see and listen* to it on video; or *read* it, *listen* to it on tape, and *see and listen* to it on video. What we found in the data material was that the described performance drops seemed to be explicable by the fact that the student had changed the test conditions *upwards* in the following hierarchy:

- read
- read *and* listen
- read *and* see and listen
- read *and* listen *and* see and listen

Our hypothesis is that the changed test condition explains the significant drop of the student’s pronunciation performance. In other words, all or most students with Danish as second language find it easier to achieve correct pronunciation if they use all of the hierarchy above as instruction before they try to pronounce a word than if they only use part of the hierarchy. However, this hypothesis is firstly not fully proved yet because of the limited data material, and secondly it is too imprecise. The indications we have are that video (see *and* listen) is central for the performance for most of the students. If the student does not use video before pronunciation, the student in other words chooses a test condition which is considerable more difficult than if the student also uses the video information. On the other hand we cannot in the available data see any clear effect of whether the student chooses or not chooses to use audio (listen). And neither can we say whether these generalisations count for all students or only for the majority of course. But as we are going to see the evidence is actually massive for whether the use or no use of video before pronunciation is an essential factor for the test score of the students.

In this way the test conditions introduce a new set of independent variables in our field data, variables co-determining the measured progress in the pronunciation training of the students. Without also looking at the test conditions for a student’s pronunciation of a word sequence it is probably not possible to estimate to which extent the student is making progress. In itself the average score of the student for a pronunciation sequence is insufficient.

Another important consequence is that a possible recommendation to a particular student about which further amount of training he should count on in order to reach a satisfying level of Danish single words pronunciation, depends on the training conditions under which the student has achieved his results so far. For instance it is to be concluded, other things being equal, that a student who in the course of training has achieved 57% with the pronunciation trainer using both audio and video, is less good at Danish pronunciation than a student who also has achieved 57% with the pronunciation trainer, but *without* the use of video. With another example a student achieving 47% without using video is probably on average better at Danish pronunciation than a student achieving 57% with the use of video.

Comparison on equal terms of two students' performances, or of the performance of the same student at different points of measurement in the course of training, demands in this way that the measured test scores have been achieved under the same test conditions. Given the limited student material of 22 students for the time being this condition constitutes a certain difficulty for relevant comparison of the performances of the students.

2.3.8 Evaluation uncertainties

The Sections above have described the complexity involved in evaluating the progress of the students based on the field test log files. Apart from that it is important to make clear that the evaluation in this report of the students' progress is subordinated a number of uncertainties that cannot be eliminated, such as:

- The concrete test situation: noise, the attitude of the student to the testing, etc.;
- Technical difficulties during the test, e.g. a failing microphone, a badly fitting head set;
- Test instructions on the training site, personal support to the student, user manual;
- Use of the listening trainer as part of the training;
- The life of the student in Denmark when not at school, the duration of the student's stay in the country so far, how long the student has been in the language school, how often the student associate with native Danes, etc.;
- The mother tongue of the student.

We do not know the concrete situation in which a student has self-tested with a word sequence. We assume that the students by and large have self-tested without disturbances during testing. But we have reason to exclude that the self-testing of the students has sometimes taken place with other persons in the test room, with background noise disturbance of the speech recognition, etc. The same counts for the individual student's attitude to the use of the pronunciation trainer that is whether the student has worked hard and determined on training and testing or whether the student has not been very serious. Students, who have done a relatively bigger training effort, have generally had, we assume in advance, a serious attitude to training and testing. This is highly confirmed by their progress results, which we are going to see. On the contrary we do not know anything for certain about students who have only completed relatively few training and test situations.

Nor do we have sufficient knowledge about the extent to which those responsible for the use of the pronunciation trainer at the training sites have taken care of the students' training and testing, discussed the user manual with them, supported during testing and assisted in dealing with the technology and the equipment, etc. Nor do we know in this connection whether the individual student from whom we have received log files have actually read the user manual.

We do not know to which extent the measured pronunciation training progress of a student is influenced by the student's use of the listening trainer in parallel to the pronunciation training.

Finally we also have to mention the important independent variables concerning the rest of the student's Danish learning before and during the pronunciation training, the language and culture background of the student, etc. We are not aware of these variables.

2.4 Graphs and examples

In this chapter we present the progress graphs according to the primary evaluation algorithm (Section 2.3.4) for all 22 students per training site.

The graphs showed in Figure 4.1, 4.2, 4.3, and 4.4 are based on tables containing measurements per student. An example of a table overview for an individual student can be seen in Table 4.1. The right column of the table contains some explanations to the reader of this report. The upper part of the table shows progress according to the primary evaluation algorithm, while the lower part shows the results of the three control measures per student. These are discussed in Section 2.5. See Table 5.1 for measurements of the students' training effort.

Attribute	**S3 = student ID**	**Explanations**
First test date	1.2.	
First test number of words	23	
Word ID	115-139 -2	word sequence
Test condition	RA	read (R) and listen (A)
Score start	29/46 = 63%	
Score middle	-	no middle score
Last test date	19.4.	
Last test number of words	25	
word ID	115-139	word sequence
Test condition	RA	read and listen
Score end	44/50 = 88%	
Trend	middle-high	trend verbally expressed
Comment	Remarkable progress	
Training weeks test start to test end	10	
Number of log files	13	
Scored sample first or early	3.2 59/98 = 60%	first sample measurement
Word ID	156-204	word sequence
Test condition	RA	read and listen
Scored middle sample	15.2 69/118 = 58%	second sample measurement
Word ID	326-384	word sequence
Test condition	RA	read and listen
Scored sample late or end	26.4 66/82 = 80%	third sample measurement
Word ID	140-180	word sequence
Test condition	RA	read and listen

Table 4.1. Example of student progress measurement.

2.4.1 Studieskolen

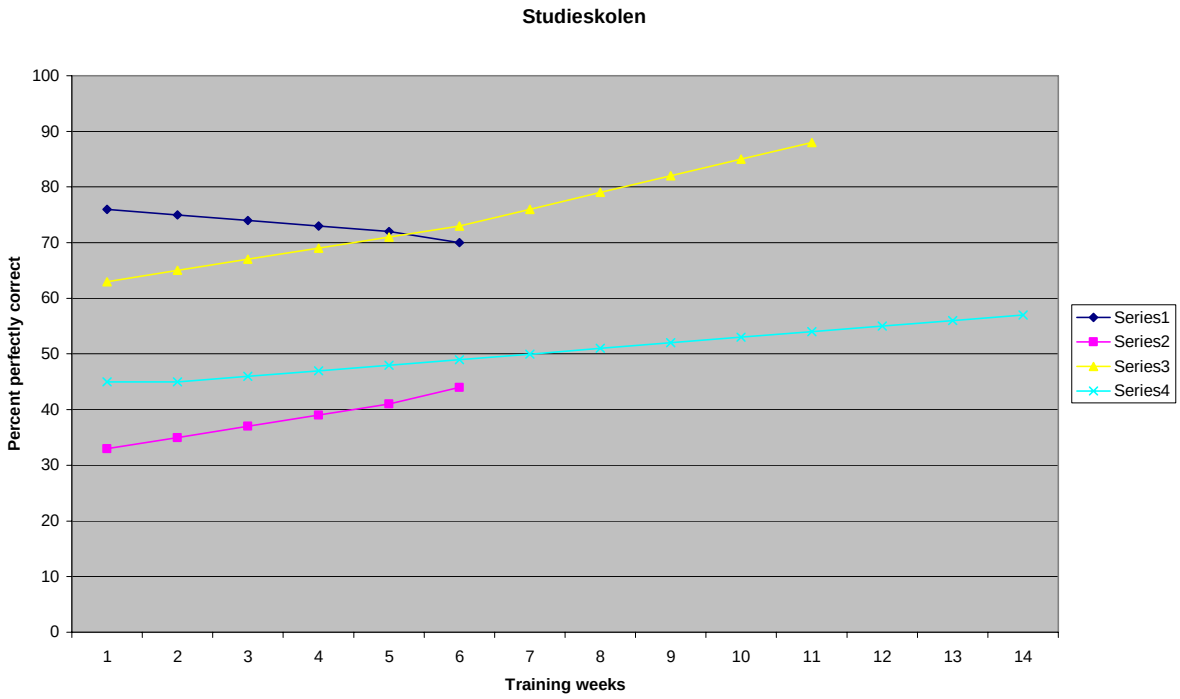


Figure 4.1. Studieskolen.

2.4.2 Monitor

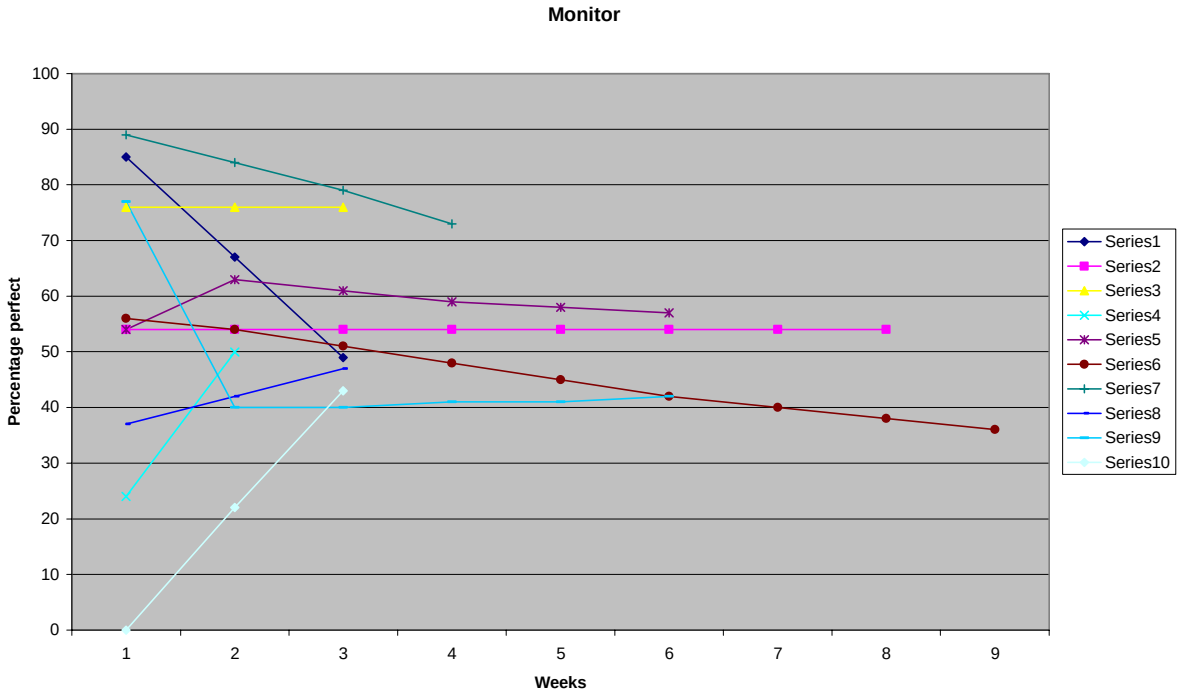


Figure 4.2. Monitor.

In all the graphs the reader can identify the students being discussed under their ID, which is combined of the first letter of the training site and a running student number. The student in Table 4.1 has in this way ID S3, that is student number 3 from Studieskolen. In the graph from Studieskolen in Figure 4.1 the number of the student corresponds to the ”series” number in the graph. However, the graphs should be printed in colour in order to make it easier to see to which student each curve corresponds.

The graphs show furthermore number of training weeks from start test to end test (X-axis), and the score of the student in start test, middle test if any, and end test (Y-axis). Please note that all other plotted points except for the start point, the end point and the middle point if any per student have been inserted by the writers to make Excel show the results in the line graphs. These other points represent therefore not measure points derived from our data.

In other respects the graph from Studieskolen is self-explanatory.

On the contrary the graph from Monitor is quite lively to watch in the beginning. This is due to the fact that all anomalies among the analyzed students can be found at Monitor that is four anomalies in total. These anomalies are presented separately in Table 5.1 and discussed in connection with the table.

2.4.3 FO

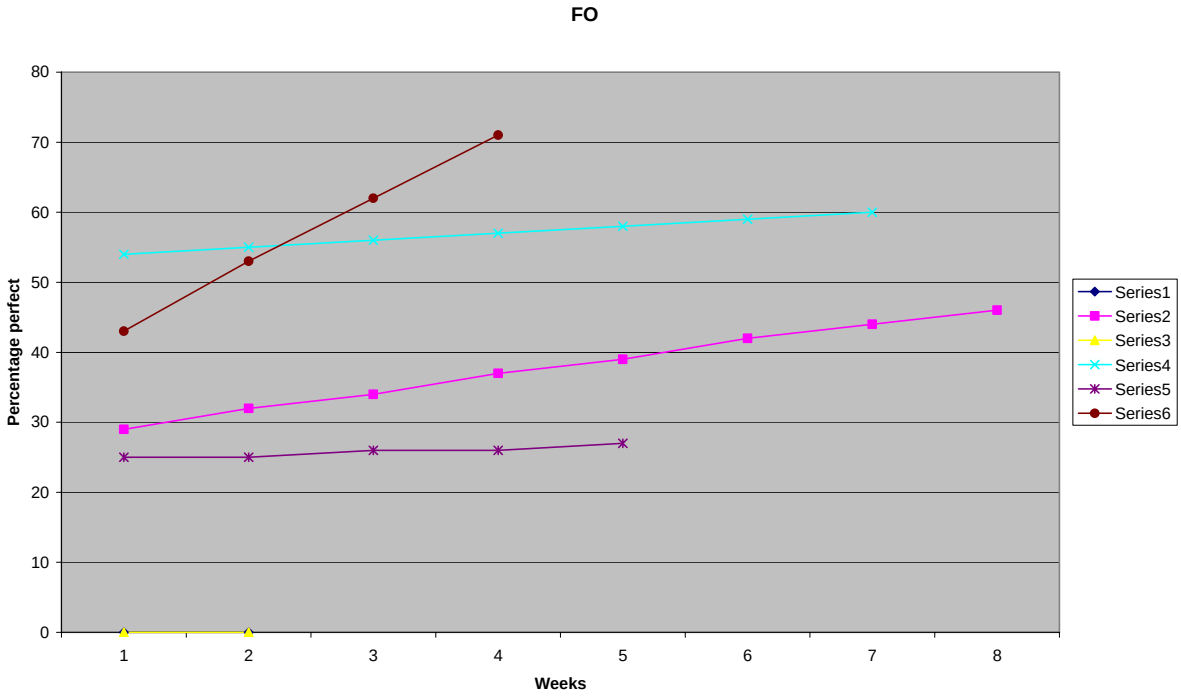
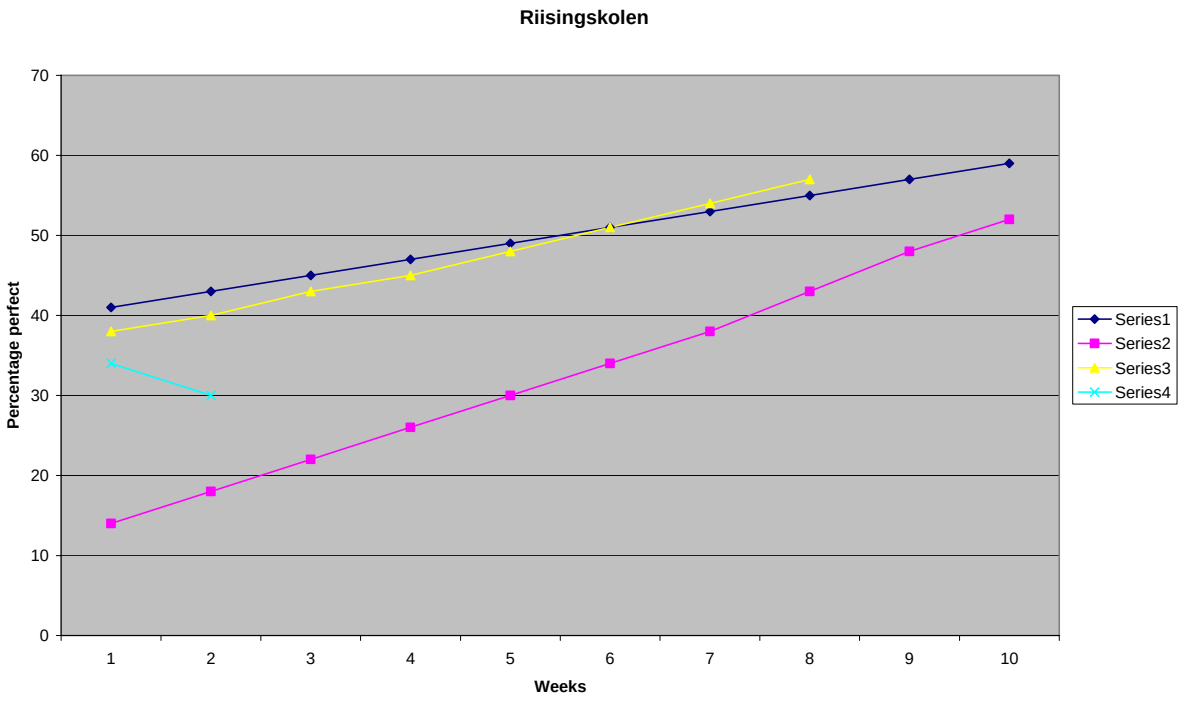


Figure 4.3. FO.

The graph from FO is by and large self-explanatory. The two 0-0 scores which can faintly be seen at the bottom of the graph between week 1 and week 2 are from two students at FO who were not included because of insufficient data, but who had already been given an ID.

2.4.4 Riisingprojektet



Figur 4.4. Riisingprojektet.

The graph from Riisingprojektet is self-explanatory. However, please note the short Series4, that is student R4. Such students, who have only trained for one week are marginal in our results because they have trained so little that it is difficult to build anything on the trend seen in their progress.

2.5 Analysis

In this chapter the results from the progress of the 22 students with the pronunciation trainer in the first field test of the project are analyzed. The superior principle of analysis has been to analyze the big data material in still deeper layers of analysis till all the questions we have been able to pose have found answers consistent with all known data material properties. In the absence of details about test persons and training sites, it has not been possible to correlate the found results with such data. On the other hand we are of the opinion that the perspectives established below in their totality are sufficiently broad and deep to substantiate the main conclusions of this report.

2.5.1 Overview of the results

Table 5.1 shows progress per student according to the primary evaluation algorithm (Section 2.3.4), the training effort measured in different ways, the training conditions at test start and end, and comments to the results. The Table is divided in two. The first part shows 18 standard cases in the data material where the students have made more or less progress under similar, but not always identical training conditions. The second part (4 students) shows different anomalies, which we would like to emphasize, as well as hypotheses to explain them. Table 5.2 explains different expressions in Table 5.1.

Progress Student	Progress in % of max. possible score	Training effort	Test condition start/end	Comments
S1	76-70% high-drop	5 weeks test effort = 11% effort, test words = 0 total effort = 18%	RA, a little V RA	Performance drops also occurs for native Danish speakers. Modest training effort. Can almost pronounce single words in Danish. Note that the end test condition is a little harder than the start test condition.
S2	33-44% low-up	5 weeks test effort = 7% effort, test words = 0 total effort = 13%	RA, a little V RA	Progress with very modest training effort, mostly without video.
S3	63-88% middle-high	10 weeks test effort = 9% effort, test words = 0 total effort = 15%	RA RA	Good progress without use of video. Can pronounce single words in Danish.
S4	45-57% low-middle	13 weeks test effort = 33% effort, test words = 1 total effort = 33%	RAV RAV	Progress with some training effort.
M2	54%-54% midle-level	7 weeks test effort = 39% effort, test words = 6 total effort = 60%	RA RAV	Considerable training effort without progress in the analyzed test files. Note that video does not seem to help this student. Explanation hypothesis: individual differences: slow learner?
M3	76-76% high-up	2 weeks test effort = 16% effort, test words = 0 total effort = 20%	R, a little A R	Can pronounce single words in Danish. Progress because audio was not used in the end test.
M4	24-50% low-middle	1 week test effort = 6% effort, test words = 1 total effort = 8%	RA, a little V R	Good progress with very modest training effort. Note that neither audio nor video was used in the end test.
M5	54-57% middle-up	5 weeks test effort = 33% effort, test words = 2 total effort = 40%	RAV RAV	Very modest progress with some training effort.
M8	37-47% low-up	2 weeks test effort = 8% effort, test words = 1 total effort = 23%	RAV RAV	Progress with modest training effort.
M10	0-43% low-up	2 weeks test effort = 15% effort, test words = 1 total effort = 30%	RAV RV	Good progress with some training effort. Note that audio was not used in the end test.
F2	29-46%	7 weeks	RV	Good progress with some training

	low-up	test effort = 29% effort, test words = 2 total effort = 43%	RV	effort.
F4	54-60% middle-up	6 weeks test effort = 21% effort, test words = 4.5 total effort = 36%	RV RA	Good progress as end test did not include video.
F5	25-27% low-up	4 weeks test effort = 2% effort, test words = 0.7 total effort = 7%	RAV RA, a little V	Modest progress with very modest training effort.
F6	43-71% low-high	3 weeks test effort = 4% effort, test words = 0.4 total effort = 12%	RV RV	There was a rather big "gap" in the end test file. We cannot exclude that the big progress is partly due to the fact that the student did not pronounce words which the student found difficult.
R1	41-59% low-middle	9 weeks test effort = 29% effort, test words = 3 total effort = 29%	RAV RA	Very good progress given that the final test condition did not include video.
R2	14-52% low-middle	9 weeks test effort = 38% effort, test words = 9.5 total effort = 38%	RA, a little V R, a little A	Very good progress given that the final test condition did not include video.
R3	38-57% low-middle	7 weeks test effort = 43% effort, test words = 7.9 total effort = 43%	R, 2/3A, 1/2V RA	Good progress as end test did not include video.
R4	34-30% low-drop	1 week test effort = 11% effort, test words = 1.5 total effort = 24%	RAV RAV	A drop after a single week of training can be without significance.
Anomalies Student	**Description**	**Training effort**	**Test condition start/end**	**Comments**
M1	85-49% high-low drop	2 weeks test effort = 6% effort, test words = 2 total effort = 13%	R R	Very modest training effort. The big drop is unexplained. Note that the student only uses text. The start score of 85% implies that the student can pronounce single words in Danish. Explanation hypotheses: unserious end test? Different persons in the two tests?
M6	56-36% middle-low drop	8 weeks test effort = 20% effort, test words = 2.4 total effort = 28%	RAV RA	Seems to be a case of what happens when the student goes from having video support to not having it. Explanation hypothesis: video not used in end test.
M7	89-73% high-drop	3 weeks test effort = 11%	R, a little V	Performance drops also occur for native Danes. Modest training effort.

		effort, test words= 0 total effort = 15%	R, a little V	Can pronounce single words in Danish. The performance drop is otherwise unexplained.
M9	77-42% high-low drop	5 weeks test effort = 30% effort, test words = 4 total effort = 40%	RV R, a little A	Seems to be a clear case of what happens when the student goes from having video support to not having it. Explanation hypothesis: video not used in end test.

Table 5.1. Test results, training effort and training conditions per student.

2.5.1.1 Explanations to Table 5.1

Term	**Explanation**
A	The student used audio to hear the words in the test sequence before they were pronounced
Student ID	Students from Studieskolen, Monitor, FO, and Riisingprojektet are marked S(n), M(n), F(n), and R(n) respectively
Full training and test syllabus	Assumed training effort = 450 words per week for 10 weeks = 4500 words
Effort, test words	Number of times the student has trained and tested with the test words between start and end test
R	The student used written text to read the words in the test sequence before they were pronounced
Test effort	Estimated number of words pronounced from start test to end test in % of full training syllabus
Total effort	Estimated number of words pronounced during all training and testing in % of full training syllabus. The students usually go through all 450 test words repeatedly, but there are individual differences in how systematically they do it and which systematic they are following
Weeks	Number of weeks between first and last test date
V	The student used video to see and listen to the words in the test sequence before they were pronounced

Table 5.2. Explanations to Table 5.1.

2.5.1.2 Comments to Table 5.1

Table 5.1 shows in fact two central sets of information. Firstly it summarizes per student the quantitative progress of the student in the measured tests that is those, which are also shown in the graphs in Section 2.4. From Table 5.1 it can also be seen under which test conditions every test is performed (Column 4). The right column in Table 5.1 comments the progress of the students. Secondly the table shows a number of central measurements of the students' training effort (Column 3). Each of these goals is explained in Table 5.2.

Apart from that Table 5.1 is divided in a (1) main group of students who carry through their tests under approximately the same test conditions, except for the fact that around half of them in their final test chose test conditions which were slightly more difficult; and (2) a group of four anomalies in the material.

2.5.1.3 Anomalies in Table 5.1

The anomalies are the following:

Student **M1** is a genuine anomaly. We have no explanation of the measured big performance drop from 85% to 49% in the test. The control measures show performances between these extremes. However, an important point is the following. If a student with the pronunciation trainer is capable of performing 89% correctly simply by reading the words to be pronounced which is the case for M1, then the student is speaking Danish at a native level and does not need to use the pronunciation trainer. The student is already at, or very close to his peak performance, and any progress or retrogression measurement of the student corresponds to measuring daily fluctuations of native Danes' pronunciation.

It appears more general that all students performing over 70% during a test *by means of reading the test words only* must already be very close to their peak performance. Therefore progress measurements of such students are practically in vain. In fact it is absurd to use the progress or retrogression of these students when measuring average progress for all students. Please note in this connection that the main group in Table 5.1 also includes two students with this profile. However, S1 is also using audio as a test condition, which is the reason why this student does not belong to the main group after all. The other student, M3, is also in the main group because this student does not only use read text in the start test, but as can be seen from the table the affiliation of this student to the main group is marginal.

For the above reasons **M7** is also an anomaly. The 89% in the start test using a little bit of video shows that M7 is close to his peak and must be assumed to pronounce Danish by and large correctly. M7's relapse during the course of test therefore corresponds to the performances of native Danes.

In some of the calculations below we will leave out M1 and M7 as these two students are already close to their peak in the start tests.

The two other anomalies are **M6** and **M9**. We explain their significant performance drop with their failure to use video in their end tests whereas they used video in the start test. These two examples show the very significant effect of changing the test conditions from using to not using the video. It would be misleading to include M6 and M9 in the main group, as the main group shows progress under relatively unchanged, but not absolutely unchanged, test conditions.

2.5.2 Training effort and results

Table 5.3 states the average training effort of the students in various ways.

It appears that the students on average have trained for 5.3 weeks between their first and last test. Their total training measured in number of pronounced words is 26.7% of a total amount of training of 4500 words. In other words we can clearly see that the students in general have trained far less than a full amount of training.

Among the students there are huge differences in the total amount of training, that is, as it appears from Table 5.1, from 7% to 60% of the full amount.

These observations mean that *the results of this report must be considered as strictly provisional regarding evaluation of the full potential of the pronunciation trainer for self-training.* We still do not have a data material showing the effect of carrying through a full amount of training for a sufficient number of students. In fact we do not have data from a full amount of training from any student. This raises the question whether it is possible from the available results to extrapolate, or in other ways estimate, the learning effect of carrying through a full amount of training.

	Number of students	Training weeks between tests	Training between tests in % of full syllabus	Repetition of test sequence in test period	Training total in % of full syllabus
Main group	18	98, average 5.44	354%, average 19.67%	41.5, average 2.31	492%, average 27.33%
Anomaly group	4	18, average 4.50	67% average 16.75%	8.4, average 2.10	96%, average 24%
All	22	116, average 5.27	421, average 19.13%	49.9, average 2.26	588%, average 26.73%

Table 5.3. Amount of training, overview.

Table 5.3 shows that the students on average have trained 19.1% of a full amount of training *between* their first and last test. The main group of 18 students in Table 5.1 has had an average progress of 14.2% between start and end test. We suggest a little correction to this average by excluding the two students whose results are on a high level in both tests that is above 70%. These students are probably already around their peak performance at the start of the test sequence. With this correction we get an average progress for the remaining 16 students in the main group of 16.4%.

2.5.2.1 The training cost for 10% progress

Based on the above figures we can calculate that the cost for **10% test score progress** in absolute percentage points added to the start score for the 16 students is a training effort of a little less than **12% of a full amount**. This is a quite positive result, a result, which on average can be carried through with a fairly moderate training effort per week in a little less than 3.5 weeks according to Table 5.3. Furthermore it should be noted that this result according to the collected field data is a **minimum result**. In fact 12% of a full amount gives more progress than 10% when the test conditions are also considered. More than half of the main group of test persons made the end test harder for themselves than the start test by choosing more difficult test conditions (Table 5.1). Unfortunately it is for the moment not possible to quantify this further pronunciation competence progress. This would demand more data from students who have self-tested with the same test sequence on different training conditions.

Furthermore it is important to understand the above good result in the light of the test conditions. So, in order to be more exact, the result is that the students *on the same or a little harder test conditions* reach their *first 10% progress* by means of *12% of the full course syllabus.*

2.5.2.2 Linear progress?

How far is it possible to extrapolate from the good result shown in Section 2.5.2.1, given that we do not have data from training courses carried fully through in which 4500 words have been trained and tested? We cannot give the answer to this at the moment, as we do not have full training amounts from a sufficient number of students. In fact none of the students have carried through a full amount of training yet.

Two extremes are in this context quite clearly wrong. The first extreme is that progress with the pronunciation trainer is linear. This is of course wrong, not only because it is not possible to obtain a result higher than a 100% perfect score, but also empirically because adults who are taught Danish as a second language often stop their progress at a state where it is still possible to hear a more or less clear accent. Our baseline of the pronunciation trainer with

Danes showed that native Danes reached a test score on the most difficult test condition (read) from a little less than 80% to almost 95%.

An estimate could therefore be that a goal for students with Danish as second language would be at least 70%. When this level is reached it is no problem to be stagnant as everybody will be able to understand the students' Danish pronunciation even though they have an accent.

The other extreme is that progress with the pronunciation trainer quickly will be stagnant after the first 10-20% progress. We believe with the results in this report to prove this wrong. Some students have improved approximately 30% with a relatively modest training effort.

Therefore the truth is probably somewhere between the two extremes. Actually we will also be able to find the truth, at least as a probable approximation, when we have a sufficient number of students who have carried through the full amount of training.

Again it must be stressed that the discussion above concerns students who are testing on identical or slightly more difficult test conditions. This means e.g. that a student who with a 70% full amount of training goes from 32% to 72% on the easiest test condition, i.e. (read + listen + see and listen) before pronunciation of every word, can still learn quite a lot before the student reaches his peak of Danish pronunciation. Therefore it would be extremely useful to have a conversion factor based on sufficient data showing how well a student, who performs a X% score on the (read + listen + see and listen) test condition, will score on average if the test condition is changed to one of the more difficult (read + listen) or (read). We do not have sufficient data to calculate an approximation for such a factor or factors.

2.5.2.3 Conclusion about progress as a function of the training effort

After analysis of the pronunciation trainer field tests we have learned some central and fundamental facts, namely that the pronunciation trainer gives the students good and fast progress of 15-20% on average in absolute percentage points which are added to their start score, given that the students are testing on identical or slightly more difficult test conditions.

We have also noted that there are two very important factors, which we know far too little about for the time being. The first is a progress estimate function allowing us to estimate the average progress for a student with a full amount of training, given that the student has carried through X% of the full syllabus. The other is a conversion function allowing us to estimate the average progress for a student who has carried through X% of the full syllabus and scored Y% improvement on the same test condition, and where we want to calculate how big progress the student will be able to make with a full amount of training on a more difficult test condition.

2.5.3 Individual and mother tongue differences

For the time being we have not access to personal data about the students. It is evident that these are very different with regard to mother tongue, the duration of their stay in Denmark so far, the extent of their language school education so far, abilities to quickly picking up pronunciation of a new language, age, etc. Therefore we cannot expect in the available material, when it is not grouped according to parameters as the ones just mentioned, to find any special correlation between amount of training between start and end test, and Danish pronunciation progress. However, we can note from Table 5.1 that quick progress is not limited to students starting with low scores, which seems quite promising. The tables 5.4 and 5.5 demonstrate this.

R3	19%	43%
M2	0%	39%
R2	38%	38%
S4	12%	33%
M5	3%	33%
F2	17%	29%
R1	18%	29%
F4	6%	21%
M10	43%	15%
R4	-4%	11%
S3	25%	9%
M8	10%	8%
S2	11%	7%
M4	26%	6%
F6	28%	4%
F5	2%	2%
Student	Progress	Test effort

Table 5.4. Progress in relation to training effort.

M10	43%	15%
R2	38%	38%
F6	28%	4%
M4	26%	6%
S3	25%	9%
R3	19%	43%
R1	18%	29%
F2	17%	29%
S4	12%	33%
S2	11%	7%
M8	10%	8%
F4	6%	21%
M5	3%	33%
F5	2%	2%
M2	0%	39%
R4	-4%	11%
Student	Progress	Test effort

Table 5.5. Training effort in relation to progress.

From Table 5.1 it is evident that the progress of the students implies huge individual differences. As one of the extremes student M2 can be mentioned who does not improve at all in spite of that this particular student (1) trains more than all other students in the data material, that is 60% of a full syllabus; (2) trains 39% of full syllabus between start and end test; and (3) even carries through the end test on a normally easier condition (read + listen + see and listen) than the start test (read + listen). The data material shows only one other student who carries through the end test on an easier test condition than the start test. All other students carry through the end test on identical or a more difficult test condition than the start test. M2 has a stable 54% test score all through the measured training duration.

As the other extreme can e.g. S3 be mentioned who with only 9% of a full training syllabus improves from 63% to fantastic 88% and on identical training conditions (read and listen). As the authors of this report have not yet been given access to details about every single student we cannot elaborate on how much these enormous differences are influenced by the mother tongue of the student. We know that the mother tongue of S3 is Polish, but we do not know the mother tongue of M2. Therefore we cannot yet elaborate on the meaning of the students' mother tongue in relation to their Danish pronunciation progress or stagnation.

However, it is evident that individual differences are also very important. Some students have a good aptitude for languages while others are less minded. As long as we do not have mother tongue information about the students, there is not much sense in trying to separate these two important factors. We probably also need data from more students than we presently have in order to do this with some indication of certainty in the conclusions.

At the moment, however, it is obvious to recommend that pronunciation trainer students are monitored to see whether they in the course of e.g. five weeks and with a training effort of 30-50% of the syllabus have or have not done substantial progress, again considering the test conditions. In case they have not improved substantially on identical test conditions, one should probably try with a language teacher in order to, through individual guidance, to be able to unblock their lack of flexibility in their Danish pronunciation training. There might also be students who are simply not capable of improving. However, we do not have any evidence at all for this presumption from our field data.

2.5.4 Students with identical test conditions

Figure 5.1 shows the progress of the seven students in the normal group in Table 5.1 who chose identical test conditions in the start and end test.

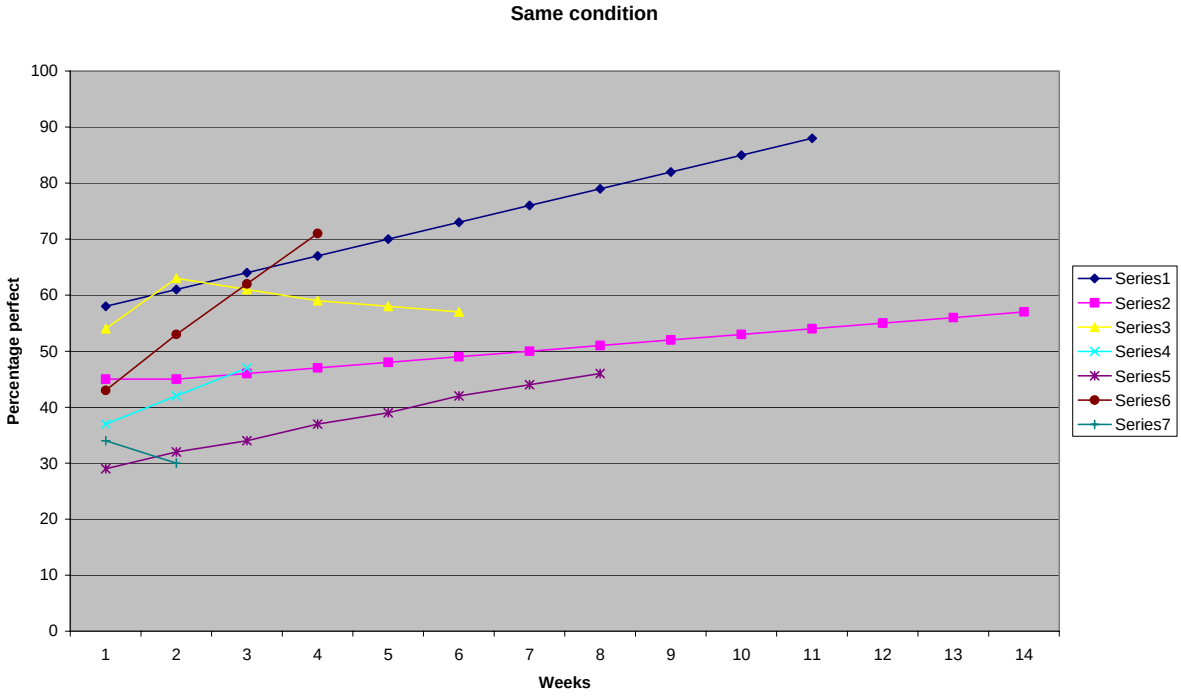


Figure 5.1. Students with identical test conditions in start and end test.

It is interesting that the average progress of these students is 14%, i.e. a little lower than the average for all students in the normal group. At first this could seem to demonstrate, expressed in self-contradictory way, that a test does *not* get more difficult by choosing a more difficult test condition. However, our hypothesis is that most students are very conscious about the difficulty difference between the test conditions. Therefore the students in Figure 5.1 are partly students who do not possess the courage to self-test on a more difficult test condition, or who have the goal of scoring as good as possible on an identical test condition. On the other hand those students who self-test on more difficult conditions at the end are those who have set higher goals with their training.

2.5.5 Control measures

We focus in this report on comparing progress per student in pronouncing *the same word sequence* on similar training conditions, cf. Section 2.3.4. The requirement of finding identical word sequences for comparison in the log files of a student means that (1) often the compared test results do not originate from very early or very late in the training. At the same time the log files we have analyzed only constitute a minor part of all the log files of a student. Therefore it is very possible that (2) the two (or three) test measure points per student could give a misleading picture of the progress of the student during the entire training period, especially given the many and to us mostly unknown independent variables which could influence every single test, cf. Section 2.3.8.

For these two reasons we have made three control measurements per student on new log files, which we have not used before, cf. Section 2.3.6. The control measurements have been made on the earliest not already measured log file, the latest not already measured log file, and a randomly chosen log file more or less in the middle between the two (in time). These three new log files will normally not show the performance of the student on the same word sequences used with the primary evaluation algorithm (Section 2.3.4). But they will show a time progression for the student, which can subsequently be compared to the already measured progression according to the primary evaluation algorithm.

The primary and secondary (sampled) test measurements per student were inserted in a table showing also the test conditions the student had had during the five test measurements in total per student represented in the table. Finally comments are made in the table per student about the relation between the two series of results. An example is shown in Table 5.3.

Student	**Test scores start to end %**	**Test conditions**	**Sampled scores %**	**Test conditions**	**Comparison sampled scores/ test scores** **Other comments**
S3.	63 88% middle-high	RA RA	60 58 80	RA RA RA	Confirm test

Table 5.3. Example of comparison between primary test scores and sampled test scores.

Our main hypotheses concerning what the comparison between the two test series, the main test and the sample test, would show was:

- Bigger dispersion in the sample test series as this in general covers a longer test duration;
- In most cases the same progress rate in main test and sample test;

- Higher end scores in the sample test than in the main test in a number of cases, given the same test conditions as the sample test often took place later in the training course;
- Same effects of different test conditions as found in the main test.

What we found was the following:

- Significantly bigger dispersion in the sample test;
- Identical progress rate in the main test and the sample test in 16 cases;
- Significantly better performance in the sample test than in the main test in three cases;
- Significantly worse performance in the sample test than in the main test in a single case;
- Confirmation of the score pattern in the anomaly cases in Table 5.1 illustrating the effect of testing with and without video: two cases.

As a conclusion the sample test shows (1) *the same result patterns* as the main test and (2) that a number of students during the total training course achieved *bigger pronunciation progress* than shown in the figures for the main test. From a higher point of view it seems that we can conclude that far most of the students have trained determined and seriously with the pronunciation trainer. Had this not been the case, the comparison between the results of the main test and the sample test would have demonstrated far more inconsistencies and apparently random fluctuations than is actually the case. This in itself is a significant result showing that the pronunciation trainer in its present version in fact inspires almost all students to make their in order to achieve a good result.

2.5.6 Training conditions - again

For us it has been a discovery in the data material to observe how much the score of the students depends on the test conditions. The students themselves seem to have a clear understanding under which test conditions it gets harder for them to achieve their best score. With the exception of merely two cases out of a total of 5 x 22 = 110 measure points, it counts for all students that they *either* self-test at the beginning and the end under *the same* test conditions or they self-test at the end under *harder* test conditions. The students doing the latter still demonstrate all progress, with the exception of the two anomalies choosing drastically to drop video completely in the end test. Furthermore we can see that most students performing a high start score do not use video or only use video to a limited extent. Native Danes do neither use video, nor audio for that matter. To them it is sufficient to read the word to be pronounced and then pronounce it according to their mental model of the pronunciation of the written word.

We draw to consequences:

- A confirmation of our hypothesis about the ranking of difficulty of the various test conditions;
- Far the most students have in fact trained seriously and determined.

2.5.7 Recommendations

2.5.7.1 User manual and student guidance at the training site

The students should in the user manual be encouraged to self-test with the same word sequence under various test conditions. When they have improved so much that they do not need video, they should do so in order to reach a higher level in the pronunciation trainer “computer game”. Their training goal is not only to reach the ”high performance” level above

70%, but finally to reach the level without using video in advance of pronunciation of the test sequence words. On the basis of the data analyzed in this report it seems in comparison less important whether the students in the difficult tests both use text and audio, text alone, or audio alone. However, the students should be encouraged to exploit these differences themselves.

The students should preferably train the full recommended syllabus, i.e. 4500 words meaning that all words in the pronunciation trainer have been pronounced 10 times.

The students should try to train the recommended syllabus during around 10 weeks. This should be feasible as a part of their Danish learning. As appears from above the analyzed students have spent comparably longer time for their training than corresponds to carrying through the full syllabus in 10 weeks.

As regards personal student instruction at the training sites we saw in Section 2.2.1 big differences between training sites where many students had superficially used the pronunciation trainer, whereas only few had used it intensely enough to deliver data for our data analysis, and training sites where as many as 57-67% of the students had delivered usable data for the analysis. An obvious reason is that the training of the students has been taken good care of at the latter mentioned training sites, and the students have been motivated to continue using the pronunciation trainer.

It is evident that we are still at a very early stage in gathering and analyzing experiences of using the pronunciation trainer in practice in the field. However, it seems at first obvious, and as already mentioned probable, from the data we have seen so far that the pronunciation trainer in the future should be integrated fully in the Danish teaching of the training sites, which would probably result in a more uniform and effective motivation of the students to use the pronunciation trainer intensely in their self-training.

2.5.7.2 The pronunciation trainer system

The work on modifying the pronunciation trainer so that the students themselves can inform the system whether they want to train or test has started. This will facilitate the analysis of their results at the same time as the students themselves will get an even better feedback on their progress than is the case with the present version of the system. We are in this connection also considering giving the students the choice between more different test conditions when they choose to self-test. Again, this would emphasize the importance of self-testing under various conditions, and it would hopefully give us better data about the quantitative progress effects when a particular student is self-testing under clearly different conditions. As we have observed many students are self-testing under various conditions, such as (read *and* listen *and* a little bit of video). These varied test conditions complicate the subsequent analysis of the students' progress.

2.6 Conclusions

The first field test with the pronunciation trainer has given a big and very valuable material about the use of the pronunciation trainer in practice at the training sites as well as the real usefulness to improve Danish pronunciation of single words with persons who have another mother tongue than Danish.

We can conclude that almost all students at the training sites have carried through serious and determined training and that they have been very conscious about the effect on their test performance of testing under various test conditions. This shows that the pronunciation trainer

actually works in practice, and makes it probable that the pronunciation trainer supports the students' motivation to train with the system.

We can also conclude that the students in general have trained insufficiently to give us the ideally required benefit of the evaluation of the pronunciation trainer's use and usefulness at the training sites. On average the students have carried through around 1/4 of the full training syllabus of 4500 words. As all training sites by and large were up and running by 1.2.2005 our data material represents at least about 12 weeks of data from the training sites being able to deliver. However, many students stopped their training much earlier than after 12 weeks. As we do not have direct access to the students at the training sites we must in the future find other means of trying to intensify the training of the students. In this connection it must be mentioned that the field training with the pronunciation trainer continued till the summer of 2005, so we hope by then to have received the required material about (almost) fully implemented training periods from at least some of the students.

It is now clear that the students are making good progress with the pronunciation trainer. On average the 16 students in the main group who do not already speak Danish at a high level, have achieved a progress of at least 16.4% in absolute percentage points above their start score, and with an average training effort of about 20% of the full recommended training syllabus. This average covers very big individual differences in the progress of the students, meaning that many students have improved much more than it appears from the average figures. We consider it to be a conservative estimate that the average progress of the students using the pronunciation trainer with a considerably bigger training effort would be able to achieve at least 40 absolute % points above their start test score, so that a student who in the start test achieves a score of 30% in the end test will achieve a score of 70% under the same test conditions as in the start test.

We are of the opinion that the results presented here give a quite convincing basis for assuming that the pronunciation trainer is a useful means of self-training Danish pronunciation for persons with another mother tongue than Danish. We have to make the reservations are that would like to see results from a sufficient number of students who have carried through the full training syllabus and that we would like to look at a number of contexts between the progress of the students during training and their personal data, including the use of the listening trainer.

With these reservations we believe that this report establishes a justifiable basis for continuing the pronunciation trainer project concerning Danish pronunciation training at sentence level.

3 DialogDesigner – A Tool for Rapid System Design and Evaluation[1]

Hans Dybkjær
Prolog Development Center A/S
H. J. Holst Vej 3C-5C
2605 Brøndby, Denmark
dybkjaer@pdc.dk

Laila Dybkjær
Natural Interactive Systems Laboratory
University of Southern Denmark
Campusvej 55, 5230 Odense M
laila@nis.sdu.dk

Abstract

As spoken dialogue systems mature, the need for rapid development tools increases. We describe such a tool that is currently being used for commercial design, specification and evaluation, and that is in the process of being developed into a complete case tool.

3.1 Introduction

Improved recognition and understanding of spoken interaction facilitate the development of higher level tools that may enhance the clarity of spoken dialogue systems (SDSs) and reduce their development time and cost. This paper describes a tool – named DialogDesigner (see also www.spokendialogue.dk/DialogDesigner) – which supports SDS developers in rapidly designing and evaluating a dialogue model. In the following Section 3.2 provides an overall description of Dialog-Designer. Sections 3.3, 3.4, 3.5 and 3.6 present different aspects of the tool functionality in terms of how to model the dialogue, get various graphical views, run a Wizard-of-Oz (WOZ) simulation session, and extract different presentations in HTML. Sections 3.7 and 3.8 describe related work on design and evaluation tools and development tools, respectively. Section 3.9 concludes the paper.

3.2 DialogDesigner

The basis in DialogDesigner is the design window where one can enter and browse a dialogue model, including prompts, conditions, and state transitions. Having entered a dialogue model there are various presentation possibilities.

One option is to view a graphical presentation of the dialogue model. This presentation can be made more or less detailed depending on what the designer wants to see. A second option is to run a WOZ simulation. This can be done with users or as part of presentations to and discussions with customers. The simulation is logged and can be saved for later analysis and commenting. The simulation log can also be used normatively to generate test scripts for use in a systematic functionality test. A third option is to extract HTML versions of the entire dialogue as well as of prompt and phrase lists.

In the following we explain the design window and the three mentioned main options, and illustrate the tool via the early design of a pizza application.

[1] This paper was published in the Proceedings of the Sixth SIGdial Workshop on Discourse and Dialogue, Lisbon, Portugal, 2005, 227-231.

3.3 Dialogue Structure and Prompts

The design window (Figure 1) has at its top three fields for administrative purposes (name of application, version and note) (1). The rest of the window concerns application design. The designer starts by entering a new group (2). A group consists of one or more dialogue states which conceptually belong together and are described by the group. A group or a state can be moved up or down in the emerging dialogue structure (3) using the arrow buttons (2). New states are entered at (4). Here one can also indicate if there is any priority condition (conditions are numbers, not Booleans) for entering the state, grammars needed for this state, and parameters that can be tested in conditions on states or transitions. No grammars are needed if the state does not take input from the user but continues directly to another state.

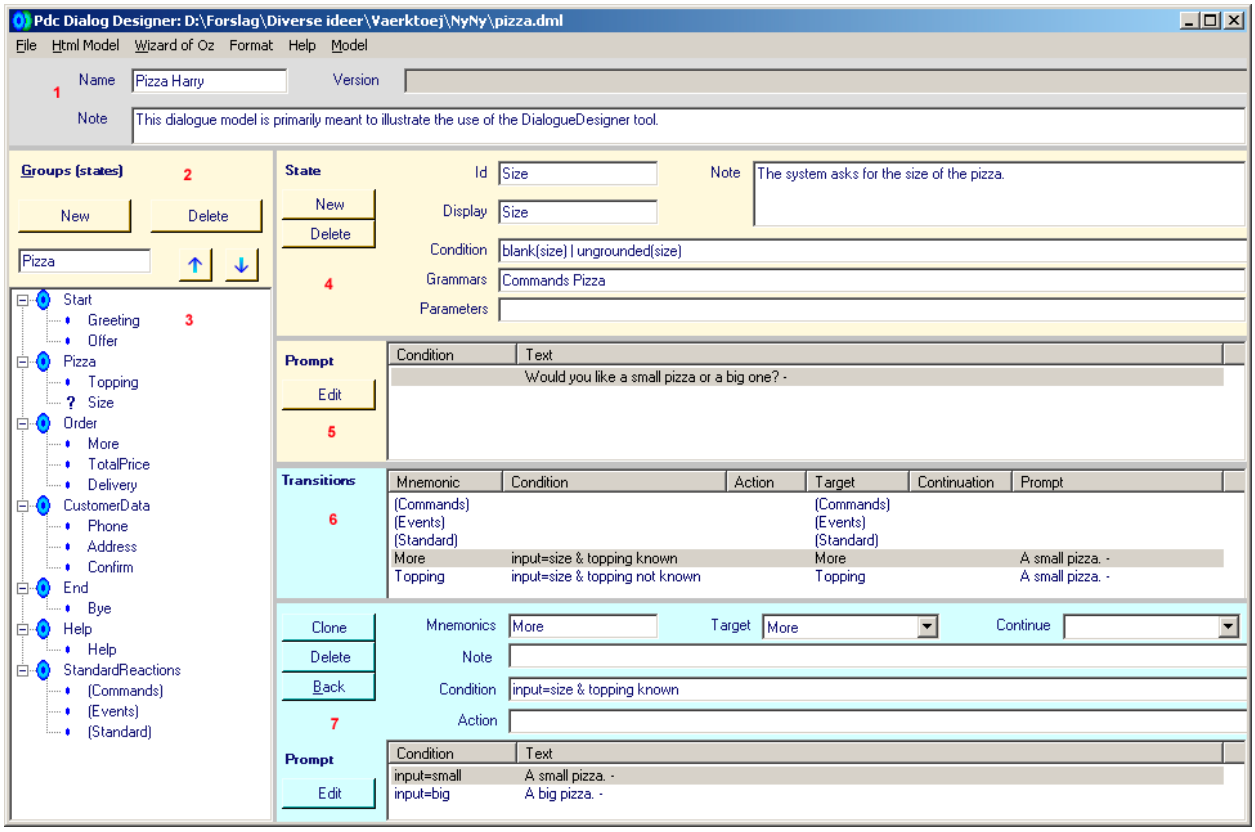


Figure 1. The design window. Red numbers are referenced in the text.

A state usually has one or more prompts attached. These are entered by clicking “edit” at (5). This leads to a window (not shown) listing all phrases already entered. New phrases can be added and one can compose a prompt for the state by selecting one or more phrases or named sets of indexed phrases and storing them. The resulting text is then shown at (5) when one returns to the design window.

To get from one state to another, transitions (6) are needed. Some transitions are globally enabled when input from the user is expected. These may include e.g. request for repetition and no input registered. When there are several such global transitions it may pay off to group them together as done in Figure 1 under the group StandardReactions. Here (Commands) contain user-initiated meta-communication commands, such as help and repeat, while (Events) contain system triggers for meta-communication, such as no input and nothing understood. (Standard) contains default domain value reactions such as price information which the user may request at any time during the dialogue. A state may have several possible transitions

leading to different new states (targets) where the choice of transition depends on the user's immediate input or on which information has been achieved so far. Transitions may target states or groups of states. In the latter case state conditions will determine which state to enter. Conditions on transitions express what must be fulfilled in order to select them. Transitions may also be accompanied by a prompt e.g. to provide feedback on the user's input or bridging to the output for the next state.

Transition information is entered at (7) where clicking on clone will enable the designer to enter a new transition. Transition prompt texts are entered in the same way as state prompt texts, as explained above.

3.4 Graphical View

Clicking Model in the top menu bar in the design window (Figure 1) opens a new window which allows the designer to see various graphical views of the dialogue (Figure 2). The graph part (7) is empty when the designer opens the window. To the left (1) are the groups and states specified in the design window. To the right (4) the designer can choose what he wants the graph to show. This should be done before he starts drawing the graph. Ticking Domain will enable all domain, i.e. task-related, transitions to be drawn. Ticking Command and System, respectively, will enable meta-transitions to be shown where System covers meta-transitions triggered by system events and Command covers user-initiated meta-transitions. Incoming and Outgoing allow the designer to see incoming and outgoing transitions, respectively, for a group or a state. Local shows transitions going out of and coming into the same state. Via shows transitions to a state that by default continues to some other state. Whenever the designer ticks one of the options Via, Incoming, Outgoing and Local, and selects a group or a state, the Outgoing (5) and Incoming (6) lists will show the transitions that will be drawn, if any.

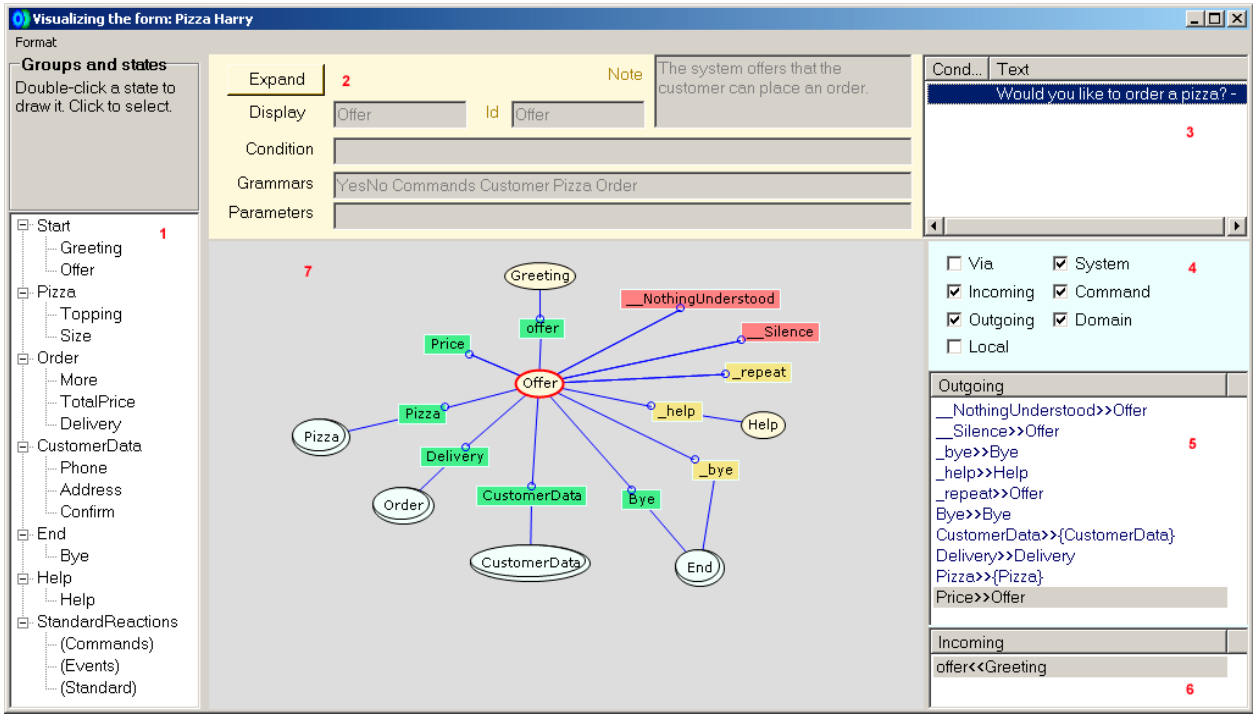


Figure 2. The graphical view. Red numbers are referenced in the text.

To draw a group or a state in the graph part of the window (7) one must double-click the group or state at (1). Groups are shown in a double ellipsis to indicate that they can be further expanded, while states are drawn in a single ellipsis. The ellipsis of a selected group or state is shown in red. To expand a selected group or a state and see its transitions as specified at (4) one must click the expand button at (2). To collapse a group again one must double-click the group at (1). Domain transition labels are green while system transitions are red and command transitions are yellow.

The graphical view is well-suited to get an overview of the dialogue structure and see connections at a more or less fine-grained level.

3.5 Wizard of Oz

In the design window (Figure 1) one may select “Wizard of Oz” -> “Woz” from the menu bar. Doing this opens a new window as shown in Figure 3. This window enables the designer to simulate a user-system interaction using the designed dialogue model.

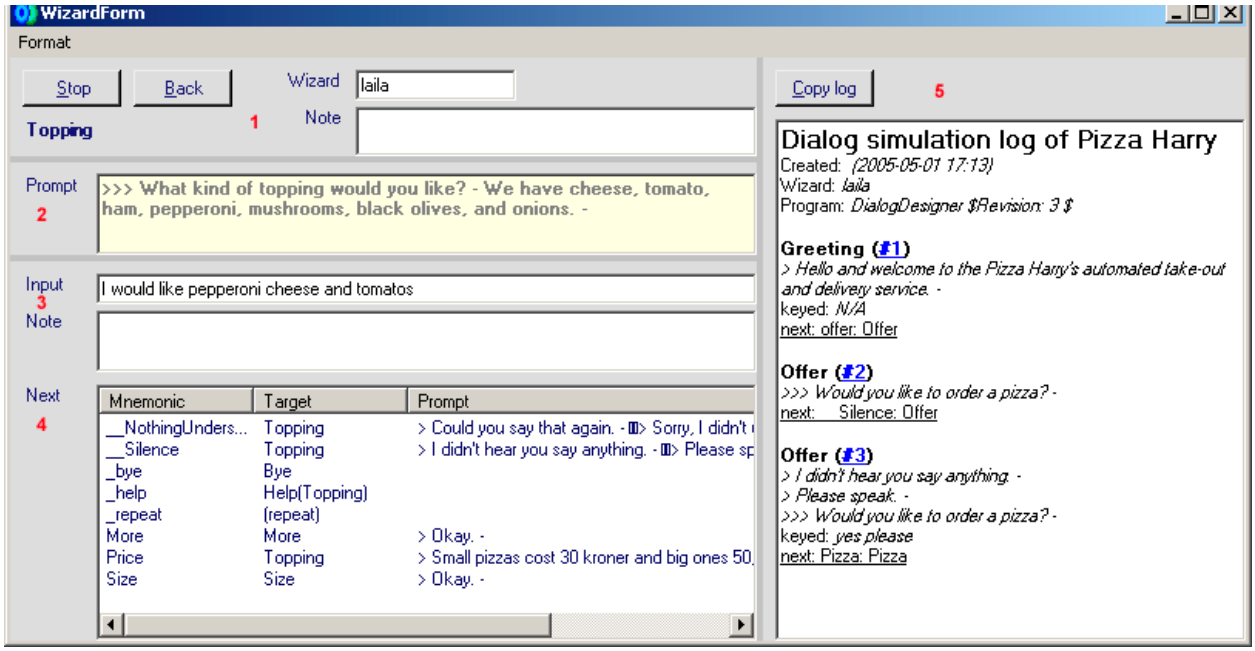


Figure 3. The simulation window. The log is stored in XML and may later be analysed in a similar window, or the RTF-format in the right-most pane may be copied to another document. Red numbers are referenced in the text.

The designer starts a dialogue by clicking Start (1, where the button now is labelled Stop because a dialogue is ongoing). This will cause the system utterance for the initial state to be displayed in the Prompt field (2). At the same time all possible transitions from this state are shown in the Next field (4). Which one to choose depends on the user’s input which is entered at (3). Entering the user’s input does not automatically cause a selection of a transition. This must be done manually. But writing down the user’s input means that the log eventually will contain a full dialogue with both system and user utterances. Such dialogues may later be used for testing the application and for further analysis. At (3) it is also possible to write notes to the current dialogue state, user input or transition.

The designer selects a transition by double-clicking on it. In doing this the previous system and user turn will be displayed in the log field at (5). At the same time the next system prompt

is shown in the Prompt field and the new transition possibilities are shown in the Next field. The designer may copy and save a log for later inspection in the analysis window.

The analysis window is opened from the design windows menu bar “Wizard of Oz” -> “Edit logs”. This window looks quite similar to the Woz window but supports the designer in inspecting, editing and commenting a previously saved log from a simulated interaction.

3.6 HTML Presentations

The HTML menu in the design window (Figure 1) gives access to a number of options for HTML presentations.

Phrase and prompt lists and a presentation of the dialogue model may be extracted in HTML. These are helpful for communicating with customers and phrase speakers. The HTML dialogue model can be used for navigating the dialogue via links, cf. Figure 4, without having access to the DialogDesigner.

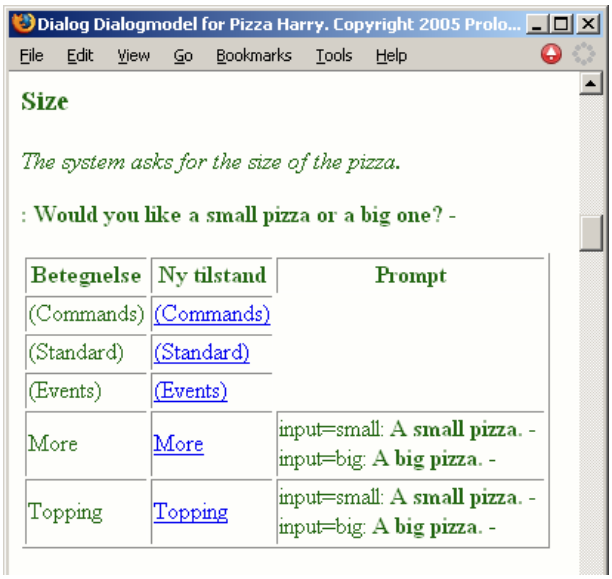


Figure 4. Excerpt of HTML presentation.

3.7 Related Design and Evaluation Tools

Other tools than DialogDesigner exist which are meant to support the design and evaluation of SDSs and which support WOZ. Two such tools are Suede [Klemmer et al. 2000], developed at the University of Washington, and the WOZ tool developed by Richard Breuer [WOZ tool] as a by-product of his work at Scansoft.

Suede offers an interface for each of the three main activities of design, test, and analysis. The design inter-face allows the designer to create example dialogue scripts and a design graph representing the general de-sign solution. For each prompt the audio output may be played if it has been recorded. The test mode enables WOZ simulation. The designer selects a prompt from a list of available prompts given the present state. The selected prompt is played to the user. Based on the user’s answer the designer selects again one among the now available prompts, etc. Simulation of recognition errors is supported. The analysis interface is similar to the design interface except for the top of the window which contains user audio input from the last session. Moreover the design graph is annotated with test data which can be played.

The WOZ tool developed by Richard Breuer offers interfaces for the three main activities of design, WOZ simulation and export. In the design mode the designer can specify the dialogue

design in terms of prompts, questions and concepts. Like in DialogDesigner but contrary to Suede this interface is textual and not graphical. However, one has - like in DialogDesigner - the option to view a graphical version of the designed dialogue model. In WOZ mode the designer chooses the output to the user from a list of possible next prompts or questions depending on the user's input. The export activity is facilitated from a menu point in the design window. There are several export possibilities, including export to XML, HTML or HDDL (a proprietary programming language used by the SpeechMania platform [Aust et al. 1995]).

Figure 5 gives a rough comparison of which features are included in DialogueDesigner, Suede and Woz tool.

Feature	DialogDesigner	Suede	Breuer	HotVoice	SpeechMania
graph view	+	+	+	+	-
graph design	*	+	-	+	-
structured prompts	+	-	-	(+)[2]	+
record prompts	-	-	(+)[4]	-	+
play prompts	*	+	+	+	-
speech recognition	-	-	(+)[4]	(+)[3]	(+)[3]
log analysis	+	+	-	-	-
regression test	*	-?	(+)[4]		+
debug	-	-	(+)[4]	+	+
WOZ	+	+	+	-	-
make test scripts	+	(+)[5]	-	-	-
phrase list	+	-	(+)[4]	-	+
prompt list	+	-	-	-	-
code generation	*	-	+	+	+
standard dialogues	-	-	-	(+)	+
state conditions	+	-	-	-	+

Figure 5. Tool comparison. +: Has feature. -: Does not have feature. ?: Unknown, *: In pipeline

3.8 Related Development Tools

IVR tools extended with recognition facilities, such as HotVoice from Dolphin and Edify, may also be seen as related work. Both these examples offer a graphical interface for dialogue flow design. In addition HotVoice also offers the possibility to edit the program text generated via the graphical interface or write the design directly in the HotVoice language. The language used by HotVoice as well as the one used by Edify are proprietary languages just like HDDL. A major difference be-tween DialogDesigner and the IVR tools is that the possibilities for designing a dialogue using an IVR tool are fairly low-level. IVR tools are fine

[2] Must be coded.
[3] Has recognition as part of the running system but recognition cannot be tested during simulation.
[4] By using SpeechMania tools on generated code.
[5] Sound must be transcribed.

for specifying dialogues as a flow diagram. However, it would be difficult to use them for the design of complex dialogues.

Spoken dialogue platforms such as SpeechMania, Envox 6 VoiceXML Studio (both also support IVR), OpenSpeech, and the CSLU Toolkit are more aimed at implementation. To different extents they offer tools like “standard dialogues” for “best practices” in user interface design, such as entering a pin code.

However, common to these tools is that they focus on the implementation rather than on the modelling and evaluation – they are not case tools. And they do not focus on presentation to customers and users.

3.9 Conclusion and Future Work

We have described DialogDesigner which is a tool in support of SDS dialogue design and evaluation. It focuses on communication and modelling flexibility as argued in [Dybkjær and Dybkjær 2004]. The HTML extracts, graph views and simulation mode provide strong support for communication with customers and domain experts which is important in real-life projects. The ability to place conditions on states, transitions and prompts provides a useful flexibility in dialogue modelling.

Three next tool development and extension steps are planned. They include features for enhanced design process support (cf. Figure 5) as well as implementation support (code generation), transcription, and synthesis. Code generation will allow the automatic generation of VoiceXML code based on the design description presented above. Automatic code generation has the potential to save considerable effort. However, it will be a challenge to flexibly support e.g. agent or problem solving approaches. For transcription we envision a tool comparable to the TranscriptionStation included in the SpeechMania platform. It requires that spoken input is recorded and that the recognised utterances are used as the basis for the transcription process. The synthesis extension must allow the user of DialogDesigner to either record output phrases for use in system simulations or use speech synthesis for the same purpose.

References

Harald Aust, Martin Oerder, F. Seide and V. Stenbiss: The Philips Automatic Train Timetable Information System. Speech Communication 17, 1995, 249-262.

CSLU Toolkit: http://cslu.cse.ogi.edu/toolkit/

Hans Dybkjær and Laila Dybkjær: Modeling Complex Spoken Dialog. IEEE Computer, August 2004, 32-40.

Edify: http://www.edify.com/

Envox: www.envox.com

HotVoice: www.dolphin.no

OpenSpeech: http://scansoft.com/products/

Scott R. Klemmer, Anoop K. Sinha, Jack Chen, James A. Landay, Nadeem Aboobaker, and Annie Wang: SUEDE: A Wizard of Oz Prototyping Tool for Speech User Interfaces. CHI Letters, The 13th Annual ACM Symposium on User Interface Software and Technology: UIST 2000. 2(2): 1-10.

WOZ tool: http://www.softdoc.de/body/home.htm

4 Usability evaluation issues in vision-based systems

Pedro Correa

Alterface, Belgium

More than any other computing paradigm, virtual environments involve the user – his senses and body – in the task. Thus, it is essential that we focus on user-centric performance measures. If a vision-based system does not make good use of the skills of the human being, or if it causes fatigue or discomfort, it will not provide overall usability despite its performance in other areas.

A large part of standard usability evaluation methodology can still be useful while evaluating novel HCI systems, thus a certain adaptation is often necessary. Testbed evaluations are still appropriate.

The standard evaluation chain is also still applicable:

- Single variable **hypothesis** generation.
- **Creation** of a very simple **application** intended to recreate the same application conditions, but focused only in the interaction (no effects, no background, no sound).
- **Testbed evaluation**. This one is the main evaluation point, and different approaches exist, we describe them below.
- **Result analysis**. The testbed evaluations will give us the needed data in order to confirm or refute the given hypotheses.

4.1 Testbed Evaluation

Testbed evaluation can be quite useful for gaining an initial understanding of interaction tasks and techniques, and for measuring the performance of various techniques in specific interaction scenarios (Doug A. Bowman's Thesis: *Interaction Techniques for Common Tasks in Immersive Virtual Environments: Design, Evaluation, and Application*).

In practice, these are some examples of aspects that can be taken into account when dealing with novel HCI techniques:

- The Learning curve: Time needed for the user to perform a first useful action.
- Fatigue: number of interactions a user can conduct one after the other without being tired/fed-up.
- Friendliness: The ease of use of the particular interaction mode, its compatibility with the physical abilities of the user.
- Effectiveness: The accuracy and completeness with which specified users can achieve specified goals in particular environments.
- Efficiency: The resources expended in relation to the accuracy and completeness of goals achieved, notably overall time spent.

We will now present some evaluation methods that have been proven to be useful up until now, and how they could be adapted to state-of-the-art vision-based systems. In the context of this report, we will focus on *user-based methods only*[6].

6 , following the IST 2002- 507382 EPOCH project report: Evaluation and Design Methodologies

We will not take into account *usability inspections methods*, which are the generic name for a set of methods based on having expert evaluators inspect or examine usability-related aspects of a user interface.

4.1.1 User-Based methods

User-based methods mainly consist in user testing where usability properties are assessed by observing how the system is actually used by some representatives of real users (Whiteside J. et al.:1988) (Dix A. et. al.: 1998). User-testing evaluation provides the trustiest evaluation, because it assesses usability through samples of real users. However, it has a number of drawbacks, such as the difficulty to properly select correct user samples and to adequately train them to manage also advanced functions of a web site (Matera M. et al.: 2002). Furthermore, it is difficult, in a limited amount of time, to reproduce actual situation of usage. This condition is called “Hawthorne effect” (Roethlisberger et al.: 1939): if the variable of the experiment are manipulated, it is possible that the productivity of the group observed decreases. Failures in creating real-life situations may lead to “artificial” conclusions rather then realistic results (Lim K.H et al.: 1996). Therefore, user-testing methods are considerable in terms of time, effort and cost. User testing is the main way for evaluating right away the look and feel of the interface, as it is possible to verify at “real-time” the reactions of the users.

Within the category of user-testing methods there are several techniques, the most important being:

- Thinking aloud
- Contextual inquiry
- Interview

Focus Groups techniques are also often used, yet we will not discuss it in this context, since they are not the most appropriate approach in order to discuss interaction techniques.

4.1.1.1 Thinking aloud

During the thinking aloud test, the user should think aloud while performing some specific task with the system. By verbalizing his thoughts, the user allows the observers to know his opinions and feeling about the application. Verbal protocols are recorded concurrently or retrospectively. The subject is probed to verbalise problems that come up. After the recording of verbal protocols, the protocols are encoded according to a previously defined encoding scheme. Verbal reports can be interpreted if the processes by which they were generated are understood. Interpretation is based on the theory that human cognition is information processing (Newell & Simon 1972, Simon 1979). Cognitive processes and their structure account for the results of verbalisations. The accuracy of verbal reports depends on the procedures used to elicit them and the relation between the requested information and the actual sequence of heeded information.

Thinking aloud allows you to understand how the user approaches the interface and what considerations the user keeps in mind when using the interface. If the user expresses that the sequence of steps dictated by the product to accomplish their task goal is different from what they expected, perhaps the interface is convoluted.

4.1.1.2 Contextual Inquiry

Contextual Inquiry is a specific type of interview for gaining data from the user. This technique aims at understanding the context in which the application is used. Contextual Inquiry (also known as "site visits") is basically a structured technique of observing and

interviewing users. It is based on the core principle that understanding the context in which a product (or service) is used (or the work is being performed) is essential for user and customer oriented design. Using contextual inquiry, you visit the workplace of prospective users to see how they work. You observe all aspects that would help define a context for their work - and thus a context for the usage of your product or service.

Contextual Inquiry is adequate in situations where the subject domain is unclear or unfamiliar to the development team, and when the context of work may have a significant effect on the new product or service. For performing a Contextual Inquiry considerable investment of time and effort may be needed in order to elicit sufficient information from the users and the environment to be studied.

Contextual Inquiry follows many of the same process steps as field observations or interviews.

Contextual inquiry is best done by a group of researchers who develop a medium- to long-term relationship with a group of organisations that are interested in providing data. According to Holtzblatt et al. the relevant steps are the following:

- "Identifying the customer: identify the groups that will be using the new technology or are using similar technology, and arrange to access organisations within the groups that give a cross section of the (potential) market.
- Arranging the visit: write to the targeted organisations identifying the purpose of the visit, a rough time-table, and how much of the employees time will be taken up by the exercise. Ensure that some feedback from the day is possible before leaving. Ensure that the participating organisations understand how many visits you intend to make over the time period of the evaluations.
- Identifying the users: a software product will affect many people throughout the organisation, not just the management or the end users. Ensure that you understand the key users in the organisation whose work will be affected by a new system or changes in the current one.
- Setting the focus: select what aspects of the users' work you wish to make the focus of each visit, and write down your starting assumptions. Make a statement of purpose for each visit, and after the visit, evaluate to what extent you have achieved your purpose.
- Carrying out the interview / observation: stay with the selected users until you have managed to answer the questions you have raised in 'setting the focus'. Very often this may involve inviting the user to directly share and comment on your notes and assumptions.
- Analysing the data: the process of analysis is interpretative and constructive. Your conclusions and ideas from one round of observations are input to the next round, and an evaluation of the results so far should be one of the purposes of subsequent visits."

4.1.1.3 Interview

Interview is an informal technique for the investigation of the users' opinions about the application, e.g. subjective satisfaction, critical incidents, anxieties which are hard to measure objectively. It is a useful method for studying what features of the application users particularly like or dislike.

Three types of interviews can be distinguished: unstructured, semi-structured and structured interviews. The type, detail and validity of the collected information vary with the type of interview.

The validity of results varies with the experience of the interviewers. The interviewer needs domain knowledge in order to ask the right questions and there is always the risk of bias in

what questions the interviewer asks and how the interviewee interprets them. Besides, Interviews are demanding in terms of the number of representative users needed. It is preferable to use questionnaires where possible. Because of the unstructured nature of an interview the result is just a report summing up the comments made by the subject in the interview.

References

Doug A. Bowman's Thesis: Interaction Techniques for Common Tasks in Immersive Virtual Environments: Design, Evaluation, and Application.

IST 2002- 507382 EPOCH project report: Evaluation and Design Methodologies (March 2005).

5 Results of iterative framework testing in haptic-based applications

Giorgos Nikolakis and Dimitrios Tzovaras

ITI-CERTH, Greece

5.1 Introduction

Usability in haptic-based systems involves several issues such as anatomy, physiology, psychology and design. Usability evaluation is performed to ensure that products and environments are comfortable, safe and efficient for people to use, as well as to ensure that a product fits the target users' needs. This section aims to provide information concerning the evaluation of a haptic-based system by identifying the criteria and methodologies used during the various steps of the application development, describing the results during each step.

Weaknesses of the evaluation criteria used for the usability evaluation of haptic-based systems have been identified in D18: "Current practice description and evaluation: Towards a framework for usability evaluation". A usability evaluation framework was proposed in order to resolve these ambiguities in the results. The framework involves evaluation steps at different development stages using sets of criteria and involving end users.

5.2 Usability Evaluation procedure

This section describes the usability evaluation procedure using the framework proposed in D18. Results produced by using the evaluation procedure are studied through the example of the usability evaluation followed for the development of a cane simulation application.

5.2.1 Preliminary usability evaluation

Before starting the development of an application the end user' s needs have to be identified in order to specify the user requirements. The preliminary usability evaluation of an application aims to **identify** whether a **haptic application can satisfy the user requirements** and to **define** the **hardware equipment** and **software tools** to be used for the application development.

Ideally during this step the users testing the system should belong both to the end users group and to have some experience on haptic devices and application development. This is important because in this step the users are actually asked to identify, if specific hardware and tools can be used to satisfy the user requirements.

The cane simulation application involved the use of a haptic glove as an interface to provide force feedback to the end user. Even before starting the development of the application the Cybertouch™ glove was evaluated using demo application provided with the devices. User's comments proved that the use of this haptic glove was not suitable for the cane simulation application. Specifically the users stated that it is not sufficient to feel vibration to their fingers when the virtual cane hits an object and they mentioned that the haptic device should provide forces to the users hand. This led to test the combination of Cyberglove and Cybergrasp devices, which were finally used to develop the application.

Number of user involved: 2

Evaluation methodology / criteria: Psychophysical criteria.

Redesign issues: CyberGrasp and Cyberglove replaced Cybertouch haptic glove.

5.2.2 Usability evaluation during development

The preliminary usability evaluation has identified the hardware equipment and software tools to be used for the development of a haptic application. The evaluation during the main development stage aims to **identify weaknesses of the software** under development and to **define directions** for the **improvement** of the application usability. This part of development and evaluation is repeated until the application satisfies the user group.

During this step the evaluation involved testing a very first version of the system. A simple test case was developed. The application included a set of candidate versions for the force feedback. The users were asked to evaluate the system and select which of the available candidates fits better their needs and to comment on how it could be improved. This was repeated several times until all the users select the same candidate.

Number of user involved: 4

Evaluation methodology / criteria: Psychophysical criteria.

Redesign issues: Approval – Disapproval of force feedback algorithms.

5.2.3 Final usability evaluation (beta – testing)

Pilot application usability evaluation is considered sufficient for applications developed for research purposes. Evaluation of a pilot application of a haptic system is usually performed in a controlled environment. This allows identifying usability issues, however it is not sufficient to detect problems that may occur in a real user environment. For this reason it is important to perform usability evaluation of a system using a large number of users. The aim of this evaluation is to **identify errors** and **bugs** in the application that cause irregular results and thus reduce the usability of the application and **provide** the developers with sufficient **information** in order **to fix these errors**.

Number of user involved: 26

Evaluation methodology / criteria: Pre-test and after test questionnaires, Time to perform the tasks and Psychophysical criteria.

Redesign issues: The evaluation of the pilot application showed that the system is usable and can be useful for cane training. However it pointed out issues that would be useful to be improved in next versions of the system. Also a set of bugs were identified and fixed during the evaluation period.

An analytical description of the evaluation is presented in sequel.

5.2.3.1 Test – Cane Simulation

Two test cases were designed for the evaluation of the cane simulation environment, an indoor and an outdoor application. Both test cases were evaluated using three different setups, namely:

- Multimodal (Both haptic and audio feedback),
- only sound,
- only haptics (in the outdoor only the traffic light sound was active).

The two test cases are described in the sequel.

Outdoors test case

The user is asked to cross a traffic light crossing using the virtual cane. The user is standing at the beginning of the test room wearing the CyberGrasp™ and a waistcoat for carrying the

Force Control Unit (FCU) for the CyberGrasp™. When the test starts, the user is asked to grasp the virtual cane. The parameters of the virtual cane (size, grasping forces, collision forces) are adjusted so that the user feels that it is similar to the real one. After grasping the cane, the user is informed that he/she is standing in the corner of a pavement (shown in Figure 1). There are two perpendicular streets, one on his/her left side and the other in his/her front. Then, he/she is asked to cross the street in front of him/her.

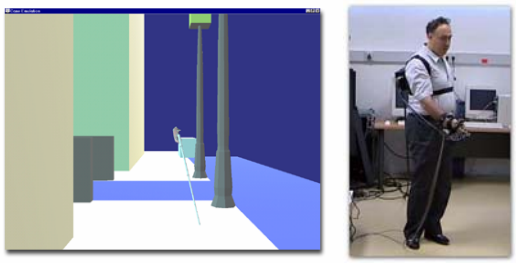


Figure 1: Cane simulation – Outdoors test – a) Virtual set-up, b) A user performing the test.

The user should walk ahead and find the traffic light located at about one meter on his/her left side. A realistic 3D sound is attached to the traffic light informing the user about the condition of the light. The user should wait close to it until the sound informs him/her to cross the street passage (green traffic light for pedestrians). When the traffic lights turn to green the user must cross the two meters wide passage until he/she finds the pavement at the other side of the street. It is also desirable that the user finds the traffic light at the other side of the street.

The specific test was considered 100% succesful if the user could observe all features in the virtual environment (i.e. find the traffic light at the beginning and end of the test, distinguish the difference between the pedestrial street and the road) and react accordingly (wait until the traffic light switches to green, and pass the street following a straight path) within a specific time frame (3 minutes).

Indoors Test case

The user is asked to navigate into an indoor environment using a virtual cane. Sound and haptic feedback are provided by the system upon collision of the cane with the virtual objects. The user is standing at the beginning of the test room wearing the CyberGrasp and a waistcoat for carrying the Force Control Unit (FCU) for the CyberGrasp. When the test starts, the user is asked to grasp the virtual cane. The parameters of the virtual cane (size, grasping forces, collision forces) are adjusted according to the characteristics of his/her cane. The goal for the user is to find the second door on his/her left side and enter the room (Figure 2). There he/she should find a chair. During his/her walk the user should find successively the wall on his left side, the first door where he/she is not supposed to enter, the wall of the second room and the door where he/she is supposed to enter. After entering the room he/she should find the chair located in his right side.

Figure 2: Cane simulation – Indoors.

5.2.3.2 Evaluation methodology / criteria

Twenty-six persons from the Local Union of Central Macedonia of the Panhellenic Accosiation for the Blind in Greece have participated in the tests. The users were selected so as to represent the following groups: blind from birth, blind at a later age, adults and children.

The 26 participants (14 male and 12 female) went through the evaluation tests program. The average age was 32.8 years, the youngest participants were 19 years old and the oldest 65 years old. 42% of the participants were blind from birth, 14.3% went blind before school age (1-5 years of age), 23.4% during schooldays (6 to 17 years) and 16.3% after school time or late youth (17 to 25 years). Also, 40% of the persons tested were students, 28% telephone operators, 12% unemployed, 6% teachers and 2% professors, librarians, educational coordinators and computer technicians. Finally, 38% of them knew about haptics and 24% had used a similar program.

The expectations of the users from the program were identified as follows: “visit new places”, “explore scenes” and “play a new game”. Some of the participants did not reply at all, others did not have any idea what to expect from the program.

The users were introduced to the hardware and software a day before participating in the tests. The introductory training, took approximately half an hour per user. The motivations for this pre-test was to introduce the users to a technology completely unknown to them, while ensuring that they feel comfortable with the environment of the laboratory.

The main evaluation study took approximately half an hour per user, including pauses. The purpose of the evaluation was not to test the reaction of a user to a haptic system. Rather, the idea was to try to obtain information about the use of such a system by a user who is somewhat familiar with the use of haptics. During the test procedure, the tasks were timed and the test leader was monitoring the performance of the users.

Test	Average Time (min) Blind from Birth	Average Time (min) Non-Blind from birth	Overall Average Time (min)	Success Ratio (%)	Percentage of users needing guidance (%)	Average degree of challenge 1=very easy 5=very difficult
1	2.04	2.17	2.12	97,41	3,80	2.65
2	1,96	1,9	1.92	92,34	3.80	2.88

Table 1: Feasibility study test evaluation results.

5.2.3.3 Redesign issues - Evaluation Results

Based on the initial specifications derived by the end user requirements, the goals of the usability evaluation study were to show that the user can use the proposed system for : a) navigating in complex environments, b) edutainment and c) interacting with haptic user interface components.

The cane applications are focusing on simulating human navigation in a virtual world, naturally; using the same perceptual cues as they do when in real world situations.

Table 1 presents the parameters used for the evaluation of the prototype, such as the time of completion/test, success ratio, percentage of users needing guidance and degree of challenge (set by the users). Results from tests show that blind people can easily navigate in a virtual environment using a cane similarly to what they do in the real world. Cane simulation was considered to be a pioneering application and results have witnessed the acceptance of the users in terms of usability, realism and extensibility of the specific application.

According to the comments of the users during the tests and the questionnaires filled by the users after the tests, the following conclusions can be drawn: It was deemed very important to utilize both acoustic and haptic feedback, as they are indispensable for the orientation.

The ANOVA 0 method was used to compare the performance of different groups of users. Four different pairs of groups were identified, according to age, gender, blindness from birth or not and employment status of the users. The time needed to complete each test, was used in order to compare the performance of the different groups. The critical value for the parameter $F_{critical}$ of the ANOVA method was calculated to be equal to 4.25 (assuming probability equal to 0.05 and degrees of freedom between groups equal to 1 and within groups equal to 24). Two groups and 26 measurements were assumed in each case and thus parameters DFS and DFG were computed to be DFS=2-1=1 and DFG=26-2=24.

The age of the users did not seem to affect their performance. Although young users were expected to have a better anticipation on using the haptic devices, older users managed to perform equally or even slightly better than younger ones. Users over 25 years old performed slightly better in the cane simulation test. On the other hand, in the cane simulation test, female users performed slightly better. According to the ANOVA method F_{cane}=2.3, F_{object}=0.12 both significantly less than $F_{critical}$.

In general, results have shown that blind from birth users had similar performance to all other user categories. Blind from birth users had slightly increased difficulty in using the virtual cane. However, this cannot be considered of high importance in order to lead to conclusions relating user performance with blindness from birth. According to ANOVA F_{cane}=0.44 and F_{object}=0.97 which are significantly less than $F_{critical}$.

Finally, the statistical analysis has shown that all employed users had finished the tests successfully. This may be a result of the self-confidence that employed users impose. The ANOVA results do not show very significant difference between the means of the groups, but for the cane simulation test F_{cane}=3.29, which is relatively close to the $F_{critical}$ value compared to other cases, being, however, still less than $F_{critical}$.

The difficulty level of the tests was reconsidered after completion, according to the percentage of the users that needed guidance and the rank that users gave to each test case. The users were asked to rank the challenge of each test using a scale between 1 (easy) and 5 (very difficult). Both tests were considered by the users to be relatively difficult. The users needed guidance to perform the tests at a percentage of 3.8%. The average rates of the challenge of the tests, according to the users, 2.65 for the outdoor and 2.88 for the indoor cane simulation test.

5.3 Conclusions

In terms of usability, we can conclude that the system can be used for educational purposes, mobility and orientation training and exploration / navigation in 3D spaces.

The main advantages of the system presented over existing virtual reality systems for the training of the blind and the visually impaired is the capability to: a) support virtual training environments for the visually impaired with large workspaces (up to 7 m-diameter hemisphere), b) implement more natural user interaction with the virtual environments (using all fingers of the user's hand) and c) propose a novel cane simulation system.

Although the proposed system expands the state-of-the-art, there still exist important technical limitations that constrain its applicability. Specifically, the system cannot prevent the user from penetrating objects in the virtual environment. The maximum workspace is limited to a 7 m - diameter hemisphere around the tracker transmitter (the 1 m limitation, caused by the

CyberGrasp™ device is solved by using a backpack so that the user can carry the CyberGrasp™ actuator enclosure). The maximum force that can be applied is limited to 12N per finger and the feedback update rate is 1KHz.

Furthermore, the following conclusions can be drawn from the evaluation of the Feasibility Study tests in terms of system usability:

- It was deemed very important to utilize both acoustic and haptic feedbacks, as they are indispensable for the orientation. The majority of the participants preferred to have both feedbacks.
- Some of them would have liked to deal with more complex scenarios.
- All people tested had no problems with the system after an explanation of the technology and some exercises to practice the application.
- The participants needed little or no guidance at all, i.e. the users had no difficulties to handle the software and the devices. On the contrary, they enjoyed completing their tasks, showed a lot of commitment and were very enthusiastic about being able to have this experience.
- No connection was found between the age that blindness occurred and the test results.
- All participants emphasized their demand to use these programs in the future.
- Multimodal feedback is very important. Sound is considered to be complementary to haptics and vice versa. Users stated that using only haptic feedback made it non-realistic and harder to navigate in the VE, while using only audio feedback required higher mental effort.
- Mixed Reality system should be developed in order to simulate accurately the way that the user grasps and uses the cane. The use of each user's real cane in the mixed simulation environment could also be an option that would offer a solution to this problem.

Concluding, the result has unanimously been that the prototype introduced was considered very promising and useful, whereas it still leaves a lot of room for improvement and supplement. Provided that further development is carried out, the system has the fundamental characteristics and capabilities to incorporate many requests of the users for a very large pool of applications. The approach chosen, fully describes the belief of blind people to facilitate and improve training practices, and to offer access to new employment opportunities. It represents an improvement of life for the blind and the visually impaired people when connected to reality training. These facts are evident from the participant's statements.

References

Scheffe, H.: The Analysis of Variance. John Wiley & Sons, 1959.

6 Evaluation of Virtual Reality Surgical Simulators

Pablo Lamata, Samuel Rodríguez and Enrique J. Gómez

GBT, Universidad Politécnica de Madrid

6.1 Introduction

Surgical simulators have nowadays two main goals: training and skill assessment [1-3]. Technology and scientific knowledge are still immature for other applications like mission rehearsal or surgical credential. Usability is then understood as the capability of the simulator to achieve training or skill assessment with effectiveness, efficiency and satisfaction, a concept very close to validation.

The work reported here deals more with the efficiency aspect of usability, understood as the resources expended in relation to the accuracy and completeness of goals achieved. Other approaches aims to understand the human sensory and cognitive capabilities [4-6]. Virtual reality technologies offer a wide spectrum of possibilities to build devices for training and skill assessment. The question is which is the best approach to reach these goals? The approach taken has been to systematise the knowledge about the resources available in virtual reality simulation.

6.2 Conceptual framework

We propose a taxonomy for the different resources available in VR simulation, which is considered as a didactic means to meet different training needs, using several didactic resources. Basically these resources are defined and classified in three main categories: Fidelity, Virtual and Evaluation resources. Fidelity resources refer to the different levels of realism offered by a simulator in its interaction and behaviour. They can be further divided into sensorial, mechanical and physiological. Computer resources are features unique to a computer simulated environment that can enhance training, like cues and instructions given to the user to guide a task, or to manage a training program. Evaluation resources are metrics to evaluate performance, follow up progress and ways to deliver constructive feedback to the user. Several subcategories and items are defined inside these main three kinds of resources [7].

6.3 Comparing laparoscopic surgical simulators

Several laparoscopic simulators are currently available. These range from simple box trainers with standardized tasks to advanced VR simulators. All of them are designed to train laparoscopic skills, but they make use of different didactic resources. This section makes use of our proposed taxonomy to study these differences.

6.3.1 Materials and methods

The following VR simulators were considered for the study (see Figure 1): “Basic Skills” package and “Suture 3.0” of MIST-VR (Mentice AB, Göteborg, Sweden), “Basic Skills 2.0” package, “Dissection” and “Gynaecologic” modules, these last two considered together, of LapSim (Surgical Science Ltd, Göteborg, Sweden), virtual tasks of ProMis (Haptica, Dublin, Ireland), Reachin Laparoscopic Trainer-RLT (Reach-In, Stockholm, Sweden), and LapMentor

(Simbionix, Israel). In addition, two generic box trainers are studied, one with physical objects and the other with ex-vivo organs.

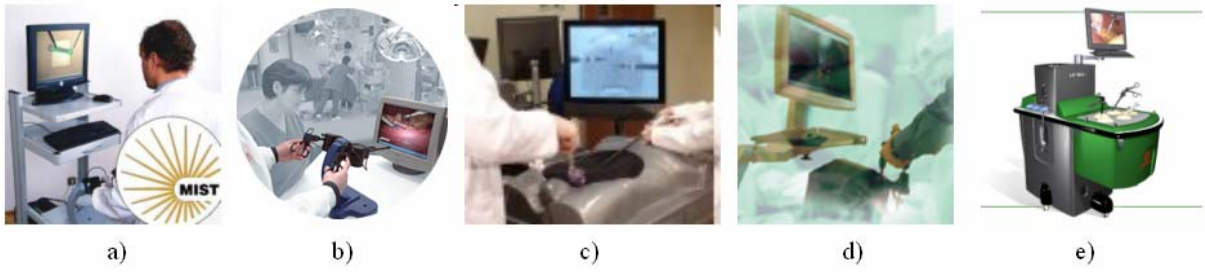


Figure 1. Commercial laparoscopic simulators analysed: (a) MIST-VR, (b) LapSim, (c) ProMis, (d) Reach-In Lap Trainer –RLT, (e) LapMentor.

We have applied our proposed taxonomy of didactic resources in laparoscopic VR in the following way. Fidelity resources were compared quantitatively, each fidelity component was studied and ranked on a scale from 0 to 10. Values were averaged in a percentage for each level of subcategories. Computer and assessment resources were only ranked as "used" or "not used", since definition of a scale addressing them is not straightforward. The chosen VR simulators were studied and an average use of these resources was calculated for each subcategory and for a global percentage of use.

6.3.2 Results

Figure 2 outlines the main results of the comparison. Different aspects of fidelity have been ranked from 0 to 10, and then averaged. Computer and assessment resources have been only counted, i.e. a 100% use of them means that the simulator uses all the different resources defined in the taxonomy.

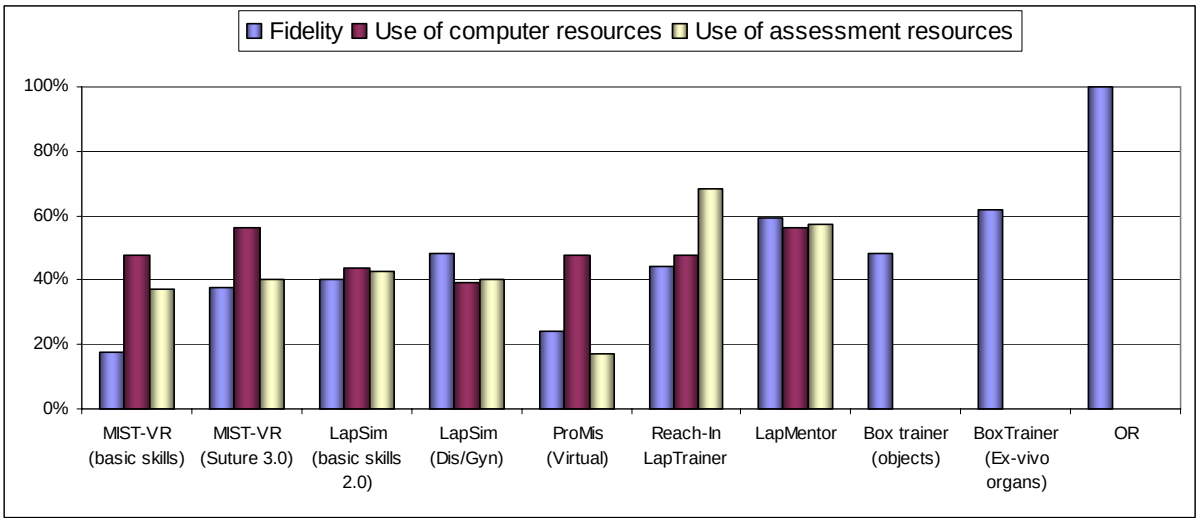


Figure 2. Fidelity and use of computer and assessment resources by laparoscopic simulators.

6.3.3 Discussion

Didactic resources offered by surgical simulators have been categorized as a framework to formulate and contrast hypotheses mainly about the efficiency, but also the effectiveness and satisfaction of surgical simulation.

This conceptual framework has been applied to compare how different laparoscopic simulators make use of available resources of virtual reality technologies. These results should now be compared to the training outcomes, leading to an estimate of the efficiency of the systems.

6.4 Conclusions

Proposed conceptual framework is a first step to assess the relationship between simulation design and training efficiency. Future research will concentrate on a thorough evaluation of the importance of different didactic resources to teach basic and more advanced laparoscopic skills. Randomized and blinded surgical trials where novice surgeons are trained in different ways is the best methodological approach.

References

[1] L.S. Feldman, V. Sherman and G.M. Fried, "Using simulators to assess laparoscopic competence: ready for widespread use?," *Surgery*, 135(1):28-42, 2004.

[2] K. Moorthy, Y. Munz, S.K. Sarker and A. Darzi, "Objective assessment of technical skills in surgery," *BMJ*, 327(7422):1032-1037, 2003.

[3] R. Aggarwal, K. Moorthy and A. Darzi, "Laparoscopic skills training and assessment," *Br J Surg*, 91(12):1549-1558, 2004.

[4] P. Lamata, E.J. Gómez, F.M. Sánchez-Margallo, F.Lamata Hernández, F.del Pozo and J.Usón Gargallo, "Study of consistency perception in laparoscopy for defining the level of fidelity in virtual reality simulation," *Surgical Endoscopy*, (in press), 2005.

[5] G. Picod, A.C. Jambon, D. Vinatier and P. Dubois, "What can the operator actually feel when performing a laparoscopy?," *Surg Endosc*, 19(1):95-100, 2005.

[6] P. Lamata, E.J. Gómez, F.M. Sánchez-Margallo, F. Lamata, M. Antolín Fernández, S. Rodríguez Bescós, A. Oltra and J. Usón, "Study of Laparoscopic Forces Perception for Defining Simulation Fidelity," *Studies in Health Technology and Information*, (in press), 2006.

[7] P. Lamata, F. Bello, R.L. Kneebone, R. Aggarwal, F. Lamata, F.M. Sánchez-Margallo, and E.J. Gómez, "Marco conceptual para el análisis, diseño y evaluación de simuladores laparoscópicos," Proc. XXIII CASEIB - Congreso Anual de la Sociedad Española de Ingeniería Biomédica, pp. 503-506, 2005.

7 Usability evaluation method for mixed reality systems in surgery

Daniela Trevisan[1], Luciana P. Nedel[1,2], Jean Vanderdonckt[1] and Benoit Macq[1]

[1]Université cathólique de Louvain (UCL)
[2]Universidade Federal do Rio Grande do Sul (UFRGS)
{trevisan,macq}@tele.ucl.ac.be, nedel@inf.ufrgs.br, vanderdonckt@isys.ucl.ac.be

7.1 Introduction

One of the most challenging aspects on the new interactive systems development lies on the ability to integrate computer-based information into the real world. This kind of systems, so called mixed reality (MR) systems, are promising since it is capable of overlaying and registering digital information on the user's workspace in a spatially meaningful way. This characteristic allows MR systems to be an extremely effective operating medium, once it is based on monitoring the state of the user and/or the environment through sensors data acquired in real time, and adapting or augmenting the computational interface to significantly improve the user performance on the task execution. When applied to the medical surgery domain, MR systems allow users to keep their environmental perception while having intuitively access to more contextual information, such as the incisions location, regions to be avoided, diseased tissues, and so on.

Human-computer interaction (HCI) research usually focuses on the design, implementation and evaluation of interactive systems in the context of a user's task and work (Dix *et al.* 1998). One of the issues to be addressed in this field is to understand the people behaviour while using interfaces, trying to identify human beings perceptive, functional and cognitive resources, and how these resources are available during the accomplishment of a computationally based task. Once these resources are identified and their activity quantified, designers of computational interfaces will be able to consider these limitations in the design of new HCI systems.

Taking medical applications as an example, the generic design of human-computer interfaces for image-guided surgery (IGS) should deal with pre and intra-operative images displayed in the surgical environment to provide surgeons with some kind of guidance (Trevisan *et al.* 2003). In this typical application of MR systems, the virtual world involves the pre-operative information, while the real world corresponds to the intra-operative live information. Both should be correctly aligned and displayed in real time.

This scenario highlights the description of user interfaces and interaction for image-guided systems is not a trivial task. The system behaviour may be able to be broken into more than one discrete mode of interaction. However, the user is engaged with a continuous view of the environment. This limitation results in two different kind of interaction – one to deal with the physical space and the other for the virtual one. Consequently, interaction discontinuities break the natural workflow, forcing the user to switch between virtual and real operation modes. Integration and fusion of different sources of information in complex scenarios (such those required in operation rooms) in an effective and meanly way are yet an open issue.

In this work we are studying the user reaction in relation to IGS applications using different kind of augmentation. For this, we specifically report the method used by us to evaluate a mixed reality IGS. This method was adapted from the testbed application method (Bowman *et al.* 2005) currently used to evaluate virtual reality applications, to be applied to mixed reality

system, in general. The method presented below was tested in the evaluation of a mixed reality maxillofacial surgery system, also developed by us.

Our main objective with this usability test is to identify theoretical and practical basis that explains how mixed reality interfaces might provide support and augmentation for interactive applications. In other words, we wish to establish a relation between the design space principles (Trevisan *et al.* 2005) and the variables tested during the experiments. To accomplish that, we intend to use the results provided by the tests to find a model that allows us to identify the contribution factor of each of the analyzed variables in the final user interaction.

The paper is organized as follows: Section 7.2 presents some related work on evaluation of 3D interaction in VR and MR environments. Then, in Section 7.3 we presented the testbed application developed to carry on the evaluation method proposed, as well as the method proposed by us to measure the system accuracy. In Section 7.4 we present all the aspects needed to proceed with the system usability evaluation and in Section 7.5 we describe our experiment preparation. In Section 7.6 we explain how we intend to compute the interaction performance of the users and in Section 7.7 we make some final comments about the work developed and our on-going research.

7.2 Related work

Some experiments regarding evaluation of 3D interactions in VR worlds can be found in previous works, as the ones developed by Poupyrev *et al.* (Popyrev 1997), Forsberg (Forsberg 1996), Witmer (Witmer 1998) and more recently by Bowman (Bowman 2002), Nedel (Nedel 2003) and Tanriverdi (Tanriverdi 2000). We can resume the different results found in this domain by saying that the 3D visualization could help the user interaction if the dominant perceptual cues (i.e. depth cues involving motion parallax, shadows, transparency, etc.) are identified and adapted for each application.

Many new interaction techniques are emerging from 3D and virtual technologies looking for more natural interactions. However, the same problems regarding the 3D visualization are still there in addition to other specific factors. So far, usability evaluation and usability design is needed. Kato *et al.* in their usability studies [Kato 00] present five main objectives for tangible interfaces. Graphics objects should match physical constraints to task requirements. Moreover, interfaces should support parallel activities when reasonable, as well as providing aids for physically-based interaction techniques based on proximity and spatial relations. Furthermore, objects should support spatial and two-handed manipulation.

Dias (Dias 2003) presents the usability evaluation results of MixIt, a testbed for authoring in Mixed Reality Environments. The experiment has conducted usability evaluations for simple manipulations using also simple rigid geometric transformations (translation, rotation, scaling). As a result, they have noticed that the rotations using the magic ring were especially problematic. Several subjects mentioned the recognition problems resulting from one hand occluding the marker on the other hand, while trying to cross hands. Marker tracking problems related to the lighting also plagued the experiment. When asked if the users found the system to be helping them making mistakes or at least not helping them avoid them, subjects gave balanced answers: 6 said “no”, while 5 said “yes”. It shows that more efforts in developing natural and intuitive interfaces to interact with Mixed Reality systems need to be investigated, yet.

A potential direction for experimentation with synergistic interaction using multiple modalities has pointed out. Recent evaluations of tangible interfaces have shown directions

for using more affective input as user's interaction options. For instance, the SenToy (Höök 2003) is a tangible interface device used to influence emotional behavior in the logic game FantasyA. Another interaction technique is the full-body interaction. Navigation in virtual and mixed environments usually requires a wired interface, some kind of console, or keyboard. The advent of perceptual interface techniques allows a new option. In the work presented by Konrad (Konrad 2003) several different interaction styles are compared, based on an analysis of the space of possible perceptual interface abstractions for full-body navigation and the results of a wizard-of-oz study of user preferences.

7.3 Testbed application

7.3.1 Study case presentation

The study case we are using in this application relates maxillofacial surgery. Patients presenting huge deformations in her/his mandible anatomy frequently should suffer an osteotomy (operation to cut off part of the bones or to append some kind of prosthesis, according to her/his injury).

In this kind of intervention the surgeon cuts the patient mandible following an imaginary line, as the ones shown in Figure 7.1. Nowadays, this class of surgery procedure takes about 3 hours and the recuperation about 2 months. Despite the fact it is a very traumatic procedure, post operatory problems can expose the patient to a new surgery and another 2 months recuperation period.

In collaboration with the Service de Stomatologie et Chirurgie Maxillo-faciale at Saint Luc Hospital, in Brussels, we proposed the development of an application using a mixed reality interface to guide surgeons in this type of interventions. The goal of this application is to increase the first surgery success, avoiding a second intervention. To achieve this objective, we are mixing real and virtual images by projecting a virtual guidance path-line on the patient mandible captured video. Then, the surgeon should cut the patient mandible paying attention to follow this path-line and avoiding touching the dental nerves. The time involved in this procedure is not important, but the accuracy is mandatory.

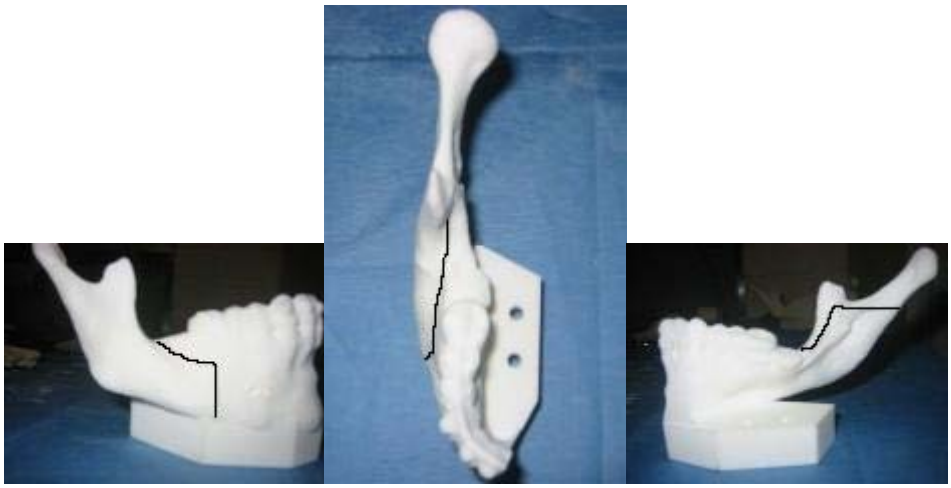


Figure 7.1: Three views of the cutting line for mandible surgery: left, upper, and right views, respectively.

7.3.2 System design

The flowchart that represents the system designed by us to support maxillofacial surgery including pre-operative CT scanning and planning to surgical guidance is shown in Figure 7.2.

The process begins with the *images acquisition* using a CT scanner. Then, a threshold is used to filter the images, separating the bones and the other soft tissues. After this processing, the images are *segmented* and a 3D model of the skull is *reconstructed* from the segmented images. Using the reconstructed 3D model and the medical doctors' experience, a path-line representing the osteotomy path is then *designed*.

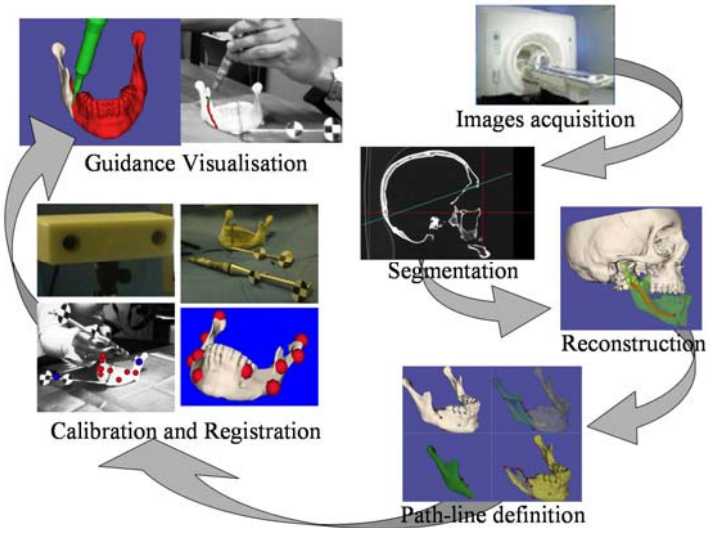


Figure 7.2: System flowchart to provide surgical guidance during maxillofacial surgery.

The next step involves the hardware *calibration* and the *registration* of real and virtual worlds. As our testbed application was conceived to be tested for a number of volunteers, it is not based on a real mandible. To simulate the same situation of an operation room, we generated mock-ups of the mandible in gypsum, based on the 3D model we reconstructed in the last step. Our real environment (the surgery scenery) involves the mock-up of the mandible, the surgeon hand and the tool used to cut the bones (as shown in Figure 7.2 upper left). The virtual environment (that should be registered with the real scene) is composed by the 3D representation of the mandible, the dental nerve, the tool marks and the path-line. After the *registration*, the final images are mixed into a common reference frame and the *visualization* and interaction with the application becomes possible.

As mentioned before, we are focusing on guidance for osteotomy procedures. Two main scenarios are explored in this study, considering virtual and augmented guidance.

The virtual guidance scenario provides the 3D visualization (in the screen) of the tracker and the reconstructed mandible. The interactive task consists in cutting the mandible following as best as possible the line-path showed in the screen. When the surgeon touches the good position (on the real object), the virtual representation of the surgical tracked tool becomes green (see Figure 7.3). Another dynamic sphere attached to the tracked tool indicates the distance from the tool-tip to the internal structure. Additional information, such as where is placed the internal structure (dental nerve), are provided by using transparency visualization effect. Manipulations with the virtual object such as rotation, scale and zooming are allowed at any time.

Figure 7.3: Virtual scenario for the simulation of an osteotomy procedure.

The augmented guidance scenario provides the visualization (in the screen) of the 3D elements superposed to live video images. This scenario presents two alternatives according to which information are displayed in the scene. In the first variant, the path-line is projected over the video image. When the user touches the correct position (on the real object), a projected sphere representing the tracked tool-tip becomes green (Figure 7.4 left). To indicate if the tool is in front of or behind the mandible mock-up we change the intensity of the green sphere. If the tracking tool is behind the mock-up and consequently occluded in the viewed image, the visualization of a semi-transparent sphere guarantees a kind of guidance for the user. The same principle is used in relation to the path-line. As shown in Figure 7.4 left, the draw representing the path-line is dotted when it indicates an occluded area (behind the visible real surface). This kind of representation was suggested by the surgeon once dotted lines have been currently used in other procedures such as pre-operative planning.

The second variant of the augmented guidance scenario counts with two more additional information: the distance indicator from the tool-tip to the internal structure (dental nerve) and the projection of this internal structure on the video image. When the marker is recognized by the system, two spheres are displayed in the screen: one attached to the tool-tip; and other, bigger, displayed in the middle of the tool (see Figure 7.4 right). The big one indicates the distance from the tool-tip to the internal structure using three colours: *gray* means *go ahead*; *blue* means *pay attention* (i.e. you are less than *3mm* from the nerve); *red* means *stop* (i.e. you will probably touch the nerve).

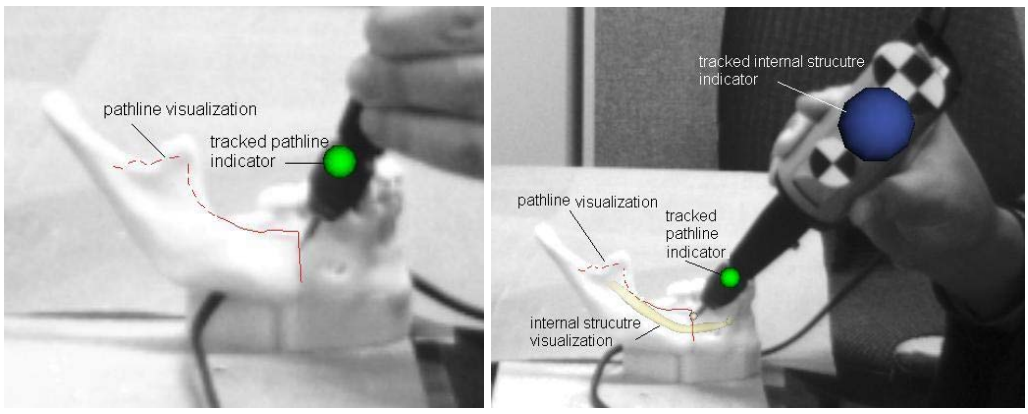


Figure 7.4: Two alternatives for the augmented guidance scenario: path-line visualization and tracked path-line indicator (left); and path-line and internal structure (dental nerve) projection, tracked path-line position, and distance indicator from the tool-tip to the internal structure (right).

7.3.3 Apparatus

In mixed reality applications, as well as in virtual reality ones, the features of the input and output devices used affect the user performance during the execution of a task (Mackenzie 1995, Poupyrev 1997). Attributes like resolution, field of vision and degrees of freedom should be considered when comparing users' performance.

We have performed our tests in a PC computer with an AMD Athlon 64 3.0Ghz processor, 1GB DDR-800 of RAM memory, a nVidia GeForce FX 5700 graphics card, a 120Gb Western Digital 7200RPM SATA hard drive and with the Ubuntu 5.04 Linux operating system distribution. We have used an LCD screen (Figure 7.5.b) for the visualization.

To capture the interaction of the user with the mock-up we have used a live video capture system based on a stereo camera from Claron Technology model S60 (Figure 7.5.a).

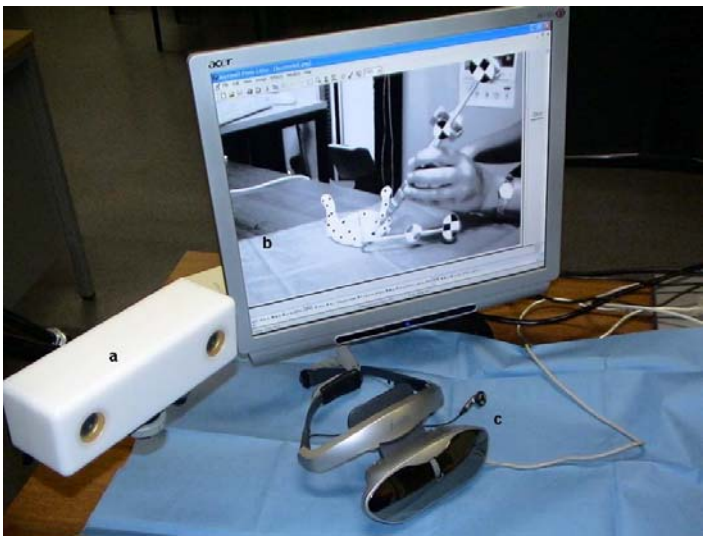


Figure 7.5: Experimental setup: displays and stereo camera

The tool used to cut the mock-ups is a mini-driller with controlled speed (Figure 7.6.d).

The mock-ups used in the experiments are two: a simple one, representing a 3D L geometry (Figure 7.6.b) and a more complex, the printed 3D mandible (Figure 7.6.c). Both present a wire passing inside, simulating the dental nerve, were built with the same material (gypsum) and printed using the Z-corp[7] machine.

To detect if the user touched the dental nerve or not during the experiment, we have developed an electronic circuit (Figure 7.6.a) connecting the wire and the tool. It emits a sound alarm and a visual feedback (using a LED) when the tool touches the wire.

[7] Z-corp Website: hhtp://www.zcorp.com

Figure 7.6: Experimental scenario tools

7.3.4 System accuracy measurement

A big problem involved in the evaluation of virtual and mixed reality system is the number of constraints imposed by its components and the huge dependencies identified among these components. The uses of non-conventional devices which the technical performance are not yet established forbid the correct evaluation of its use. For example, during the evaluation of a new interactive technique adapted for head-mounted displays, how to identify the origin of the good (or bad) user performance? Is the new technique really good (or bad)? Is the device sufficiently precise for the proposed task? Is the user familiar with this kind of device, if comparing with the conventional ones? As it is yet difficult to individualize the several components of a MR system, it is also hard to answer the questions above.

Trying to minimize the intervention of each component of the system in the final evaluation, we calculated our system performance by estimating the global and local errors.

The local error (e_l) is calculated by computing the sum of the error estimation for all factors that can insert noise in the system: the tracker (e_{tr}), the user interaction (e_{ui}), the registration (e_{reg}) and the printed model (e_{pm}).

$$e_l = e_{tr} + e_{ui} + e_{reg} + e_{pm}$$

The global error (e_g) is provided by calculating the distance from the points picked by the user to the registered object. We can say that

$$e_g \approx e_l$$

As our model of the mandible is quite complex and hard to assess, we have also modelled and reproduced a simpler mock-up in form of L (as shown in Figure 5.6) which abstracts the mandible format.

Accuracy, or trueness, is the closeness of measured values to their expected quantity as defined by some standard. The difference between the true value and the value reported by the instrument is called the measurement error. In measuring a spatial position, the measurement error is a 3-vector *(dx,dy,dz)*, where each component is the difference between the measured position and the true position in the corresponding axis. The error value is often reduced to a single number, which corresponds to the Euclidean length of the error vector, i.e., $\sqrt{dx^2 + dy^2 + dz^2}$.

There is no standard way of expressing or reporting 3-D orientation errors. Referring to a single angle implies a selection of a single vector whose measured and true orientations are

compared. The reported error value would thus be critically affected by the choice of vector. For example, the component of the orientation error that affects the roll around the selected vector would have no influence on the error value. Perhaps the most useful way to quantify orientation errors is to report on the largest angular error for any vector, or some approximation of it, but such common measure has yet to be established in the optical tracking industry.

Due to a variety of uncontrolled factors, repeated measurements of the same quantity under different conditions generate slightly different values. It is, therefore, common to take a statistical view of accuracy, describing it in terms of error distribution around the true value.

A number of statistical values may be used to quantify error distribution in a single number. The maximum error (or its close relative, the median error) is easy to define, calculate and understand. Unfortunately, it is not a useful term in characterizing measuring performance, since its value critically depends on the number of samples taken and, due to random factors, may vary widely from one experiment to another unless the number of samples is very large, sometimes impractically so. The average error is also easy to define and calculate. It does not, however, reflect well enough how *safe* it is to rely on the reported value. In averaging, small errors are assigned the same importance as large errors, while users often consider large errors more important than small ones. The average error, therefore, tends to generate lower values than what users feel is the *true* inaccuracy of the instrument.

Perhaps the most useful statistical error quantification is the % confidence interval (CI), defined as the value below which a specified % of the samples fall. For example, the value reported for the 95% confidence interval (95% CI) implies that only 5% of the samples had higher error values. Unfortunately, CI values have not been widely reported in the past, and there is no consensus regarding what specific interval to use.

The most frequently used measure of spatial measurement accuracy is the Root Mean Squared Error (RMSE), calculated by averaging the squares of the individual error values, then taking the square root of the average. If all the errors have the same magnitude, RSME generates the same value as the average error. When some of the errors are larger than others, however, they are given more weight, resulting in the RMSE being larger than the average.

Error estimation is calculated as follows.

7.3.4.1 Tracking error

When the accurate value is not known, it is often useful to report on the repeatability (or preciseness) of measurements in terms of their standard deviation (STD). Standard deviation is identical to RMSE when the average is the true value, i.e., it is calculated by taking the square root of the average of the squares of deviations from the average value. When the average is off the true value by distance *d*, the following relationship applies (assuming a normal error distribution):

$$RMSE = \sqrt{d^2 + STD^2}$$

The causes for measurement errors in optical pose sensors are numerous and their effects are often complex and interdependent. A useful way to discuss them is by dividing into 3 categories: (1) calibration; (2) drift; and (3) jitter.

Calibration errors are introduced in the camera calibration process at the factory, and remain constant for years unless the physical structure of the camera changes, e.g., due to a physical shock. Drift represents the effects of variations in the operating environment, e.g., temperature, lighting, or marker position/orientation. Jitter is a momentary deviation caused by optical or electrical noise in the image capture and analog-to-digital conversion circuitry.

Claron Technology guarantees that the calibration error of each camera shipped falls bellow 0.25mm RMSE and the jitter's standard deviation under a selection of typical measurement conditions (i.e taking account the field of measurement approximately 75 cm, 150 Lux, 20 ms shutter) is *0.007mm* RMSE for static target and *0.07mm* RMSE for moving target.

To control drift problems, we access the hazard messages provided by the MTC library. For instance, drift can be caused by a rapid internal temperature change in camera. In this case a thermal instability hazard is added to measurements until camera becomes thermally stable ($\approx$15 minutes following power-up or 2 minutes following activation, whichever is later).

Besides, small drift may result when part of an Xpoint is shadowed at certain orientations. This condition is detected and a *Shadow over XP* hazard is appended to the marker's measurement.

A relevant factor for the accuracy of the tracked object is how to place the markers. Due to the lever effect, the farther apart the Xpoints at the ends of a vector, the lower the error in measuring the vector's orientation in space, and, therefore, the lower the error in extrapolating object positions of interest, such as a tool-tip. The drawing below (Figure 7.7) demonstrates this effect by showing a tool tip position error range for two different placements of Xpoints. To minimize extrapolation error, one end of the longer vector should be placed as close to the tool-tip as possible, and the other as further away from it as possible.

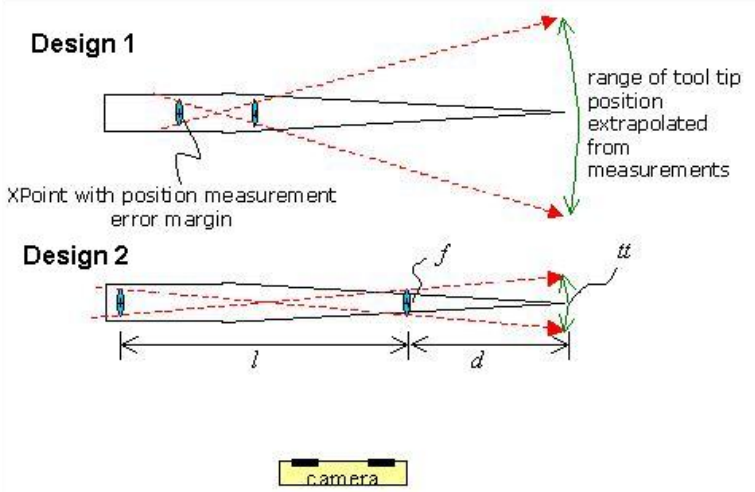


Figure 7.7: Relation between tool-tip accuracy and markers placement

Assuming the tool-tip is approximately aligned with the long vector of the marker, and the vector's length is *l* (refer to the Figure 7.7), and given a position RMS error at each marker Xpoint e_f and a distance *d* between *f* and the tool-tip point *tt*, the tool-tip RMS error e_{tt} is approximately:

$$e_{tt} \approx e_f + 1.5 * e_f * d / l$$

For this reason we have tested two designs of markers for the tracked tool shown, as shown in Figure 7.8. We have computed 1000 pose for each design and the RMS errors for the positions shown in Figure 7.7 are depicted in Table 7.1.

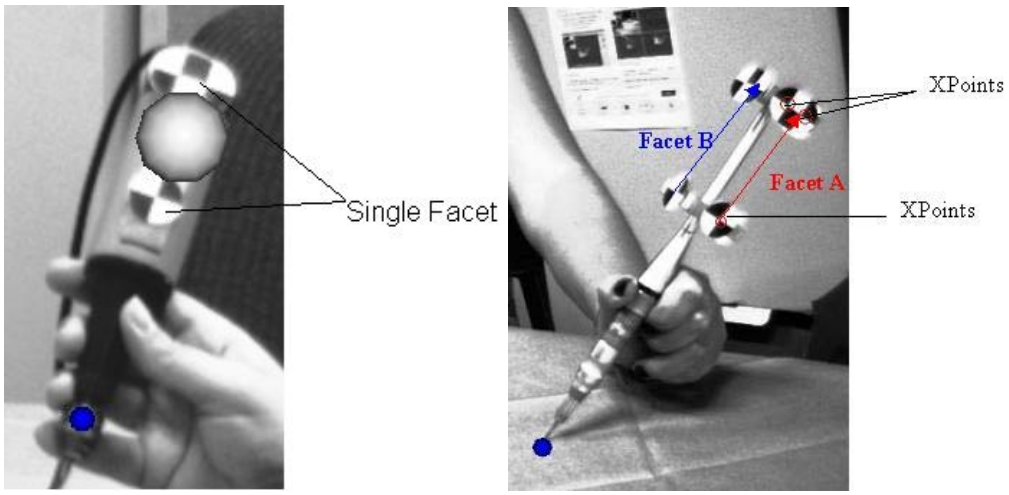


Figure 7.8: Markers attached to the tracked tool: one facet marker (left); and a multi-facet marker.

Design	*l(mm)*	*d(mm)*	e_f	e_{tt}	Tool-tip poses captured (RMS)
1	60.55	118.91	0.17	0.5564	0.5680
2	117.1933	58.5695	0.20	0.3470	0.3620

Table 7.1: Tool-tip precision error for 1000 pose

From this analysis, the marker configuration for Design 2 (Table 7.1) presents a better accuracy and we will use such design for our tests.

If *d* is small relative to *l*, the tip error is dominated by the Xpoint error (roughly, calibration error + jitter, typically in the range *0.2 mm - 0.4 mm* RMS). If *d* is large relative to *l*, the error is dominated by the ratio *dl*. As a rule of thumb, to ensure that RMS error at the tip is sub-millimeter, keep $l \geq d$.

7.3.4.2 User interaction error

We call user interaction error, the imprecision of the user while interacting with the system and is directly related to the dexterous ability of each user. To evaluate the precision of such interaction we have performed a sequence of 15 tests with 15 users that provide us the RMS and the standard deviation for the user interaction error.

The test consists in picking the 4 corners of one of the faces of the L mock-up. As the real distance and angles between these points are known (Figure 7.9 (left)), we can compare it with the distances and angles given by the acquired 3D points (Figure 7.9 (right)). As these points were acquired by the users while using the tracking system and the tracking error e_{tr} is known, the user interaction error e_{ui} should be the values collected during the experiments, minus the tracking error, or yet *0.048mm*.

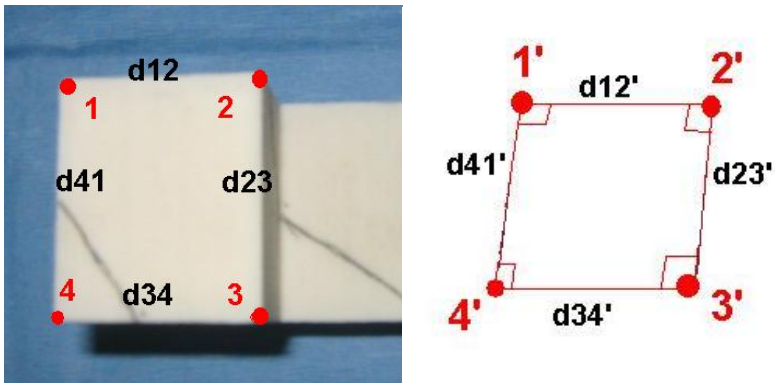


Figure 7.9: Distances and angles between real measures and virtual ones are compared: points selected by the user on the real object surface (left); and virtual representation of the selected points (right).

If the user that will use the system performs such calibration procedure before the registration procedure, the system will be able to compute a personalized performance for this user.

7.3.4.3 Registration error

The efficiency and accuracy of the registration method used is highly dependent on the choice of a number of parameters both at the acquisition stage and during data processing. Such method was previously validated by Noirhomme *et al.* (Noirhomme *et al.* 2004), where registration has sub-pixel accuracy in a study with simulated data, while they obtained a point to surface RMS error of *1.17mm* to *0.38mm* in a 24 subject study.

In this section, we study the influence of three such parameters: the number of points one should acquire to characterize the object surface, the pattern along which those points should be acquired, and the resolution needed for the Euclidean distance map.

For the resolution, we assume the values found by Quentin *et al.* where a resolution similar to the internal slice resolution of the image yields good results. Improving the distance transformation resolution further does not benefit to the accuracy.

For the number of points and which pattern to consider we have performed a sequence of tests and we found 20 points along the contour of the mandible has produced acceptable registration matching.

7.3.4.4 Printed model error

The objects (i.e. the phantoms) used in this work were produced by rapid prototyping based on a principle of manufacture per addition of matter using the Z-corp[8]. This work is done in collaboration with CRIF-WTCM[9].

The final printed object were scanned (using laser scanner) and compared with the initial 3D model using the Inspector Software[10] to align and registered the models. Figure 7.10 is showing the objects overlapped with a precision of *0.20 mm*.

[8] Website: http://www.zcorp.com/
[9] Website: http://www.wtcm.be/
[10] Website: http:// www.metrics.be

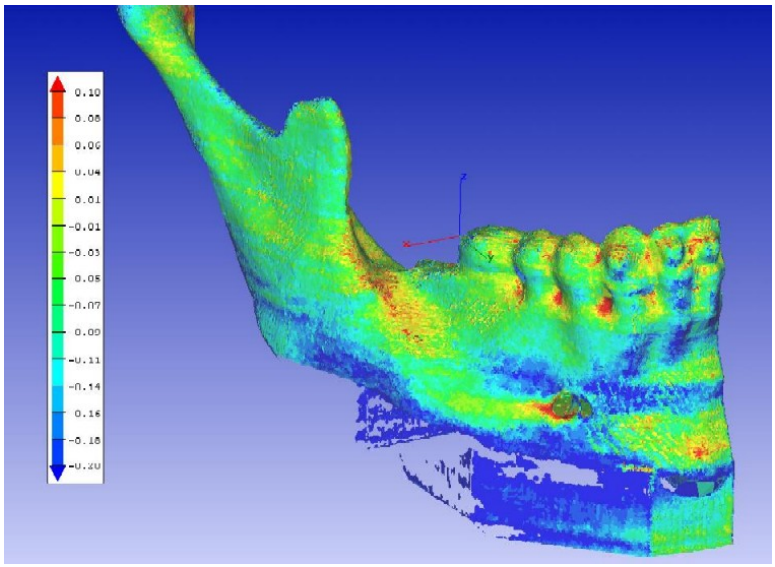


Figure 7.10: Registration between printed and virtual objects indicates printed model error of *0.20 mm.*

7.4 Evaluating the system usability

7.4.1 Scenarios considered

For the evaluation tests, we considered three different scenarios:

- No guidance scenario
- Virtual guidance scenario
- Augmented guidance scenario

In the no guidance scenario none computer guidance is provided to the user. Three different views of the path-line (Figure 7.1) are shown to the user in a paper printed version and the user should cut a real mandible just checking this piece of paper. This scenario corresponds to the real currently situation and will serve as reference scenario to measure the value added by the guidance provided by the other scenarios. The other two scenarios were previously explained in Section 7.3.2.

7.4.2 Hypotheses

In this section we describe the hypotheses we intend to prove with the tests. We considered only single variable hypotheses, all of them relevant in the evaluation of the continuity criteria. Each hypothesis is based on the user performance which, in our case, is defined as *the precision for the task execution.*

H1 - The kind of guidance will influence the task performance and the user satisfaction.

H1.1 - precision on the task execution will increase according to the volume of guidance (information) given

H1.1.1 - none guidance will provide the worst task precision

H1.1.2 - the path-line guidance will provide lower task precision than path-line + internal structure + distance indicator

H1.1.3 - the path-line + internal structure + distance indicator will provide the best task precision.

H1.2 - time of task execution will increase according to the volume of guidance (information) given

H1.2.1 - none guidance will present the shortest time for task execution

H1.2.2 - the path-line guidance will present shorter time than pathline + internal structure + distance indicator guidance

H1.2.3 - the path-line + internal structure + distance indicator guidance will present the biggest time of task execution

H1.3 - the global task performance will increase according to the volume of guidance information given

H1.3.1 - none guidance will provide the worst task performance

H1.3.2 - the path-line guidance will provide lower task performance than path-line + internal structure + distance indicator

H1.3.3 - the path-line + internal structure + distance indicator will provide the best task performance

H1.4 - user satisfaction will increase according to the volume of guidance information given

H1.4.1 - none guidance will provide the worst user satisfaction level

H1.4.2 - the path-line guidance will provide lower user satisfaction level than path-line + internal structure + distance indicator

H1.4.3 - the path-line + internal structure + distance indicator will provide the best user satisfaction.

H2 - The kind of visualization used will influence the task performance and the user satisfaction

H2.1 - the task execution precision will increase according to the kind of visualization

H2.1.1 - the three paper printed views of the path-line will provide the worst task precision

H2.1.2 - the virtual view will provide lower task precision than the augmented one

H2.1.3 - the augmented view will provide the best task precision

H2.2 - time for task execution will increase according to the kind of visualization

H2.2.1 - the three paper printed views of the path-line will present the shortest time for task execution

H2.2.2 - the virtual view will present shorter time for task execution than the augmented view

H2.2.3 - the augmented view will present the biggest time for task execution

H2.3 - the global task performance will increase according to the kind of visualization

H2.3.1 - the three paper printed views of the path-line will provide the worst task performance

H2.3.2 - the virtual visualization will provide lower task performance than augmented visualization

H2.3.3 - the augmented visualization will provide the best task performance

H2.4 - user satisfaction will increase according to the kind of visualization

H2.4.1 - the three paper printed views of the path-line will provide the worst user satisfaction level

H2.4.2 - the virtual visualization will provide lower user satisfaction level than augmented visualization

H2.4.3 - the augmented visualization will provide the best user satisfaction level

H3 - the kind of mock-up will influence the task performance and the user satisfaction

H3.1 - simple mock-up will provide better precision on task execution than complex ones

H3.2 - simple mock-up will provide shorter time on task execution than complex ones

H3.3 - simple mock-up will provide better global task performance than complex ones

H3.4 - simple mock-up will provide better user satisfaction than complex ones

7.4.3 Independent variables

Independent variables are the experiment variables that are manipulated to generate different conditions to compare. In general, good examples of independent variables in interfaces are interface style and help level (Dix *et al.* 1998). In virtual reality environments, examples are the size of the selectable objects and the visualization strategy used. In this mixed reality experience, we established the following variables to test the hypotheses:

- ***Kind of guidance.*** Used to prove the first hypothesis, concerns the strategy used to help the user to better accomplish the task (to cut a mock-up following a precise path-line).The available guidance possibilities are: (1) no guidance at all; (2) the visualization of the path-line with a visual feedback of the tracked line; (3) the visualization of the path-line with a visual feedback of the tracked line and the internal structure (dental nerve) visualization + the visual feedback of the tracked internal structure;
- ***Kind of visualization.*** This variable allows the direct test of the second hypothesis. The three possible visualization strategies considered are: (1) three paper printed views of the path-line over the mandible; (2) the virtual visualization; (3) the augmented visualization;
- ***Kind of mock-up.*** This variable is used to test the third hypothesis and involves the design of the real mock-up used. For our tests, we considered two mock-ups: (1) a 3D object with the form of an L; and (2) another one representing a real mandible.

7.4.4 Dependent variables

Dependent variables are the measures taken as the performance indicative or level of acceptance of the technique by the users. These measures can be objective, as the time taken to accomplish a task, or subjective, collected from post-test questionnaires answered by the subjects.

The dependent variables used in our experiment were:

- Precision on task execution. We measure the user precision by calculating the distance between the path-line and the line drawn by the user;
- Time to perform the task;
- Task performance: Precision/Time;
- User satisfaction: based on the value attributed to the *frustration* factor during the application of the NasaTLX post-test;
- Workload. Based on the result given by the subjective NasaTLX post-test.

7.5 The experiment

7.5.1 Task description

The task is always the same for the three scenarios described before in Section 7.4.1 and consists in cut the mock-up reproducing as best as possible the indicated trajectory without touching the internal hidden structure (dental nerve). There is no limit time to accomplish the task.

Before to start the execution of the main task, the user has free time for training, consisting in performing sub-tasks that allow the user to understand how the system will provide guidance in each scenario. In the same time, we can decompose the task complexity and isolate variables involved in the final interaction. Then, the main task was decomposed into two sub-tasks: user perception calibration and motor calibration.

7.5.1.1 User perception calibration

In the perception calibration task the user should select over the mock-up surface arbitrary points indicated in the virtual or augmented scene. This task is done once without guidance (i.e. the user perceives the point in the virtual or augmented scene and performs the task) and then with guidance (i.e. when the user is pointing the good position, a small sphere representing the tracked tool-tip becomes green). This task is performed in both guided scenarios: virtual and augmented. In the augmented scenario, the tracked tool-tip becomes transparent and the path-line visualization is dotted to indicate the surface occlusion. Time and precision (i.e. performance) for the user task accomplishment are computed.

7.5.1.2 User motor calibration

In the motor calibration procedure, the user should cut a trajectory drawn over the mock-up surface. In this case, the mock-up resistance is the same used to execute the main task. Time and precision (i.e. performance) for the user task accomplishment are computed.

7.5.2 Pre-Tests

Based on the three scenarios stated we have designed 8 different situations (as illustrated in Table 7.2) which were tested during the pre-tests with 5 volunteers. As a result of these preliminary tests, some visualization parameters were adjusted, such as, level of object transparency, position of the distance indicators in the scene and the values range used to track distances (i.e. path-line and internal structure distance). For the tracked path-line distance we have found the range of 2mm (1mm to the left and 1mm to the right) a good one for setting colours guidance. For the tracked internal structure distance we have delimited the range from 2mm to 5mm as the “pay attention zone” and less than 2mm as the “stop zone”. In this case, the surgeon has suggested giving the “red” visual feedback before touching the internal structure once it would be too late in the real scenario.

Scenario	Kind of visualization	Kind of guidance	Kind of mock-up	Kind of display
1	None	(1)	L	None
2	None	(1)	Mandible	None
3	Virtual	(3)	L	LCD screen
4	Virtual	(3)	Mandible	LCD screen

5	Augmented	(3)	L	LCD screen
6	Augmented	(3)	Mandible	LCD screen
7	Augmented	(2)	L	LCD screen
8	Augmented	(2)	Mandible	LCD screen

Table 7.2: List of the scenarios used during the pre-tests session. The numbers in the "kind of guidance" column correspond to: (1) no guidance; (2) the visualization of the path-line with a visual feedback of the tracked line; (3) the visualization of the path-line with a visual feedback of the tracked line and the internal structure (dental nerve) visualization + the visual feedback of the tracked internal structure.

7.5.3 Subjects and procedure

Thirty two users, 30 right-handed and 2 left-handed, volunteered for experience the tasks: 30 males and 2 females, with a mean age of 24 years.

Each participant tested only one of the designed scenarios. Practice was balanced using a total of 16 L mock-ups and 16 mandible mock-ups. Each scenario was tested 4 times. Before each session, the mock-up should be calibrated and registered. Session length varied according to the index of difficulty given by the kind of mock-up, the kind of guidance, and the level of the individual motor coordination of each user.

The subjects were tested individually. They began the session with a familiarisation period and performing some simple tasks (such as perception calibration and motor calibration) taking account the kind of scenario which is being used. Considering the virtual 3D scenario, users were carefully informed how to use the manipulation commands (i.e. translation, rotation and zoom), but they were left free to use it or not during experimental trial.

At the end of the test, each user was asked to answer the NasaTLX questionnaire.

7.5.4 Logging

For each user and each task, we are saving the time spent to accomplish the task, the original path-line and the path-line drawn by the user.

7.6 Computing interaction

Some much known metrics have been used to compute performance and movement time in many experiments. For instance, the Fitts' law (Fitts 1954) is an effective method of modelling rapid, aimed movements, where one appendage (like a hand) starts at rest at a specific start position, and moves to rest within a target area. The law can be used to assist in the design of user interfaces. It can also be used to predict the performance of operators using a complex system, assist in allocating tasks to operators, and predict movement times for assembly line work. However, Fitts' law predicts movement in only one dimension. Fitts' original experiments tested human performance in making horizontal moves toward a target. Both the amplitude of the move and the width of the terminating region were measured along the same axis. It follows that the method is inherently one-dimensional. So, when dealing with two or three dimensional target acquisition tasks, new interpretations of "target width" must be considered. Another major deficiency is the absence of a consistent technique for dealing with errors. Researchers have developed a method to handle errors, but it has been largely ignored because of its complexity. In summary, using such approach for any particular design problem we are working on the maths won't tell us which part of the parameter space we are operating in. We are not interested just in find a number representing the final user

performance; we want to be able to decompose such result by identifying all parameters involved in the final performance such as perceptive, cognitive and functional parameters.

For each tested scenario, the 3D distance between the reference and the final path performed by the user were calculated providing the final user performance. Both paths (i.e. reference and final) were acquired in the same way, i.e. digitalizing points under the surface trajectory using the tracked tool. Even task was performed in depth we use such information (i.e. distance from tool-tip to the internal structure) just to guide the user during the execution and then we can control if the internal structure was touched or not and then control the level of task execution success.

Once the final user performance was calculated including user perceptive performance and user motor performance for a given system accuracy, we wish to identify all variables involved in the interaction and express it in terms of continuous interaction.

User interaction can be expressed by the following components:

$$I \approx \alpha V_i + \beta G_j + \gamma O_k + \sigma D_l + \varphi W_{ijkl} + \varepsilon$$

, where numbers are representing the values that the variable can assume:

V_i = kind of visualization (i= 1, 2, 3)

G_j = kind of guidance (j= 1, 2, 3)

O_k = kind of object (k= 1, 2)

D_l = kind of display (l = 1)

W_{ijkl} = workload index for a given visualization i, guidance j, mock-up k and display l

$\alpha, \beta, \gamma, \sigma, \varphi$ = contribution factors for the final interaction I

ε = error factor

To solve such linear system we applied a SVD (Singular Value Decomposition) method which uses the least squares and pseudo-inverse computation (Peters *et al.* 1970). Least squares is a mathematical optimization technique that attempts to find a "best fit" to a set of data by attempting to minimize the sum of the squares of the ordinate differences (called residuals) between the fitted function and the data.

In mathematical terms, we want to find a solution for the equation

$$\left[\underline{A}^T \underline{A}\right] C = \underline{A}^T I$$

where $\underline{A}$ is an m-by-n matrix of variables (with $m > n$) and C and I are n-respectively (vector of constants $\alpha, \beta, \gamma, \sigma, \varphi$) m-dimensional (vector of final user interactions) column vectors.

7.7 Final comments

In this document we presented the strategy we developed to evaluate the usability of mixed reality systems, precisely, the use of mixed reality in surgery. First of all, we briefly presented the problem we are trying to solve and the application developed to be used as a testbed

application. Then, we explain how we calculated the errors inherent from the computer system and devices used, that is used for measuring the system accuracy.

Finally, the experiment is described in detail and the method we will use to evaluate the system interaction is presented. In our on-going work we are doing the tests with the users. After that, we will tabulate the tests results and extract our conclusions. With this procedure we intend to highlight the influence of each system component (kind of visualization, kind of guidance, kind of object, kind of display and workload) in the final user performance. The results obtained from this evaluation will be then combined with the suggestions proposed by the use of DesMiR (Trevisan *et al.* 2005) design space.

References

Bowman, D., Gabbard, J., and Hix, D. A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods. Presence: Teleoperators and Virtual Environments, vol. 11, no. 4, 2002, pp. 404-424.

Bowman, D.A., Kruijff, E., LaViola, J.J. and Poupyrev, I, 3D User Interfaces – Theory and Practice, Addison-Wesley, ISBN 0-201-75867-9, 2005.

Dias, M., Jorge, J., Carvalho,J., Santos, P. Luzio, J. Usability Evaluation of Tangible User Interfaces for Augmented Reality, The Second IEEE International Augmented Reality Toolkit Workshop, 7th of October 2003, Tokyo, Japan.

Dix, A., Finlay, J., Abouxd, G. and Beale, R., Human-computer Interaction, Prentice Hall, England, 2nd ed., 1998.

Fitts P. M. The information capacity of the human motor system in controlling the amplitude of movement, Journal of Experimental Psychology, vol.47 pp.381-391, 1954.

Forsberg A., Herndon K., Zeleznik R., Effective Techniques for Selecting Objects in Immersive Virtual Environments, Proc. ACM UIST'96 Symposium on UserInterface Software and Technology (UIST), 1996.

Höök, K., Bullock, A., Paiva, A., Vala M., Chaves, R., Prada, R. "FantasyA and SenToy", Computer Human Interaction (CHI'03), extended abstracts on Human factors in computer systems Ft. Lauderdale, Florida, USA, 2003, pp. 804 – 805.

Kato, H., Billinghurst, M, Poupyrev, I., Imamoto, K.,Tachibana, "Virtual Object Manipulation on a Table-Top AR Environment", Proceedings of the International Symposium on Augmented Reality (ISAR 2000). Oct. 5-6, 2000.

Konrad, T., Demirdjian, D., Darrell, T. Gesture + play: full-body interaction for virtual environments, CHI '03 extended abstracts on Human factors in computer systems Ft. Lauderdale, Florida, USA, 2003, pp. 620 – 621.

MacKenzie, I. "Input devices and interaction techniques for advanced computing", Virtual Environments and Advanced Interface Design. Edited by Barfield, W. and Furness III, T. Oxford University Press, NY, 1995, pp. 437-470.

Nedel, Luciana P.; Freitas, Carla M.D.S.; Jacob, L.J.; Pimenta, M.. Testing the Use of Egocentric Interactive Techniques in Immersive Virtual Environments. In: INTERACT 2003 - NINTH IFIP TC13 INTERNATIONAL CONFERENCE ON HUMAN-COMPUTER INTERACTION, 2003, Zurich. Human-Computer Interaction: INTERACT´03. Amsterdam: IOS Press, 2003. p. 471-478.

Noirhomme Quentin and Ferrant Matthieu and Vandermeeren Yves and Olivier Etienne and Macq Benoit and Barette Olivier, Registration and Real-Time Visualization of Transcranial Magnetic Stimulation With 3-D MR Images, IEEE Transactions on Biomedical Engineering, vol.51, pp.1994-2005, Nov. 2004.

Peters, G. and Wilkinson, J.H. "The least squares problem and pseudo-inverses", Comp. J., 13:309-316, 1970.

Poupyrev, I. et al. “A framework and testbed for studying manipulation techniques for immersive VR”. ACM Symposium on Virtual Reality Software and Technology 1997, Swiss Federal Institute of Technology, Lausanne, Switzerland, September 1997.

Tanriverdi, V. and Jacob, R. J. K. Interacting with Eye Movements In Virtual Environments, In proceedings of CHI’00, ACM, pp.265-272, 2000.

Trevisan, D.,Vanderdonckt, J., Macq, B. And Raftopoulous, C. “Modeling interaction for image-guided procedures”, In: Proceedings of International Conference on Medical Imaging SPIE2003 (San Diego, 15-20 February, 2003), v. 5029, p. 108-118, International Society for Optical Engineering, 2003.

Trevisan, D., Vanderdonckt, J. and Macq, B. “Conceptualizing mixed spaces of interaction for designing continuous interaction”, Virtual Reality, v.8, p.83-95, 2005.

Witmer, B. G. and Singer, M. J. Measuring Presence in Virtual Environments. A presence questionnaire. Presence, Vol. 7, No. 3 pp. 225–240, June 1998.

8 Remote Usability Analysis of MultiModal Information Regarding User Behaviour[11]

Fabio Paternò, Angela Piruzza and Carmen Santoro

ISTI-CNR, Pisa (ITALY)

Abstract

In this section we describe MultiModal WebRemUsine, a tool for remote usability evaluation of web sites that considers information from log files, videos recorded during user tests, and data collected by an eye-tracker. The tool performs an automatic evaluation of the usability of the considered web site by comparing such information (which describes the actual behaviour of the users) with that contained in the task model associated with the pages (which describes the expected behaviour of the user). The results of the analysis performed by the tool are provided to the evaluators in terms of task not completed, errors occurring during the performance of tasks, time for completing a task, etc. These results are provided along with information regarding the user behaviour during the task performance. Using such data, evaluators should be in a position to identify problematic parts of the website and make improvements, when necessary. An example of application of the proposed method is also shown.

Author Keywords

Remote usability evaluation, websites, multimodal data.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

8.1 Introduction

The dissemination of Web applications is enormous and still growing. The great penetration of Web sites raises a number of challenges for usability evaluators. In this section we discuss what information can be provided by automatic tools able to process multimodal information on users gathered from different sources, so as to provide the most effective remote usability evaluation of websites. The collected information ranges from browser logs to videos to eye-tracking data, and the approach proposed tries to integrate such data in order to derive the most complete information for analysing, interpreting and evaluating the users while visiting a website, by taking into account the factors that might affect the performance of the users.

The proposed approach is supported by a tool – MultiModal WebRemUsine, which has been improved over the years in order to include and handle more and more information and provide additional features. In one of its first versions [4], the tool was just able to automatically analyse the information contained in Web browser logs and compare it with task models specifying the ideal behaviour of users interacting with the application and

[11] This paper was published in the Proceedings of the International COST 294 Workshop on User Interface Quality Models, Rome, Italy, 2005.

representing the actual Web site design. The goal was to identify where users interactions deviate from those envisioned by the system design and represented in the model. However, such information can be rather limited because when users visit a webpage, their attention can be captured by different areas of the same page and this information cannot be derived just analysing log files, which are only able to track physical interactions of the user with the application (e.g., scrolling, clicking, etc.). Eye tracking is a technique able to allow for deriving the current area of interest of the user by following the user's gaze. Thus, it helps evaluators in discovering the navigation strategies of the users visiting the web site and analysing the impact of different areas of the page. This enables easy identification of possible problematic parts of the page.

However, there are situations in which even the eye-tracker data may result inadequate to provide sufficient information for effective evaluation. Indeed, a user may look at the same portion of the page for quite different reasons, and such reasons could not be discovered by just analysing eye-tracking information. For instance, users might delay in staring at a certain point of the page because they might not be aware of having found the requested information and still look for it or, alternatively, they are aware of having achieved the goal and thus are interested in further reading the information found. In both cases, a video-based analysis might provide useful information for interpreting the different impacts of the same page on users: for instance, it might highlight situations where, although the logged information and user's gaze might make evaluators conclude that the user has successfully completed the expected task, a puzzled expression of the user should force the evaluator to re-interpret the collected data and derive, more realistically, that the user has completed the task but might not have realised it, which is still a signal of a usability problem in the page. So, this simple example shows how important is integrating all the data that is possible to capture on the user behaviour (and possibly also on the environmental/contextual conditions in which the user interaction takes place) in order to perform the most comprehensive evaluation.

Moreover, it is worth pointing out that, apart from the data provided by the eye-tracker, which still remains a rather expensive technology, the approach proposed has the remarkable advantage to allow evaluators to identify usability problems even if the analysis is performed remotely, which might contribute to keep at minimum the evaluation costs and allows the users to remain in their familiar environments during the evaluation, improving the trustworthiness of the evaluation itself.

Here we first discuss related work, next we indicate the architecture and the method underlying our environment and recall the main features of the first versions of WebRemUsine. Then, we present the multimodal data about the user behaviour need for performing the evaluation, and report on some example application. Lastly, some conclusions along with indications for future work are provided.

8.2 Related Work

Creating a Web site allows millions of potential users, who have diverse goals and knowledge levels, to access the information that it contains. While a Web site can easily be developed using one of the many tools available able to generate HTML from various types of specifications, obtaining usable Web sites is still difficult. Indeed, when users navigate through the Web they often encounter problems in finding the desired information or performing the desired task. With over 30 million Web sites in existence, Web sites have become the most prevalent and varied form of human-computer interface. At the same time, with so many Web pages being designed and maintained, there will never be a sufficient number of professionals to adequately address usability issues without automation [2] as a

critical component of their approach. For these reasons, interest in automatic support for usability evaluation of Web sites is rapidly increasing [6][1].

With the advent of the Web and the refinement of instrumentation and monitoring tools, user interactions are being captured on a much larger scale than ever before. In order to obtain meaningful evaluation it is important that users interact with the application in their daily environment. Since it is impractical to have evaluators directly observe users' interactions, interest in remote evaluation has been increasing. In addition, some studies [7] have confirmed the validity of remote evaluation in the field of Web site usability. Some work [3] in this area has been oriented to using audio and video capture for qualitative usability testing. In our case we provide more quantitative data and the support for their intelligent analysis. Here we present a tool for performing remote usability evaluation of Web applications. We describe how it supports analysis of task performance and Web pages accesses by the users.

8.3 The Architecture

In its first version, the core of the system architecture of our system was mainly composed of three modules (see Figure 1): the ConcurTaskTrees editor (publicly available at http://giove.cnuce.cnr.it/ctte.html) developed in our group; the logging tool that has been implemented by a combination of Javascript and applet Java to record user interactions; and WebRemUSINE, a Java tool able to perform an analysis of the files generated by the logging tool using the task model created with the CTTE tool.

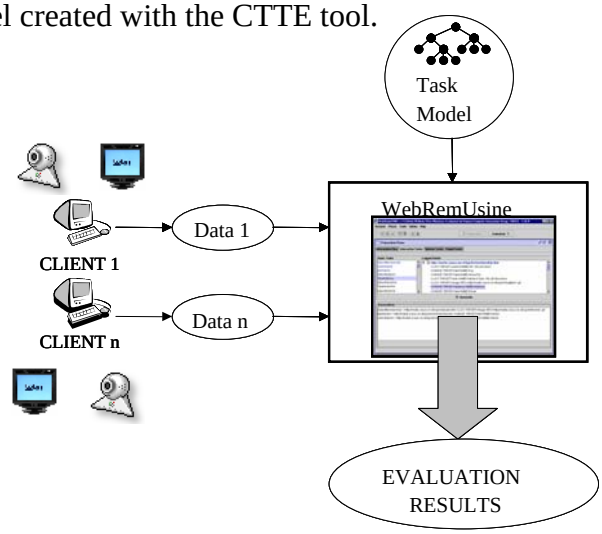


Figure 1. An Overview of the Architecture

The ConcurTaskTrees Environment (CTTE) is a tool for editing and analysing task models, which describe the activities to perform in order to reach user's goals. CTTE supports the ConcurTaskTrees (CTT) notation [5] for specifying task models, by providing a graphical representation of the hierarchical logical structure of the task model. In particular, it is possible to specify a number of flexible temporal relations among tasks (concurrency, choice, enabling, disabling, suspend-resume, order-independence, optionality, …) and for each task it is possible to indicate the objects to be manipulated and a number of other attributes. The notation also allows designers to indicate how the performance of the task should be allocated (to the user, to the system, to their interaction) through different icons.

The logging tool stores various events detected by a browser, using Javascripts that are encapsulated in the HTML pages and executed by the browser. When the browser detects an event, it notifies the script for handling it. By exploiting this communication, the script can

capture the events detected by the browser and add a temporal indication. Then, a Java applet stores the log files directly in the application server. Our browser logging tool overcomes the limitations of other approaches to logging: server logs have limited validity since various page accesses are hidden to them because of browser cache memory and they do not detect the local interactions with the user interface elements (check-boxes, type in fields, …) that also in the case of proxy-based approaches cannot be detected.

The WebRemUSINE analysis can detect usability problems such as tasks with long performance or tasks not performed according to the task model corresponding to the Web site design. The task model describes how activities can be performed according to the current design and implementation. Either the designer or the evaluator can develop it. Since ConcurTaskTrees has various temporal and logical operators, it is possible to describe all the various paths the user can follow to accomplish a task. Deviations from the expected paths can be detected in the logs and represent useful information for identifying any pages that create problems to the user. Moreover, the tool analysis provides information concerning both tasks (such as time performance, errors, …) and Web pages (such as download and visit times, …). These results allow the evaluator to analyse the usability of the Web site from both viewpoints, for example comparing the time to perform a task with that for loading the pages involved in such a performance. WebRemUSINE also identifies the sequences of tasks performed and pages visited and is able to identify patterns of use, to evaluate if the user has performed the correct sequence of tasks according to the current goal and to count the useless actions performed. In addition, it is also able to indicate what tasks have been completed, those started but not completed and those never tried. This information is also useful for Web pages: never accessed Web pages can indicate that either such pages are not interesting or that are difficult to reach. All these results can be provided for both a single user session and a group of sessions. The latter case is useful to understand if a certain problem occurs often or is limited to specific users in particular circumstances.

8.4 The Method

The method proposed is composed of two main phases, the preparation and the evaluation.

8.4.1 Preparation Phase

The main goal of the preparation phase is to create an association between the basic tasks of the task model and the events that can be generated during a session with the Web site. This association allows the tool to use the semantic information contained in the task model to analyse the sequence of user interactions. Basic tasks are tasks that cannot be further decomposed, whereas high-level tasks are complex activities composed of sub-activities. The log files are composed of sets of events. If an event is not associated with any basic task, it means that either the task model is not sufficiently detailed, or the action is erroneous because the application design does not call for its occurrence. For example, when a user sends a form then two events are stored in the log: one associated with the selection of the Submit button and the other one with the actual transmission of the form. Thus, in the task model two basic tasks are required: one interaction task for the button selection and one system task for the form transmission, otherwise the model is incomplete.

In the logs, there are three types of events: user-generated events (such as click, change), page-generated events (associated with loading and sending of pages and forms) and events associated with the change in the target task by the user, which is explicitly indicated through selection from the list of supported tasks.

Tasks can belong to three different categories according to the allocation of their performance: user tasks are only internal cognitive activities that thus cannot be captured in system logs, interaction tasks are associated with user interactions (click, change, …) and system tasks are associated with the internal browser generated events. In addition, the high-level tasks in the model are those that can be selected as target tasks by the user. Each event is associated with a single task whereas a task can be performed through different events. For example, the movement from one field to another one within a form can be performed by mouse, arrow key or Tab key. The one-to-many association between tasks and events is also useful to simplify the task model when large Web sites are considered so that we need only one task in the model to represent the performance of the same task on multiple Web pages.

8.4.2 Evaluation Phase

In the evaluation phase the proper automatic analysis is performed, where WebRemUSINE examines the logged data with the support of the task model and provides a number of results concerning the performed tasks, errors, loading time. WebRemUSINE displays all results in various formats both textual and graphical. Such information generated is analysed by the evaluators to identify usability problems and possible improvements in the interface design.

During the test phase all the user actions are automatically recorded, including those associated to goal achievement. The evaluation performed by WebRemUsine mainly consists in analysing such sequences of actions to determine whether the user has correctly performed the tasks complying with the temporal relationships defined in the task model or some errors occurred. In addition, the tool evaluates whether the user is able to reach the goals and if the actions performed are actually useful to reach the predefined goals, or a precondition error occurred, which means that the execution task order did not respect the relations defined in the system design model.

In addition to the detailed analysis of the sequence of tasks performed by the user, evaluators are provided with some results that provide an overall view of the entire session considered:

- The basic tasks that are performed correctly and how many times they have been performed correctly.
- The basic tasks that the user tried to perform when they were disabled, thus generating a precondition error, and the number of times the error occurred.
- The list of tasks never performed either because never tried or because of precondition errors.
- The patterns (sequences of repeated tasks) occurred during the session ad their frequency.

Such information allows the evaluator to identify what tasks are easily performed and what tasks create problems to the user. Moreover, revealing tasks that are never performed can be useful to identify parts of the application that are difficult to comprehend or reach. On the basis of such information the evaluator can decide to redesign the site in order to reduce the number and complexity of the activities to be performed.

From the log analysis the tool can generate various types of results:

- *Success*: the user has been able to perform a set of basic tasks required to accomplish the target task and thus achieve the goal.
- *Failure*: the users starts the performance of the target task but is not able to complete it;
- *Useless uncritical task*: the user performs a task that is not strictly useful to accomplish the target task but does not prevent its completion.

- *Deviation from the target task*: in a situation where the target task is enabled and the user performs a basic task whose effect is to disable it. This shows a problematic situation since the user is getting farther away from the main goal in addition to performing useless actions.
- *Inaccessible task*: when the user is never able to enable a certain target task.

A further type of information considered during the evaluation regards the task execution time, and the duration is calculated for both high level and basic tasks. The set of results regarding the execution time can provide information useful to understand what the most complicated tasks are or what tasks require, in any event, longer time to be performed. Longer execution time does not always imply complicated tasks, for example in some cases downloading time can be particularly high. WebRemusine provides also detailed information regarding downloading time so that evaluators can know its influence on the performance time. A further type of evaluation concerns the time associated with actions that generate errors. By analysing when errors occur, it is possible to determine if the user's performance improves over the session. If the errors concentrate during the initial part of the test and their number decreases over time, we can assume that the user interface is easy to learn.

8.4.3 Using Multimodal Information on Users: Data from Videos and Eye-tracker

In the previous sections we have provided an overview of the system, and, from the point of view of the client side, we have only considered the data coming from the log files. However, such information revealed insufficient to perform an effective evaluation because there might be a number of factors that might affect the performance of a user while interacting with an application and cannot be described just recording log files. Then, integration of such data with videos recorded during the sessions and data from eye-tracker was used to enrich the multimodal information that the system is able to gather on the user.

As for the videos, an association between task and video is automatically performed by the tool, thanks to the information regarding the starting/ending time of the different tasks. Indeed, as the whole session is recorded by a webcam, thanks to such times it is possible to split the video associated with the whole user session into different fragments related to the completion of the various tasks, together with the possibility of visualizing the related video with a suitable player in the tool, that can be activated/stopped by the evaluator. It is worth pointing out that, in order to do this it was necessary to modify the process of recording log files so as to save the information about such times. The data from videos are important in that they can provide more 'contextual' information during the performance of a task. Indeed, since the evaluation is remotely performed, the evaluator is not in a position to understand if any condition might have disturbed the performance of a task while the user visits the web site in his/her own environment. For instance, a high time (or, at least, a time higher than expected) for completing a task might not necessarily be brought about by a usability problem, but it may be caused by interruptions during the session occurring in the user's environment and due to some external factors. Another useful information that can be gained from videos are the comments of the users, that sometimes can reveal that users are aware of having performed an error but not in a position to undo the relate actions.

While videos provide more 'contextual' information regarding users, giving the means for correctly interpreting the user's actions, the eye-tracker provides technical measurements and traces of the routes that users follow while visiting a website. Indeed, the eye-tracker records the pauses and 'jumps' that normally occur when a user looks at a page (respectively called fixations and saccades), together with the so-called 'scanpaths', the traced routes of sequences of fixations and saccades, revealing the path that the user followed during the visit of the page. By superimposing the scanpath on the page it is possible to understand the strategies used by

the user to visit the page. Indeed, the evaluator can understand where the user paused to look and, alternatively, the areas of the page that did not attract his/her attention. Moreover, having in mind the target task that the user should achieve, it might be relevant to analyse the areas around the links that should be followed in order to reach the expected goal, according to the task description specified in the task model. For instance, it might be relevant to analyse the time users spent looking at these area (duration of fixations), as well as the number of accesses to such areas (if any), which is given by the number of fixations. By examining such variables it is possible to understand if users found difficulties in exploring the page: for instance, if the time duration is high, it might be a sign of user's difficulty in elaborating the information; also the number of fixations might be a sign of problems because the occurrence of some fixations on certain areas might indicate that the users are confused and are not able to find the information that they are looking for. Moreover, also the scanpath might give useful information for the evaluation: for instance, a long scanpath might indicate that the structure underlying the page is rather complicated, while, vice versa, a very short scan path indicates a rather poor structure of the page.

In addition, it is worth pointing out that, in order to manage the information associated with the eye-tracker, it was needed to modify also the logging tool and, in particular, handling scrolling events. Indeed, as soon as a scroll event is recorded, also the extent of the shifting is recorded with respect to the top and bottom corner of the page, so as to reconstruct the actual area that the user was currently looking at. Then, as we considered different sources of data for the evaluation and in order to provide the most flexibility possible, different evaluation options were considered and included in the tool. In particular, we identified a number of such evaluation options on the basis of the information that each type of evaluation considers

i)Information derived only from analysing log files;

ii)Data from log files and users' video recorded during the tests;

iii)Log files information and eye-tracker data;

iv)Log files information, videos, and eye-tracker data.

8.5 An Example

In this section we show an example of application of the proposed evaluation method. The website we considered (http://www.pisaonline.it) provides general information about Pisa (in Figure 2 the homepage is shown). The website is divided into four main sections: "Pisa da Visitare" (Visiting Pisa), "Pisa da Vivere" (Living in Pisa), Pisa da Studiare" (Studying in Pisa) e "Pisa Aziende" (Companies in Pisa).

A number of target tasks were identified, some examples are "Trova Info su Tartufo Bianco" (Find information about the white truffle), "Trova gradazione alcoolica del Chianti", (Find alcoholic content of Chianti), and so on. We involved in the test participants aging between 21 and 36.

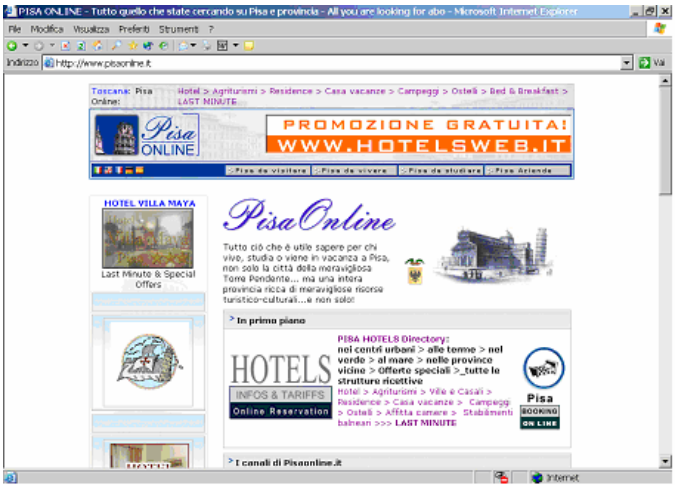


Figure 2: The homepage of the site evaluated

In Figure 3, the scanpath length of the home page of the web site is provided: as you can note, it is rather long (16554 pixels) and, looking at the superimposed page image (Figure 5) visualized for the task "Find restaurant" it is possible to see that the gaze of the user moves from one point to another indicating a possible confusion of the user who seems not able to find the right link.

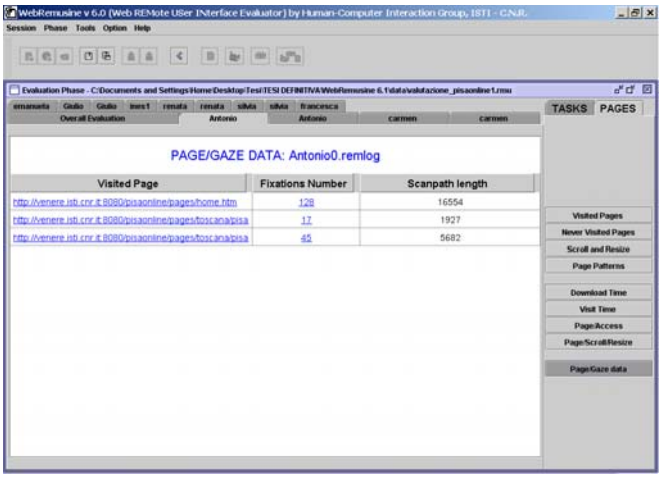


Figure 3: The window of the tool associated with Page/Gaze data

Another user who made the greatest number of errors during the navigation and selected the target task "Find information on the white truffle" associated with the abstract task "Access to Ulisse" performed a number of errors of precondition, which means that carried out a number of tasks that are no necessary according to the designer's task model for achieving the task goal. The first of such errors has been carried out when the user was visiting the page "pisa_aziende.html" (regarding the industrial companies in the area) and it is possible to note that the user pauses at looking at the area of the page dedicated to the restaurants (there is a fixation with a certain duration), while s/he does not look at all to the right link, the link related to "Pisa da visitare", which, according to the designer's task model, represents the right route for completing the selected task. From this it can be derived that the user was wrong at interpreting the name and interpreted it as a name of a restaurant.

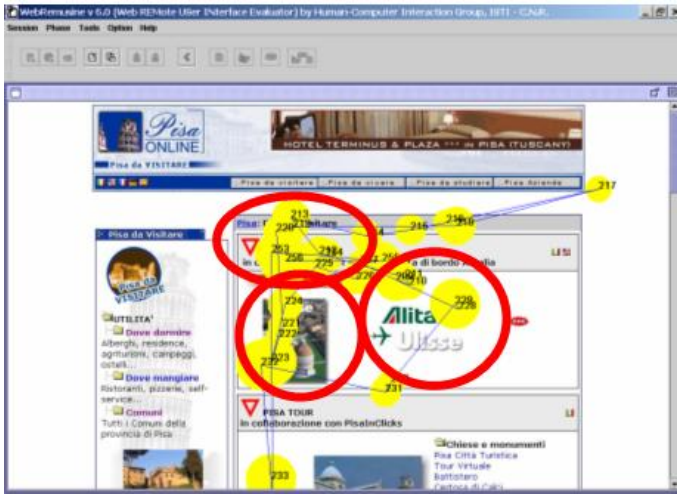


Figure 4: The ambiguity of links related to the section dedicated to "Ulisse"

The same user spent long time before selecting the "Pisa da Visitare" link and, from the associated scanpath (Figure 4) it is evident that s/he found it difficult to identify the right link among those associated with "Ulisse" (which is the Alitalia magazine). Indeed, there is a textual link (with label "Ulisse"), another textual link with a different label ("Alitalia Ulisse"), and also an icon with an image associated to Ulisse, and the information obtainable with the last two links is different from that that can be accessed through the first link.

Moreover, the experiment highlighted that the majority of users did not select the link associated with returning back to the homepage , which is rather surprising due to the relevancy of this page within the entire site. The occurrence of such behaviour in almost all users made evaluators think that the link itself could be unclear. Indeed, this intuition was reinforced by the image related to the scanpath of users on the page shown in Figure 4, where it is possible to note that the user did not pause on looking at the (image) link for going back to the homepage, although it appeared on the top-left part of the web page. A possible explanation for such behaviour was that the imagelink had been confused with a bare decorative image.

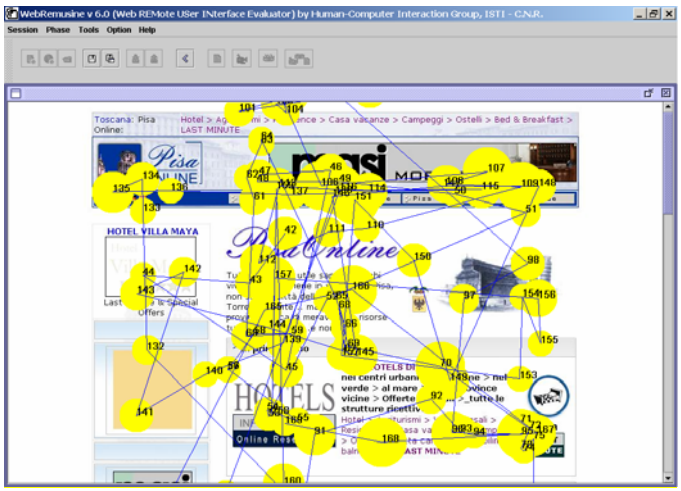


Figure 5: Scanpath of "Find Restaurant"

In addition, users who selected "Find alchoolic content of Chianti" ("Trova gradazione alcolica del Chianti", see Figure 6) as target task, which is related to the high-level task

"Scegli vino" (Select wine) tended to find such information within the section called "Companies in Pisa", rather than, more correctly, within the "Living in Pisa", where the right link actually is. Figure 5 shows that there are many fixations (202,204,205,222,etc.) on the link associated with "Pisa Aziende" ("Companies in Pisa"). This highlighted that the logic followed by users in finding such information while exploring the page is different from that followed by designers.

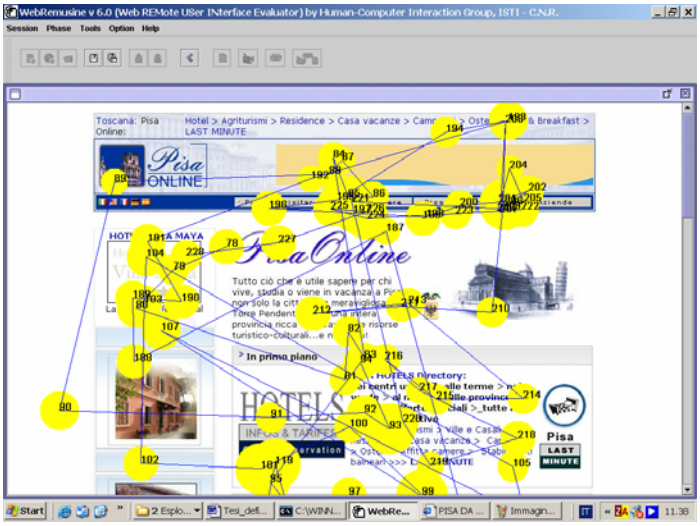


Figure 6: Scanpath of "Find Alchoolic Content of Chianti"

In another experiment we analysed a different site regarding a publishing house and mainly focused on data recorded by videos. In Figure 7 you can see the evaluation of task/time performed by the tool, regarding a user who explicitly declared at the end of the task that she was wrong at completing the task.

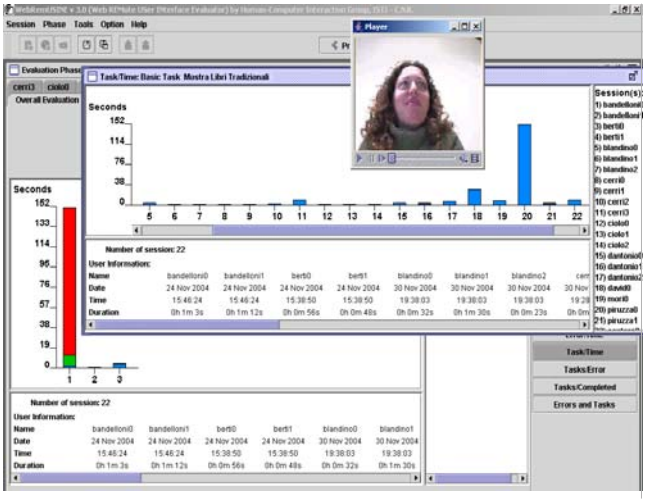


Figure 7: Task/Time Information with Video of the user

The data from the videos were useful to detect some usability problems. For instance, by examining the tasks that were wrongly performed, it was possible to have further details on the facial expressions of the users, who sometimes seemed to be confident of their choices, while other times seemed to be quite confused and doubtful, and this information is important

when evaluating the user behaviour. Particularly useful information was gained from videos as far as the execution time, which sometimes seemed to be higher that expected: the analysis of the video revealed that users pause at looking the page, then they happen to comment on it, so this is important to understand that it is not completely effective measuring the degree of difficulty of a task by just calculating the time spent in performing it, because users often are distracted/attracted by portions of the page that are not relevant for carrying out the concerned task, only by curiosity.

8.6 Conclusions

The advances in technology is more and more allowing evaluators to afford sophisticated hardware and software able to collect information about remote users interacting with webpages. We propose a method for remote evaluation of websites that, through a combination of different sources of data coming from the client side (log files, videos and eye-tracker data) allows the evaluator to get detailed information about the behaviour of the users. More specifically, the evaluator can be informed on the page the user was visiting to perform a certain task, the sequence of actions carried out by the user on the page, together with detailed information of the route traced by the user gazing at the page and background videos on the user during the session.

Such composite information is the input of an automatic tool that has been developed in our group and has been shown to be effective in providing evaluators with means for discovering possible problematic areas of the website.

Future work will be dedicated to extend the data regarding the user behaviour and state, including the emotional state, in order to have a more complete analysis of what happens during task accomplishment and better identify the potential usability issues

8.7 Acknowledgements

We thank the SIMILAR European Network of Excellence for partly supporting this work.

References

[1] Card, S., Pirolli, P., Van der Wege, M., Morrison, J., Reeder, R., Schraedley, P., Boshart, J., (2001) Information Scent as a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method for Web Usability, Proceedings ACM CHI 2001, pp.498-504.

[2] Ivory M. Y., Hearst M. A., (2001) The state of the art in automating usability evaluation of user interfaces. ACM Computing Surveys, 33(4), pp. 470-516, December 2001.

[3] Lister M., (2003) Streaming Format Software for Usability Testing, Proceedings ACM CHI 2003, Extended Abstracts, pp.632-633.

[4] L.Paganelli, F.Paternò, Tools for Remote Usability Evaluation of Web Applications through Browser Logs and Task Models, Behavior Research Methods, Instruments, and Computers, The Psychonomic Society Publications, 2003, 35 (3), pp.369-378, August 2003.

[5] Paternò, F., (1999) Model-based design and evaluation of interactive applications, Springer Verlag, 1999. ISBN 1-85233-155-0.

[6] Scholtz, J., Laskowski, S., Downey L., (1998) Developing usability tools and techniques for designing and testing web sites. Proceedings HFWeb'98 (Basking Ridge, NJ, June 1998). http://www.research.att.com/conf/hfweb/ proceedings/scholtz/index.html

[7] Tullis, T, Fleischman, S., McNulty, M, Cianchette, C. and Bergel, M., (2002). An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. Usability Professionals Conference, Pennsylvania, 2002.