DICIT: Evaluation of a Distant-talking Speech Interface for Television

Timo Sowa[1], **Fiorenza Arisio**[2], **Luca Cristoforetti**[3]

[1]Elektrobit Automotive, [2]Amuser, [3]Fondazione Bruno Kessler (FBK)-irst
[1]Am Wolfsmantel 46, 91058 Erlangen GERMANY,
[2]Via Val della Torre 4, 10149 Torino, ITALY,
[3]Via Sommarive 18, 38123 Povo (TN), ITALY
[1]timo.sowa@elektrobit.com, [2]fiorenza.arisio@amuser.com, [3]cristofo@fbk.eu

Abstract

The EC-funded project DICIT developed distant-talking interfaces for interactive TV. The final DICIT prototype system processes multimodal user input by speech and remote control. It was designed to understand both natural language and command-and-control-style speech input. We conducted an evaluation campaign to examine the usability and performance of the prototype. The task-oriented evaluation involved naïve test persons and consisted of a subjective part with a usability questionnaire and an objective part. We used three groups of objective metrics to assess the system: one group related to speech component performance, one related to interface design and user awareness, and a final group related to task-based effectiveness and usability. These metrics were acquired with a dedicated transcription and annotation tool. The evaluation revealed a quite positive subjective assessments of the system and reasonable objective results. We report how the objective metrics helped us to determine problems in specific areas and to distinguish design-related issues from technical problems. The metrics computed over modality-specific groups also show that speech input gives a usability advantage over remote control for certain types of tasks.

1. Introduction

The DICIT project addresses the development of advanced technologies for speech/acoustic processing and interpretation based on multi-microphone devices. It focuses on a novel concept of interface to TV-based home entertainment. One hallmark of this concept is natural language aiming for easier access to complex functions than the remote control (RC). Another one is the use of distant microphones relieving users from wearing any cumbersome devices and allowing them to move without restrictions.

The main topic of this article is a task-oriented evaluation of the DICIT prototype system with test persons. The aim of this study was to assess the overall usability, to evaluate the performance of components, and to assess aspects of the design. Due to the system setup and its features a tailored set of evaluation metrics had to be invented. In the next section we describe the DICIT prototype and some example interactions. We report on related work about speech system evaluation and compare DICIT to other systems in Section 3. The methodology applied in the evaluation campaign is described in Section 4 which is followed by a section about the metrics. In Section 6 we describe the tool used for annotation and analysis. The results of the campaign are described and discussed in Section 7.

2. The DICIT System

The DICIT project produced a first (interim) and a final interactive TV prototype. In this paper we always refer to the final version and its evaluation when talking about the prototype, the system, or simply about DICIT. The first prototype has also been evaluated for usability using a comparable methodology. The prototype's novel and outstanding attribute is control via speech input from the far field in addition to the remote control which is the standard for home entertainment. The DICIT system lets users give commands to "from the sofa" or from any other position within a configurable area. The functions cover basic tasks such as switching between channels or modifying the volume. Yet the most important feature is the EPG *(electronic program guide)* including a program list and a filter for the criteria "channel", "genre", "day", and "time". Also, program titles which change dynamically in the EPG are speakable. Three language-specific variants were built: for English, for Italian, and for German. Speech input, screen texts, and speech output are tailored to the respective language in each variant.

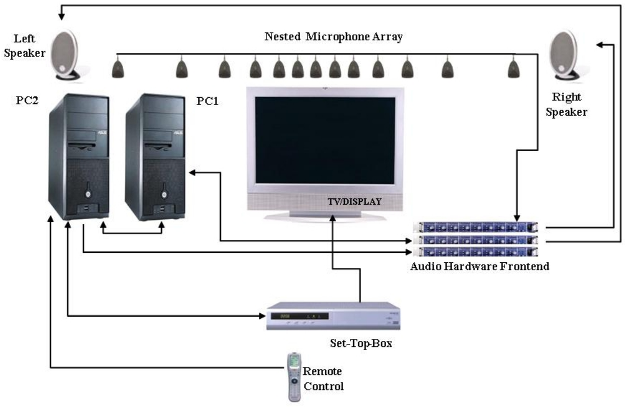
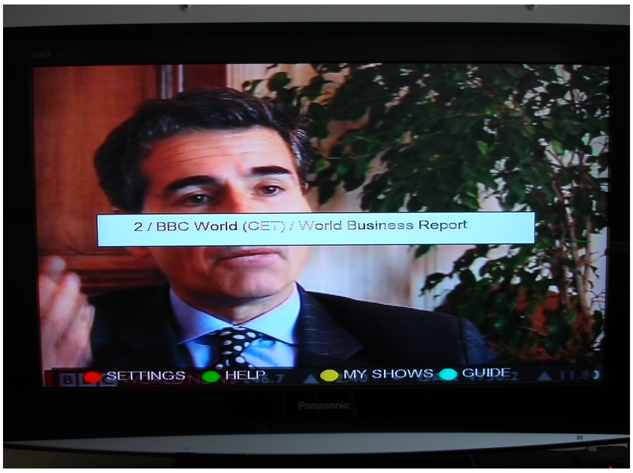


Figure 1: Prototype schema.

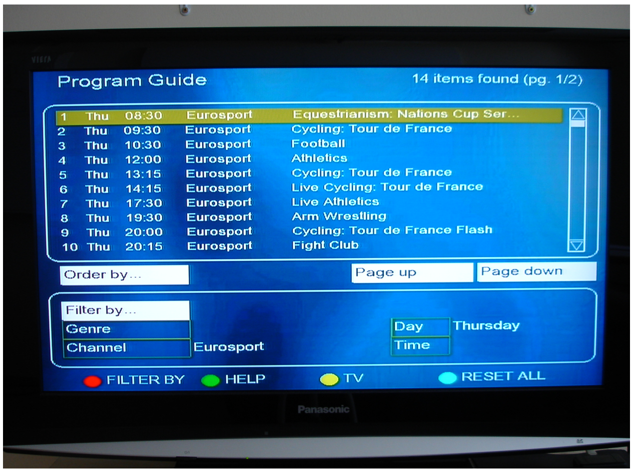
2.1. System Components

DICIT's main hardware components are two PCs, a microphone array, a Set-Top-Box (STB), and a TV set (Fig. 1). **PC 1** preprocesses the audio signals coming from the microphone array. This is done using different techniques in a cascade. At first the user is localized in the room *(source localization)* and a virtual microphone is directed towards



2 / BBC World (CET) / World Business Report

Figure 2: TV screen with channel/program information overlaid.

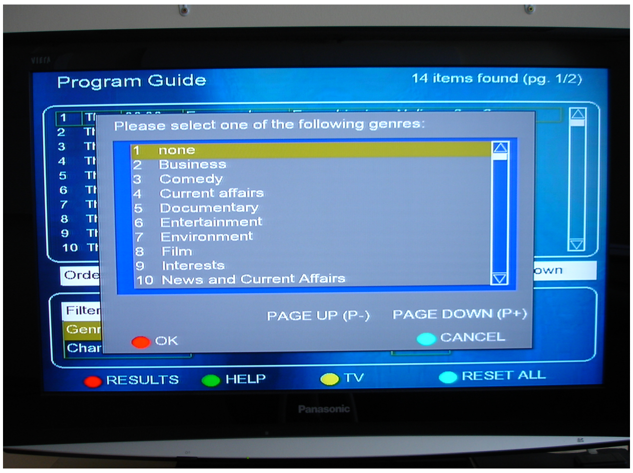


Program Guide 14 items found (pg. 1/2)

1	Thu	08:30	Eurosport	Equestrianism: Nations Cup Ser...
2	Thu	09:30	Eurosport	Cycling: Tour de France
3	Thu	10:30	Eurosport	Football
4	Thu	12:00	Eurosport	Athletics
5	Thu	13:15	Eurosport	Cycling: Tour de France
6	Thu	14:15	Eurosport	Live Cycling: Tour de France
7	Thu	17:30	Eurosport	Live Athletics
8	Thu	19:30	Eurosport	Arm Wrestling
9	Thu	20:00	Eurosport	Cycling: Tour de France Flash
10	Thu	20:15	Eurosport	Fight Club

Filter by...
Genre
Channel Eurosport
Day Thursday
Time

Figure 3: Electronic program guide (EPG) with two filter values set.



Please select one of the following genres:

1 none
2 Business
3 Comedy
4 Current affairs
5 Documentary
6 Entertainment
7 Environment
8 Film
9 Interests
10 News and Current Affairs

Figure 4: Genre selection via overlay popup screen.

him, fusing together all the signals of the microphone array *(beamforming)*. The resulting signal could be contaminated by the audio TV output and needs to be cleaned *(echo cancellation)*. The last step is to identify relevant voice segments *(smart speech filtering)* and to send them to the second PC, where high-level processing takes place. **PC 2** is the mind of the system. An *automatic speech recognizer* (ASR) produces a word chain which is interpreted by a *natural language understanding* (NLU) module. The NLU module employs statistical models mapping recognized word chains onto parametrized actions. A *dialogue manager* (DM) executes the actions (requested via voice or remote control by the user) according to the current dialogue state, commands the STB, and plays audio messages to the user.

2.2. Sample Interaction

The DICIT system either shows the TV screen or a menu screen; either of which can be overlaid with additional selection or information boxes. All visual displays are rendered by the Set-Top-Box on the dialogue manager's request. Suppose a user is currently watching TV and instructs the system to "display the current channel" with this voice command. Then an overlay info box displaying the name of the program and the channel would appear for a couple of seconds (Fig. 2). Then assume the user says "what's on Eurosport on Thursday". DICIT would switch to the EPG main screen showing a list of Eurosport shows on Thursday in the upper part and the filter criteria in the lower part (Fig. 3). Note that two filter values (*channel* and *day*) are set while the other two are still empty. Next the user decides to refine the filter and add a specific genre by saying "select genre". The system responds by saying "please make your choice" and displaying a selection box with the list of genres (Fig. 4). The user may choose by speaking one of the (dynamic) entries or the line number, browse by saying "next page", or cancel. Even "cursor down" and "OK" would work as speech commands to select the second entry.

The example illustrates that DICIT allows both complex speech commands, sometimes called *one-shot* or *multi-slot* commands, but also simple speech input. Note that DICIT does not engage the user in a pure speech-driven dialogue. In every state of the interaction the user may switch the input modality and may continue, to resume the example, by selecting the genre "Entertainment" with button 6 on the RC.

3. Related Work

A lot of progress on methodologies for the evaluation of dialogue systems and their components was made in the past two decades. Metrics to assess speech recognition and understanding in dialogue systems are more or less standardized. Furui (2008) provides a summary of such metrics. For continuous speech recognition and understanding typically measures based on the alignment of recognized word chains or concept structures to reference structures are used. The ratio of mismatches (insertions, deletions, substitutions — possibly weighted) to the length or size of the reference yields suitable metrics such as WER (word error rate). Besides such accuracy measures one particularly important factor is what Furui calls *situation awareness*. This refers to whether users know what they can say in a given situation or not. The factor and corresponding metrics thus provide information about the design of the system.

Methodologies and metrics to assess dialogue management strategies can be found in (Danieli and Gerbino, 1995). The PARADISE framework suggested in (Walker et al., 1998) ties subjective and objective metrics together and provides a method to determine objective predictors for subjective usability. Evaluation has also become an integral part of projects dealing with speech understanding and dialogue systems. Examples for evaluation campaigns of research prototypes were discussed, for instance, by Bernsen and Dybkjær (2008) who illustrate the methodology applied to the edutainment system HCA and the on board system SENECA SDS for cars. Lamel et al. (1998) describe the evaluation of the multimodal (speech and touch) information kiosk MASK to be used for travel inquiries. The evaluation campaigns for SENECA and MASK share some similarities with the work presented here, because one of the common aims was to find out whether speech and/or multimodality is superior to "standard" non-speech input modalities in the respective application domains.

As for systems similar to DICIT there are some voice-controlled devices already commercially available. Products such as VoiceMe by Hotech can learn about 100 different voice commands and bind them to arbitrary sequences of infrared signals to control the TV/VCR or any other infrared device. VoiceMe allows distant talking up to five meters. The Remote by Amulet Devices has a built-in microphone which transmits the speech signal to a PC/Set-Top-Box. The receiving system runs a software based on Windows Media Center. Amulet Remote accepts speech input for TV and VCR functions such as switching channels or selecting a recorded program. Speech recognition is activated by tilting/holding the RC in an upright position. In contrast to DICIT these systems seem to be quite limited with respect to vocabulary size, flexibility of grammar, and complexity of the functions to be voice-controlled. They do not offer "natural language" in the sense that users may express a command in many different ways. Another difference between DICIT and most commercial speech systems (as well as some research systems) is that DICIT does not require the user to activate the system or explicitly open the microphone and start recognition (via "push-to-talk" button or similar means).

4. Experimental Methods

The basic paradigm for the evaluation campaign is a user study with naïve participants, i.e., subjects who are neither involved in the development of the system nor have extensive background knowledge about speech technology. In order to get results about the expected performance and adequacy of the interface in everyday life, participants should operate the system in a typical environment of use, and they should not be disturbed or influenced by the experimenter or another person. The evaluation of the final prototype took place at all DICIT partner sites in Italy (three sites), Germany (two sites), the Czech Republic, and the USA (one site each). This was done to test the three language-versions with native speakers and to check whether different environmental conditions have an impact on the results. At each site, dedicated experiment rooms arranged in a comparable way were set up (Fig. 5). The subject was sitting or standing in front of the TV at 2.5 meters, with the microphone array placed above the TV at 1.5 meters from the ground.[1]

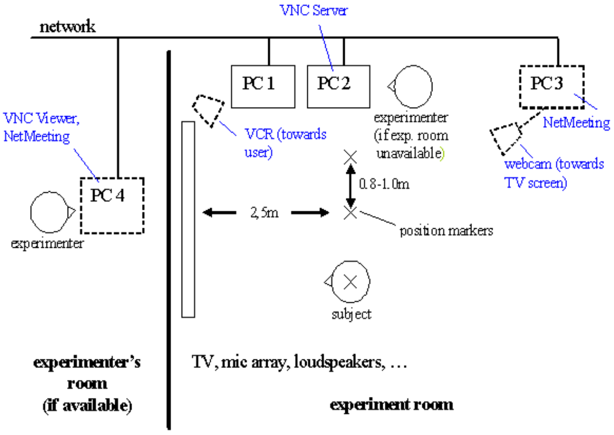


Figure 5: Room setup.

An important aspect which has a potential impact on measurements of performance is the influence of the rooms. Although we tried to do the evaluation in similar rooms in terms of geometry, they had different characteristics. While some rooms showed a good noise insulation and low background noise, others were noisy office rooms with running computers and air conditioning inside. Also the reverberation time was quite different from site to site. Given all these elements, we tried to equalize the behavior of the prototype through a detailed calibration (Marquardt et al., 2009). This procedure involved the use of a sound level meter to fine adjust hardware gains and software tools to determine the best parameters of the system running in that particular room.

The experimenter could monitor the evaluation from another room or stayed in the experiment room outside the subject's view and outside the system's recognition area. All the experiments have been recorded by a video camera pointed to the user. We used a task-based paradigm that covers some of the most frequent tasks TV and EPG users are confronted with.

4.1. Subjects

The total number of subjects is 171 of which 50 were native speakers of Italian, 51 native speakers of German, 15 native speakers of US (American) English, 18 native speakers of other kinds of English, and 37 nonnative speakers of English. Special care was taken that the subject set covers all age groups, people with different educational background, and that people professionally involved in speech processing or multimedia were excluded.

[1] One site in Germany had a slightly different setup, since a video beamer was used instead of a TV. The mic array was placed underneath the projection area.

4.2. Procedure

The overall duration of a session, which depended a lot on the subject, was between 90 and 180 minutes — though most subjects needed 120 minutes and more. The usability test session consists of five main parts as described in the following. Note that the durations for the parts are only rough estimates:

1. Training phase (35 minutes) – Subjects were given the opportunity to train using the system.

2. Usability test (45 minutes) – For the main usability evaluation phase the subject was asked to solve 16 tasks of five types. The tasks were ordered with increasing complexity. Each task type covers one typical use case of the DICIT system and had to be solved within 120 seconds.

3. Complete Questionnaire (15 minutes) — Subjects were asked to complete a usability questionnaire after the test to elicit attitudes about the application in general. This test is what Sauro (2010) calls *perception satisfaction* in contrast to *performance satisfaction* which is related to single task performance.

4. Acoustic frontend test (20 minutes) – In order to test the localization algorithm of the acoustic frontend, subjects were asked to stand, give speech commands prompted by the experimenter and change the position after each command.

5. Finalize Questionnaire (5 minutes) – Finally we asked for some personal data.

4.3. Training

To have a complete system's overview, a video was shown before the subjects began to use it, to let them know all the system's features. At some evaluation sites the video was not available. Here the experimenter gave a live demo of the system. The demo followed the same script used to create video to make sure that the contents of video and demonstrations are comparable. After the video/live demo subjects were given the opportunity to train using the system, with some hands-on experience before the real test part of the experiment started. They were invited first to control DICIT with the RC (with speech recognition turned off), and then with voice commands. Subjects could freely "play around" with the system, without any particular goal, but the experimenter suggested to try some important features in case they did not explore them by themselves.

4.4. Tasks

In order to evaluate whether there is some real advantage of speech input as compared to a "traditional" TV set, and to understand if voice as a shortcut for complex functions is easier to use than RC, each task had to be solved under either of three conditions: using only voice commands (V), only the remote control (RC), or with free choice of the modalities (VRC). The VRC mode is to find out whether the two modalities integrate smoothly and whether people make use of both when given the choice. Modalities were balanced across tasks, i.e., the same task was solved almost equally often under each condition. Each modality (V, RC, VRC) was used once for each task. When the task was to be solved by voice only or voice plus RC, subjects were requested to avoid simply reading the task description. The task wording and sentence structure was deliberately variegated to avoid any impact on the commands chosen by the subjects. For instance, one task to set filter values in the EPG was formulated as a command: "Try to find out what shows you can watch on CNN on Sunday!" Another task of the same kind was formulated indirectly as a question: "Would you mind searching for some Tuesday afternoon programs about traveling?" We used the following five task types:

- Task Type 1 – Using Basic Commands in the TV Screen (go to a specific menu screen, adjust volume, change channels).

- Task Type 2 – Modifying the Settings (change the DICIT voice, switch off the system voice, change the prompts' style).

- Task Type 3 – Filtering the Program List in the EPG Screen (get the schedule for a specific station, find shows on a specific channel and day, find some programs that belong to a certain genre in a certain time span, and search for specific genres on a certain day and time of day).

- Task Type 4 – Doing a Program Search from the TV Screen (use a single command ("one-shot") to find the schedule for a specific day, then search for programs of a certain genre on a specific channel at a certain time of day).

- Task Type 5 – Programming/Selecting Specific Shows in the EPG Screen (try to program in advance a certain show that was specified with 3–4 different items, like hour, day, genre or channel, or only with the title).

4.5. Questionnaire

The questionnaire consists of 71 questions according to the criteria of DIN EN ISO 9241-110 (ISO, 2006). The first part is about specific parts of the DICIT system, such as screen, voice output, and voice input. Further questions concern users' expectations about the appeal of DICIT, an overall impression of the system, and concludes with a semantic differential. The second part contains questions on the frontend evaluation, statistical questions and questions regarding habits watching TV. With questions on the appeal of DICIT we ask whether a subject would buy the system if it was commercially available, how much he/she would pay, and whether the system meets his/her expectations. The questions thus aim towards a hypothetical introduction of this technology in the consumer market.

5. Objective Metrics

The choice of objective metrics is guided by the aims of the evaluation campaign. The metrics thus capture speech component performance, task-related usability and effectiveness, and design- and awareness-related issues as follows.

5.1. Speech Component Performance

These metrics are used to evaluate the successive stages of speech input processing. Metrics to assess acoustic preprocessing are mandatory. This is because speech input for distant recordings is noisier than in close-talk situations. Furthermore DICIT uses no explicit "start" signal for speech input. Thus, acoustic processing could be a significant source of error on its own — which is why we test it with the first two metrics. The scope of this group of metrics is a single spoken utterance.

Speech event detection reflects the ability of the system to capture valid speech input by the user and reject other sources of sound. Values per utterance can be *OK*, if the user's speech has been detected correctly, *missed*, if valid speech input is not taken up, or *false positive*, if the system detected something which is not speech by the subject (e.g., background noise).

Segmentation provides information about DICIT's ability to find the correct beginning and end of a spoken utterance. Values can be *OK*, if the audio fragment contains a complete utterance, *cut*, if something is missing, *split*, if something is missing which can be found in the preceding or successive utterance, *cut/split* if a snippet is both cut and split, or *joined*, if the audio fragment contains two separate utterances.

Word recognition rate (WRR) is a numerical value computed in the usual way with $WRR = 1 - \frac{I+D+S}{N}$, where I, D, S are the number of word insertions, deletions, and substitutions, and N is the number of words in the reference.

Action classification rate (ACR) assesses the "natural language"-mapping from recognition result to an action. Values can be *correct*, if the action and its parameters are correctly classified, and *incorrect* otherwise.

System reaction finally evaluates the response of the system viewed as a "black box". Possible values are *correct*, if the system performs the action the user intended, *incorrect* otherwise.

Note that some speech performance metrics are directly influenced by the room acoustics: **speech event detection** and **segmentation** depend on the smart speech filtering that is mainly influenced by background noise. The reverberation time influences source localization and hence the beamforming. The **word recognition rate** depends on the speech recognizer and is influenced by both background noise and reverberation time. Acoustic models for the speech recognizer were based only on data recorded in a single room. This implied slightly worse performance in other rooms.

5.2. Interface Design and User Awareness

Metrics on interface-design and awareness primarily relate to the appropriateness of an input by the user in different senses. Assuming that all subjects use DICIT in a goal-oriented, constructive way (as requested by the experimenters), any "inappropriate" input would point at a mismatch between the subject's mental model of the application on one hand and reality on the other. The consequence for a system designer would be either to make the user aware of the problematic part of system's working, i.e., training, or to change the design as to meet an untrained user's expectations (Dix et al., 2004, pp. 49–51). In any case we consider it important to identify inappropriate input in order to clearly separate technical problems from errors caused by design or insufficient awareness. Thus we have introduced the following metrics with the scope of a single spoken utterance (or RC input for goal focus).

Plausibility represents whether a subject's utterance generally addresses the capabilities of the system — regardless of the current dialogue state. Values can be *plausible* or *implausible*. Implausible utterances are, for instance, out-of-domain utterances or self-talk. This measure reflects the awareness for principal limitations of DICIT.

Coverage assesses the availability of the system request expressed in the user's utterance in the current state of the system. Values are *available*, if the intended action was available in the current dialogue state, or *unavailable*, if not. It represents *situation awareness* as described in Section 3.

Goal focus assesses whether the execution of a single action (by speech or RC) brought the subject closer to the goal of the current task. Values are *closer* and *not closer* respectively. We consider this metric to be a measure for the awareness regarding the effect of DICIT's functions.

5.3. Task-related Usability and Effectiveness

Task-related metrics to asses the adequacy of an interface for certain tasks are used here as in many other evaluations of speech systems. In contrast to the other two groups they do not refer only to speech input, but to all kinds of task executions regardless of modality such that comparisons between modalities are possible.

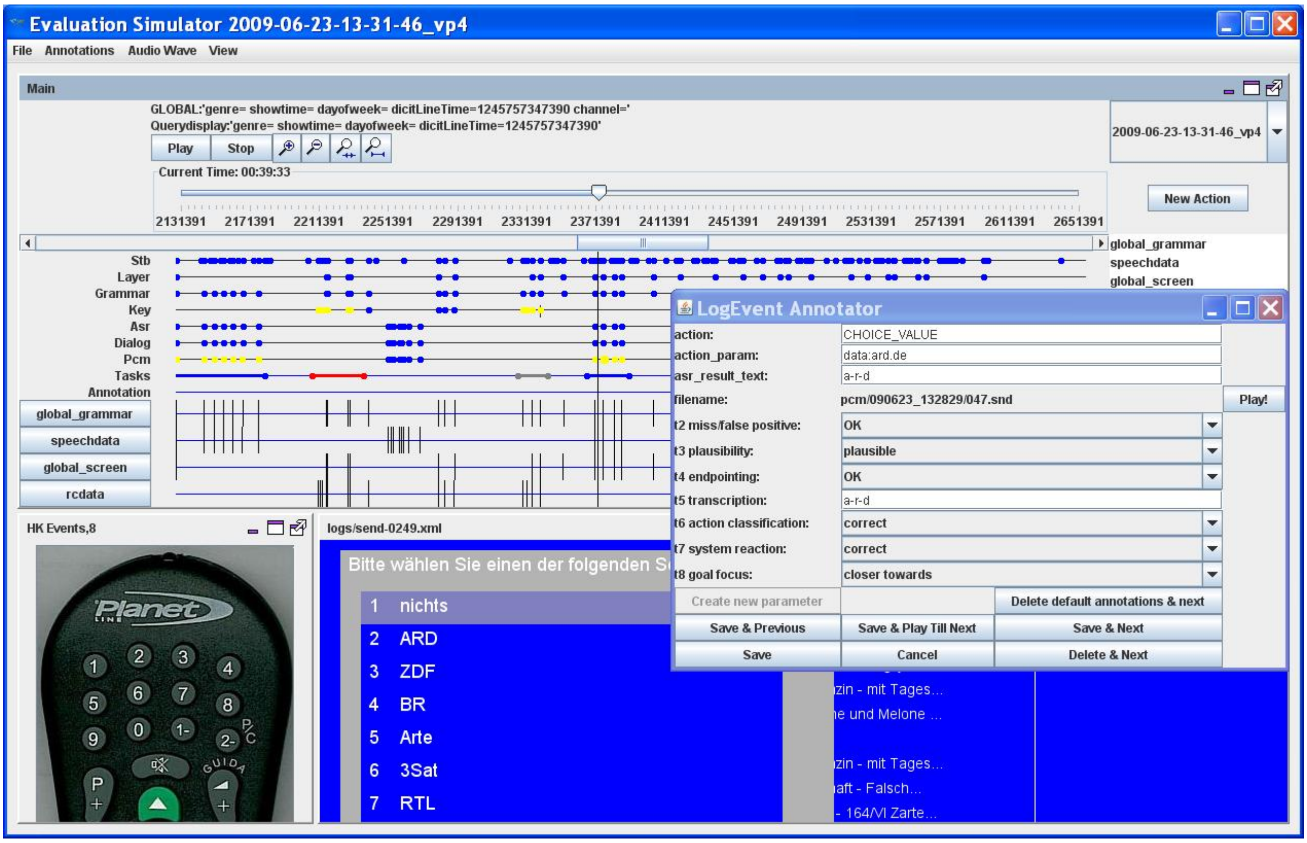
Task completion rate (TCR) signals whether a task been solved, i.e., whether a defined end-state of the system was reached.

Task completion time (TCT) is the time to successfully complete a task.

Number of turns (NOT) is the amount of user inputs (in either modality) needed to complete a task.

6. Tool-support for Annotation and Analysis

For data logging, annotation, and analysis of the metrics described in the previous section tools were developed that greatly simplified our work. During the experiments a data logger stored all user input, system output, and important processing steps of the dialogue manager in log files. An annotation and analysis tool reads the log files and displays user/system interactions in a multi-tier time-aligned view. An earlier version of the tool is described in (Wesseling et



GLOBAL:'genre= showtime= dayofweek= dicitLineTime=1245757347390 channel='
Querydisplay:'genre= showtime= dayofweek= dicitLineTime=1245757347390'

Current Time: 00:39:33

action: CHOICE_VALUE
action_param: data:ard.de
asr_result_text: a-r-d
filename: pcm/090623_132829/047.snd
t2 miss/false positive: OK
t3 plausibility: plausible
t4 endpointing: OK
t5 transcription: a-r-d
t6 action classification: correct
t7 system reaction: correct
t8 goal focus: closer towards

Figure 6: Main screen of the tool for annotation and analysis. Upper part: time-aligned event view, lower part: RC and TV screen simulation, smaller popup window for speech annotation.

al., 2008). It is flexibly configurable and can also be used for other systems than DICIT.

Fig. 6 shows the main screen of the annotation and analysis tool. The upper part shows logged events as small dots. Each event type is displayed in a separate tier. By right-clicking on a dot with the mouse an annotation window for the corresponding event is opened. Such a window for a speech event is shown in the screen shot. Here the annotator can see the recognition result (a-r-d, a German channel name), the action executed by the system (CHOICE_VALUE, for selecting an entry in a list by speaking it's name), and the parameter (data:ard.de, for the channel). Speech input can be replayed with the "play" button. The annotator enters a transcription of the speech snippet and sets the variables for the metrics discussed before, for instance "action classification".[2] Each speech input was transcribed using the tool. From the logged recognition result and transcription the tool computes the word recognition rate as defined before. With convenience functions like "save & next" the currently edited annotation is saved and the tool proceeds to the next event in the same tier. Already annotated items are displayed with yellow (light) dots.

The simplified annotation procedure requires that speech events were correctly detected and segmented by the system. However, this cannot be guaranteed, since valid speech inputs may have been missed. In order to capture missed events a special feature of the tool can be used which displays the wave form of a complete session. Those portions of the wave form which belong to speech events are marked. The annotator can now inspect and listen to unmarked parts in which significant sound activity is visible. Using this method, listening to the complete session became unnecessary and that saved time.

7. Results and Discussion

7.1. Subjective Part (Questionnaire)

The overall tendency for the responses to questions regarding system usability, screen design, and speech interaction was positive, or very positive (depending on the language). Subjects like the idea of speech interaction in a TV scenario and, at least on average, did not have particular trouble with the screen design and the usage of the prototype. Subjectively the voice interaction is preferred over the remote control, also for simple tasks such as changing the channel or volume, but mainly for using the EPG and searching programs. As for the general opinion, the subjects' experiences with the DICIT prototype are positive: they think that it is easy and fun to use and they attribute the adjectives "original", "friendly", "organized", and "polite" to the DICIT system. Even though the objective results indicate that the system's speech capabilities are far from being perfect (see below), many subjects were surprised by the naturalness and complexity of speech input DICIT accepts (in compar-

[2] A selection field for coverage does not appear in the annotation window. Coverage is implicitly selected with the values for action classification.

Metric	Total	Male	Female
Speech event detection	96.2%	95.9%	96.7%
missed	2.8%	3.1%	2.5%
false positive	1.0%	1.0%	0.9%
Segmentation	92.3%	93.3%	90.8%
cut	1.1%	1.2%	1.0%
split	5.8%	4.7%	7.3%
cut/split	0.7%	0.7%	0.7%
joined	0.1%	0.0%	0.1%
Word recognition rate	56.7%	61.0%	51.0%
native speakers	61.8%		
Action classification	62.3%	65.7%	57.9%
System reaction *correct*	60.3%		
Plausibility	94.2%	95.0%	93.1%
Coverage	96.4%	96.3%	96.4%
Goal focus	70.8%	72.6%	68.7%
speech	69.8%		
RC	71.7%		
Task completion rate	83.9%	84.2%	83.5%
Task completion time (s)	43.9	43.8	44.1
Number of turns	8.6	8.5	8.7

Table 1: Summary of the objective evaluation.

ison to the RC as input device). The subjective results show an astonishingly positive attitude towards the DICIT speech interface and subjects rated speech comprehension abilities on average with 7 on a 1–10 scale (10 is best).

7.2. Objective Part

The complete data set analyzed for objective metrics comprises 11.181 spoken utterances, 16.431 key strokes, and 2.152 task executions by 140 subjects. Table 1 shows a summary of the metrics. The upper part shows speech component performance. On average, the system reacted correctly for about 60% of all detected speech inputs with about 62% correct action classifications. It is no contradiction that ACR values are higher than corresponding WRR values, because in statistical recognition not all words contribute to the "core" meaning of a phrase. Even if all words but one meaning-bearing word are misrecognized, the statistical method could still classify the action correctly. There is room for improvement of the rate of correct system reactions, but the 60% seem to be acceptable for users considering the subjective feedback and ratings for the speech comprehension abilities (see above).

The objective metrics show higher WRRs for men than for women (usually female voices are recognized worse than male voices). The difference is also present in ACR, but a weak WRR is possibly not the only reason for this. Another reason could be differences between men and women regarding the attitude/habits towards using technical devices (as stated in the questionnaire).

The metrics assessing the acoustic frontend further show that there are more missed inputs than false positives. Ideally these two values should have been equal so the system could be tuned a bit more sensitive. Another, more significant issue is segmentation. In particular, there are about 6% split utterances. In connection with the very small join rate it implies that the system reacted too quickly to small speech pauses and stopped recognition too early.

The design- and awareness-related metrics in the middle part reveal possible problems with the subjects' mental models of the application. From the values for plausibility and coverage we infer that subjects are aware of general limitations in terms of speech input and of the constraints in certain dialogue situations. However, the rate of goal-directed input is just 71% with little difference between the input modalities. Note that for speech input the 69.8% refer to *correctly* executed actions. This metric indicates that, regardless of modality, subjects had problems understanding the effect of an action. One consequence could be to provide better explanations or training and an error recovery procedure for complex functions such as the program filter.

With a task completion rate of about 84% it could be shown that the DICIT interface is at least effective such that most tasks could be solved by most subjects within a reasonable time span. Table 2 shows task completion rates, times, and number of turns according to modality and task. Note that only 14 of the 16 tasks are included, because two tasks could only be solved with speech. To improve readability, absolute RC results (percent for TCR, seconds for TCT, number for NOT) are provided as a baseline. For the conditions V and VRC only the difference to RC is provided. Note that improvements are positive numbers for TCR, but negative numbers for TCT and NOT.

When comparing task completion rates (columns 2–4) between modalities the results are inconclusive. It is not evident that V or VRC always leads to significantly better or worse task completion than RC. Tasks 5c and 5d are exceptions. Here subjects had to look for a specific program in a long list. With the remote they browsed in the list (up to 10 pages of 10 entries each) and it happened often that subjects were not fast enough or simply overlooked the entry. With speech they could just say the name of the program even if it was not currently displayed, and a the system was looking for a match in the entire list.

As for task completion times (columns 5–7) the results are inconclusive for the simpler task groups 1 and 2. For some tasks, TCT even increased under condition V and likewise under VRC. The latter seems to suggest that people try using speech even though it is less efficient for some tasks. However, the table shows a very clear decrease of TCT for voice and multimodal input for the more complex task groups 3 and 5 (with just one exception for VRC).

The most significant improvement when comparing the conditions including voice to RC only is the number of turns (columns 8–10). For every task except one where the turn value remains the same, NOT is lower for conditions V and VRC than for RC. The decrease is particularly clear for task groups 3 and 5 where it amounts to more than 70%.

Since DICIT allows short and simple, but also complex *one-shot* speech requests (cf. 2.2.), we evaluated whether users preferred one of these two input styles over the other. For that purpose we counted the length, i.e., the number of words per speech utterance. It turned out that about 37% of all utterances have a length of one. More than 60% were one- or two-word utterances. So the majority did not take advantage of complex speech inputs, but preferred simple

Task	TCR (%)			TCT (seconds)			NOT		
	RC	V	VRC	RC	V	VRC	RC	V	VRC
1a	98.0	-3.9	-5.1	39.7	+15.8	+8.8	7.5	-1.2	0.0
1b	97.6	-1.5	-1.5	23.6	-6.7	-4.2	10.9	-8.8	-6.8
1c	96.0	-0.9	-1.9	37.7	+14.9	+10.7	7.1	-0.8	-1.3
2a	92.7	-0.4	+3.3	40.1	+0.6	-0.2	9.3	-2.3	-3.2
2b	96.1	+3.9	-3.4	37.1	-5.6	-1.3	8.3	-3.7	-1.7
2c	91.7	-8.8	-3.2	44.4	+7.7	+12.9	10.7	-3.1	-1.3
3a	92.0	-1.8	+1.5	47.5	-9.3	-6.2	14.9	-10.4	-8.5
3b	80.4	+9.6	-0.4	59.9	-23.0	-21.9	20.5	-14.9	-14.3
3c	84.4	-6.8	-3.9	54.9	-7.2	-2.4	21.0	-15.0	-11.7
3d	77.6	-0.1	-0.9	78.3	-19.7	-23.8	28.3	-20.5	-18.1
5a	50.0	-1.3	+8.5	76.5	-24.2	+0.1	25.5	-18.5	-12.7
5b	60.0	-8.7	-0.4	84.6	-16.7	-18.8	26.4	-17.1	-14.2
5c	44.7	+32.8	+31.4	57.1	-22.2	-24.2	18.9	-14.2	-13.4
5d	59.5	+25.5	+25.3	56.1	-19.5	-13.9	16.1	-10.9	-8.3

Table 2: Task completion rates, times, and number of turns.

commands. This “step by step” behavior is probably due to an interaction style that imitates using the remote control instead of the more powerful “shortcuts” provided with the voice interface. However, the number of complex utterances is not negligible and it is much higher for the current prototype than for the first prototype (which was inferior in terms of speech comprehension).

8. Conclusions

The usability and performance of the distant-talking TV system DICIT has been evaluated in an extensive study in three languages and with 171 subjects of different age groups. We have chosen an evaluation methodology which covers both an analysis of the system’s design and user awareness as well as an analysis of the functional components. That way, design-related issues could be separated from technical issues. As for technical metrics we found that, besides speech recognition, speech segmentation was one significant source of error. The metrics related to user awareness highlighted deficits regarding training or understanding of the system’s functions. We received quite positive subjective feedback from the subjects and could show that speech input in the TV/home entertainment domain positively affects usability when compared to remote control input alone. Results on utterance lengths indicate that short commands should be available and need to work properly before a speech system can be extended into the direction of natural language and multi-slot commands.

9. Acknowledgements

This work was partially funded by the Commission of the EC, Information Society Technologies (IST), FP6 IST-034624, under DICIT.

10. References

N. O. Bernsen and L. Dybkjær. 2008. Spoken dialogue systems evaluation. In L. Dybkjær, Holmer Hemsen, and Wolfgang Minker, editors, *Evaluation of Text and Speech Systems*, pages 185–219.

M. Danieli and E. Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 34–39, Stanford.

A. Dix, J. Finlay, G. D. Abowd, and R. Beale. 2004. *Human-Computer Interaction*. Pearson Education, Harlow, England, 3rd edition.

S. Furui. 2008. Speech and speaker recognition evaluation. In L. Dybkjær, Holmer Hemsen, and Wolfgang Minker, editors, *Evaluation of Text and Speech Systems*, pages 1–27.

ISO. 2006. Ergonomics of human-system interaction – part 110: Dialogue principles. ISO Standard 9241-110, International Organization for Standardization.

L. Lamel, S. Bennacef, J. L. Gauvain, H. Dartigues, and J. N. Temem. 1998. User evaluation of the MASK kiosk. In *Proceedings of ICSLP’98*, pages 2875–2878, Sydney.

L. Marquardt, E. Mabande, A. Lombard, K. Reindl, Y. Zheng, M. Schneider, A. Brutti, P. Svaizer, and W. Kellermann. 2009. MC-AEC and BSS algorithms for an advanced distant-talking ASR front-end optimized for an interactive TV scenario. DICIT Project Deliverable 3.2, DICIT Consortium.

J. Sauro. 2010. Performance satisfaction and perception satisfaction. http://www.measuringusability.com/blog/-test-task-sat.php. Web blog on measuring usability.

M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. 1998. Paradise: A general framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of ACL*, pages 271–280, Madrid.

H. Wesseling, M. Bezold, and N. Beringer. 2008. Automatic evaluation tool for multimodal dialogue systems. In *Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems (LNAI 5078)*, pages 297–305, Berlin. Springer.