D6.4: Final Report on Multimodal Experiments Part I: Evaluation of the SAMMIE System

Hartmut Mutschler, BEF
Frank Steffens, Andreas Korthauer, BOSCH

Final 1.1

Distribution: public

TALK

Talk and Look:
Tools for Ambient Linguistic Knowledge
IST 507802 Deliverable 6.4

25 January 2007

Project funded by the European Community
under the Sixth Framework Programme for
Research and Technological Development.

The deliverable identification sheet is to be found on the reverse of this page.

Project ref no.	IST-507802
Project acronym	TALK
Project full title	Talk and Look: Tools for ambient linguistic knowledge
Instrument	STREP
Thematic Priority	Information Society Technologies
Start date / duration	01 January 2004 / 36 Months

Security	Consortium internal
Contractual date of delivery	M36 = December 2006
Actual date of delivery	20 December 2006
Deliverable number	D6.4 (Part I)
Deliverable title	Final Report on Multimodal Experiments – Part I: Evaluation of the SAMMIE system
Type	Report
Status & version	Final 1.1
Number of pages	96
Contributing WP	WP6
WP / Task responsible	BOSCH
Other contributors	BEF (c/o BOSCH), BMW, DFKI, USAAR
Author(s)	Hartmut Mutschler, Frank Steffens, Andreas Korthauer
EC project officer	Evangelia Markidou
Keywords	Evaluation, final in-car showcase, multimodal experiments

The partners in TALK are:

Saarland University USAAR
University of Edinburgh HCRC UEDIN
University of Gothenburg UGOT
University of Cambridge UCAM
University of Seville USE
Deutsches Forschungszentrum für Künstliche Intelligenz DFKI
Linguamatics LING
BMW Forschung und Technik GmbH BMW
Robert Bosch GmbH BOSCH

For copies of reports, updates on project activities and other TALK-related information, contact:
The TALK Project Co-ordinator
Prof. Manfred Pinkal
Computerlinguistik,
Fachrichtung 4.7 Allgemeine Linguistik
Postfach 15 11 50
66041 Saarbrücken, Germany
pinkal@coli.uni-sb.de
Phone +49 (681) 302-4343 - Fax +49 (681) 302-4351

Copies of reports and other material can also be accessed via the project's administration homepage, http://www.talk-project.org

Contents

Executive Summary

The TALK deliverable D6.4 splits into two parts:

1. The first part concentrates on the evaluation of the final SAMMIE in-car system.
2. In the second part we report on the data collection experiments SAMMIE, MIMUS and SACTI. Moreover, we present results from the evaluation experiments using the TownInfo system.

This part of deliverable D6.4 reports on the results of the evaluation of the final SAMMIE in-car system. A user test was performed in an experimental car with the SAMMIE system and a Command&Control-like reference system (C&C). The SAMMIE dialogue system with its evaluated variants and the C&C system as well as their integration into the BMW car is described in detail in TALK deliverable D5.3 [1].

The objectives of the evaluation study were to find out, how efficient the Final In-Car Showcase SAMMIE system for the interaction with a MP3 system in a car is being used and to what extent it is accepted. 21 Subjects performed two runs with SAMMIE and the C&C system on a 19 – 35 km course with 7 – 10 tasks. The experimental design also allowed for a comparison with the corresponding evaluation of the In-car Baseline system (cf. TALK deliverable D6.3 [2]).

When directly comparing the results of both studies, it is important to note that the evaluation conditions for the Baseline system were different from the evaluation of the final SAMMIE system:

- The Baseline system has been evaluated as a laboratory prototype using a simulated driving task, whereas the final evaluation took place in the BMW car under real driving conditions on the road.
- Due to the missing vestibular feedback of acceleration, the simulated driving task in the Baseline study was unfamiliar and more demanding to some Subjects than the real driving task in the final evaluation.
- A head set with a close-talk microphone was used for the Baseline system versus a far-talk microphone array for the final SAMMIE evaluation, resulting in noisier speech signals for the speech recognition and language understanding.
- The conditions for task completion were more restrictive in the final evaluation, as the tasks were linked to fixed segments of the experimental course, i.e. tasks were considered as failed if not successfully completed within the given course segment.

Following is a summary of the main results for the final SAMMIE evaluation:

Task completion: The task completion rate (TCR) reached a level of about 80%. This has to be interpreted as a general high level, considering the partly tight time and driving conditions. The tasks with SAMMIE were completed somewhat (but not significantly) more frequently than the tasks with C&C. The SAMMIE TCR was about 6% above the baseline TCR. Considering the different conditions of the present study with a tighter schedule for the tasks to be performed, this is a clear advantage of the SAMMIE system over the Baseline system. Often a combination of understanding, dialogue and system problems was the reason for not completed tasks, particularly by less experienced Subjects.

Dialogue efficiency: Frequently the users did not choose the direct and shortest dialogue and they took a considerable number of iDrive actions. Significantly more turns on average were necessary to complete a task with the C&C system (5,4 turns) than with the SAMMIE system (4,9 turns). Considering the complexity of most of the tasks, this still seems to be an acceptable level.

One task performed with SAMMIE and C&C took about 40 – 50 s on the average. The minimal task durations were about 10 s – 12 s. The comparable tasks in the baseline study, however, took clearly longer.

Driving quality and mental load: About 2,5 driving errors per minute occurred without a pronounced difference between the systems (SAMMIE and C&C). Lane departures and low speeds were the most frequent driving errors and can be attributed to the visual distraction when observing the display. The subjectively judged driving quality was nearly equal for both systems, which confirms the objective driving quality results. A comparison with the Baseline system is not applicable.

The mental load was on a generally low level of about two (scale 1 – 5). There was no difference in mental load between the systems. Higher scores resulted from operating the MP3 system within a demanding traffic situation and in the context of dialogue or speech recognition problems.

Modality preference: Basically, the multimodal combination of speech and manual input was extensively used. At the beginning of a task, there was a very clear preference for speech input with both systems. With ongoing interactions while performing the tasks, there was a clear reduction in speech preference. MP3 experienced Subjects tended to use speech more than the less experienced Subjects and vice versa for iDrive.

Subjective ratings: With the present systems by far most of the Subjects tended to a positive judgement of the multimodal interaction systems. I.e., there was a clear improvement concerning the subjective overall impression from the Baseline to the SAMMIE systems, the more so as the present systems were judged to be easier to use than the baseline system.

SAMMIE was assessed to be less distracting and more comfortable than the C&C and Baseline system. The decision for a certain modality and the change between modalities was easy for most of the Subjects. This is an important result in favour of the concept of multimodality, since a change between modalities at pleasure is easily possible.

Overall, speech output and the display were judged relatively positively. The information output, however, was not fully accepted with regard to liking, support, information distribution and assistance.

Concerning the dialogue there was a tendency to a positive judgment. SAMMIE was generally better judged than C&C. We used statements from the COMMUNICATOR evaluations [4] to assess aspects of the dialogue. The best scores got the statement concerning the understanding of what the system said. Restrictions referred to the statements, that it was easy to get the information which the user wanted and that the system worked as expected. The Subjects who participated already in the baseline study often stated spontaneously an increased performance of the present systems as compared to the Baseline system.

Subjectively the most important advantage of the multimodal input was avoiding the problems of one modality by choosing the other. Consequently, the “free choice of the operation mode” was rated positively by a considerable part of the Subjects.

Recommendations: Finally recommendations are given concerning the multimodal interaction concept, system performance and system output. The most important ones are the following:

- Pursue the concept of multimodality with free choice of modality at any time.
- Keep the concept of barge-in by Push-to-Talk button and possibly extend the concept with respect to modality changes from speech to iDrive
- Further improvements of speech recognition and language understanding performance are needed with regard to acoustic conditions, large vocabulary and grammar coverage. This is considered an important aspect of multimodal systems featuring speech dialogue.
- Reduce amount and length of the speech output to the necessary information.
- Keep the display as it is but leave out unnecessary information.

1 INTRODUCTION

Within the TALK project the multimodal interaction system SAMMIE (TALK In-Car Showcase) had to be evaluated within a user field test. The objectives were to analyse

- the usage of the multimodal systems (choice of modality),
- the dialogue efficiency (Task Completion Rate, number of turns, dialogue times),
- the acceptance of the system (questionnaires with subjective evaluation),
- the efficiency of the speech system (false reactions, rejections).
- The influence onto driving quality (driving errors, driving scores).

The main variable was the multimodal interaction system. The Full SAMMIE system had to be compared to the Command&Control (C&C) system as the reference system as well as to the baseline system. The Non-Adaptive (NA) SAMMIE system should be included into the evaluation, too.
For a more detailed description of the evaluated system variants and their specific features see deliverable D5.3 [1]; the results of the baseline system evaluation can be found in deliverable D6.3 [2].

The study was conceived as critical experiment. I.e. hypotheses were defined on the basis of the results of the baseline study and other deliberations. Moreover, additional results were expected concerning the multimodality and efficiency of the SAMMIE system.

Essential aspects of the methods were:

- system variants
- experimental set-up
- experimental course
- Subjects
- evaluation tasks
- experimental design, realization
- measurements, questionnaires

Following hypotheses were established:

1. Users prefer speech input more with the SAMMIE system than with the C&C system
2. Users with much MP3 experience tend to manual operation
3. Users with much MP3 experience achieve a higher operation efficiency, particularly with a lower number of turns
4. Users get a higher Task Completion Rate with SAMMIE than with C&C
5. Users are faster with the SAMMIE system than with the C&C system
6. The number of turns is higher with C&C than with SAMMIE
7. SAMMIE needs less iDrive actions
8. The number of system errors with SAMMIE is only marginally higher than with C&C
9. SAMMIE distracts the user less from driving than C&C
10. The SAMMIE system leads to a higher user acceptance than the C&C system
11. Users can assess well what the system has understood

2 EVALUATION DESIGN

2.1 Experimental set-up

The basic components of the experimental set-up were (s. Figure 1 and Figure 2):

- Experimental car (BMW 335)
- SAMMIE system with microphones, loudspeaker and iDrive
- Two cameras for Subject and traffic scene
- Split screen and video recorder
- Additional electronics
- Data recorder, keyboard, writing-pad

The exterior elements of the SAMMIE system were a microphone for speech input and the iDrive device for manual input. [1] In contrast to the baseline system, the microphone could be opened by the user by means of activating the Push-To-Talk button (PTT) at the steering wheel or automatically by the system during the dialogue. Opening and closing of the microphone was indicated by slightly different acoustical signals and a large green/red microphone icon on the display. With an additional button the dialogue could be interrupted optionally. The MP3 display (SAMMIE display) showed the MP3 elements and the list of artists / songs / albums etc. (s. Figure 3).

The iDrive button allowed several operations: Turning (2 directions), pushing (1 direction) and shifting (4 directions). Turning induced scrolling of the cursor and pushing activated the pronounced item. Shifting upwards led to a higher menu level or another former display presentation. Shifting downwards paused the playing song. Shifting to the left or right side changed to the preceding or next song.

The Subject camera recorded the Subject including his body motions and manual activities. The scene camera recorded the traffic scene.

The split screen displayed the Subject, the traffic scene as well as the MP3 display, together with the actual date and time. The split image was recorded by a VHS video recorder.

The time of the laptop, the video recorder and the extra clock for the supervisor were synchronized to get a uniform time base.

The supervisor was sitting beside the Subject and was guiding through the course. He monitored the driving safety by observing and warning in potentially dangerous situations. [2] Moreover, he noted some essential data (rough task times, chosen modalities, task completion) and identified most of the driving errors, which he signalised to the experimenter for registration. He noted all relevant times and events on an experimental sheet.

The experimenter was sitting behind the supervisor. She supervised the experimental set-up, announced the tasks and activated F-keys on the keyboard to stamp the exact times of task beginning and ending. Moreover, she registered the driving errors by means of a data recorder (communicator). In the event of system crash or hang-up she activated a reset.

[1] “iDrive” is used as a synonym for ergocommander

[2] Dangerous situations occurred very rarely and accidents could be avoided easily by this additional control.

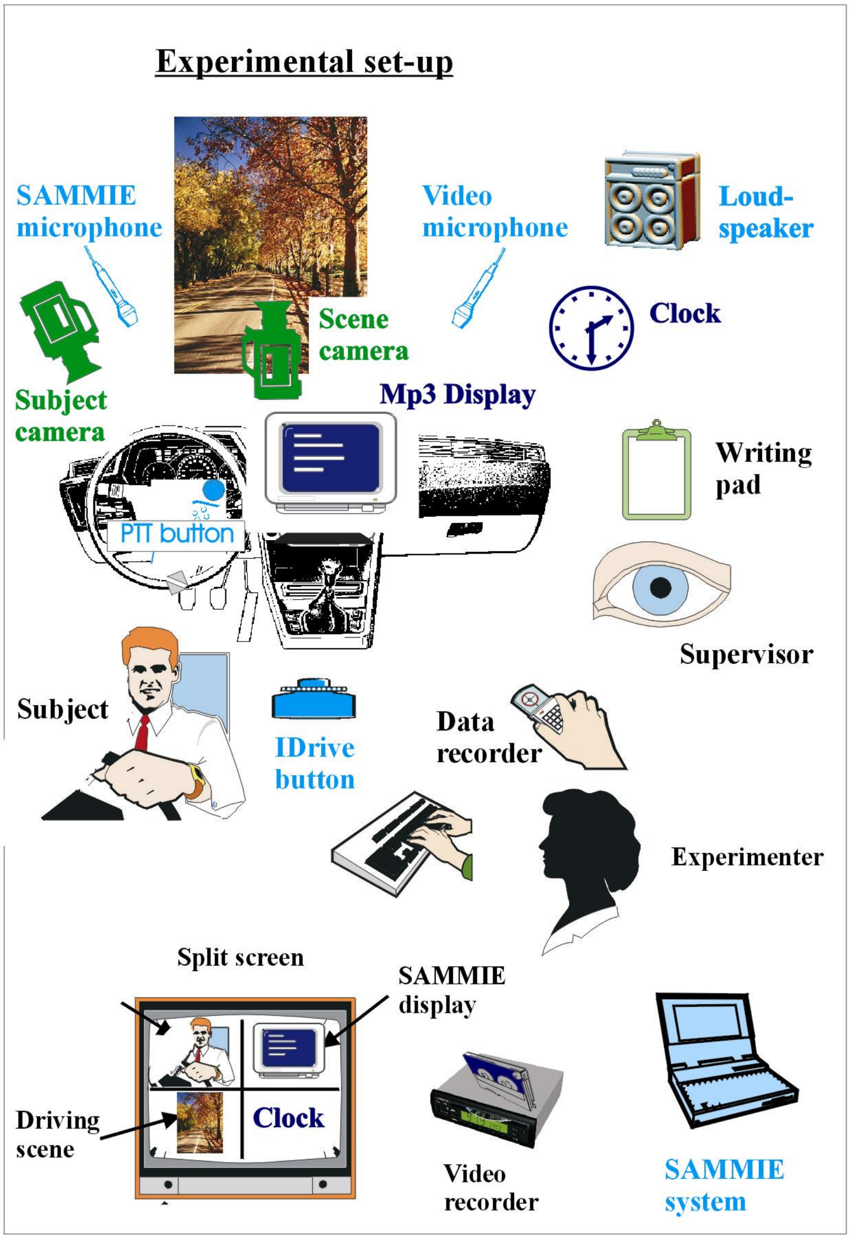


Figure 1: Experimental set-up

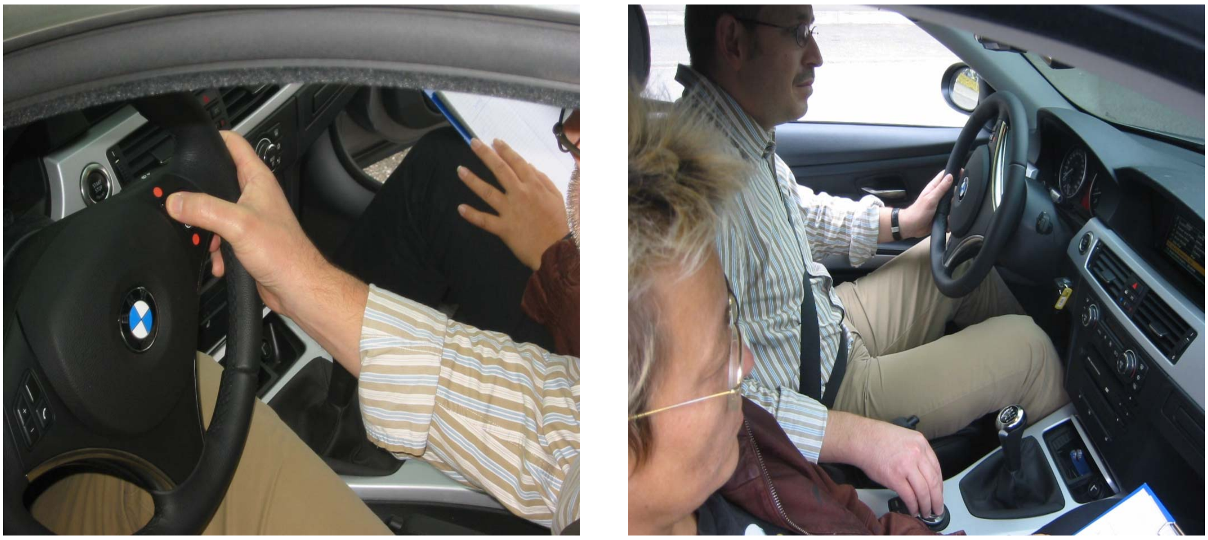
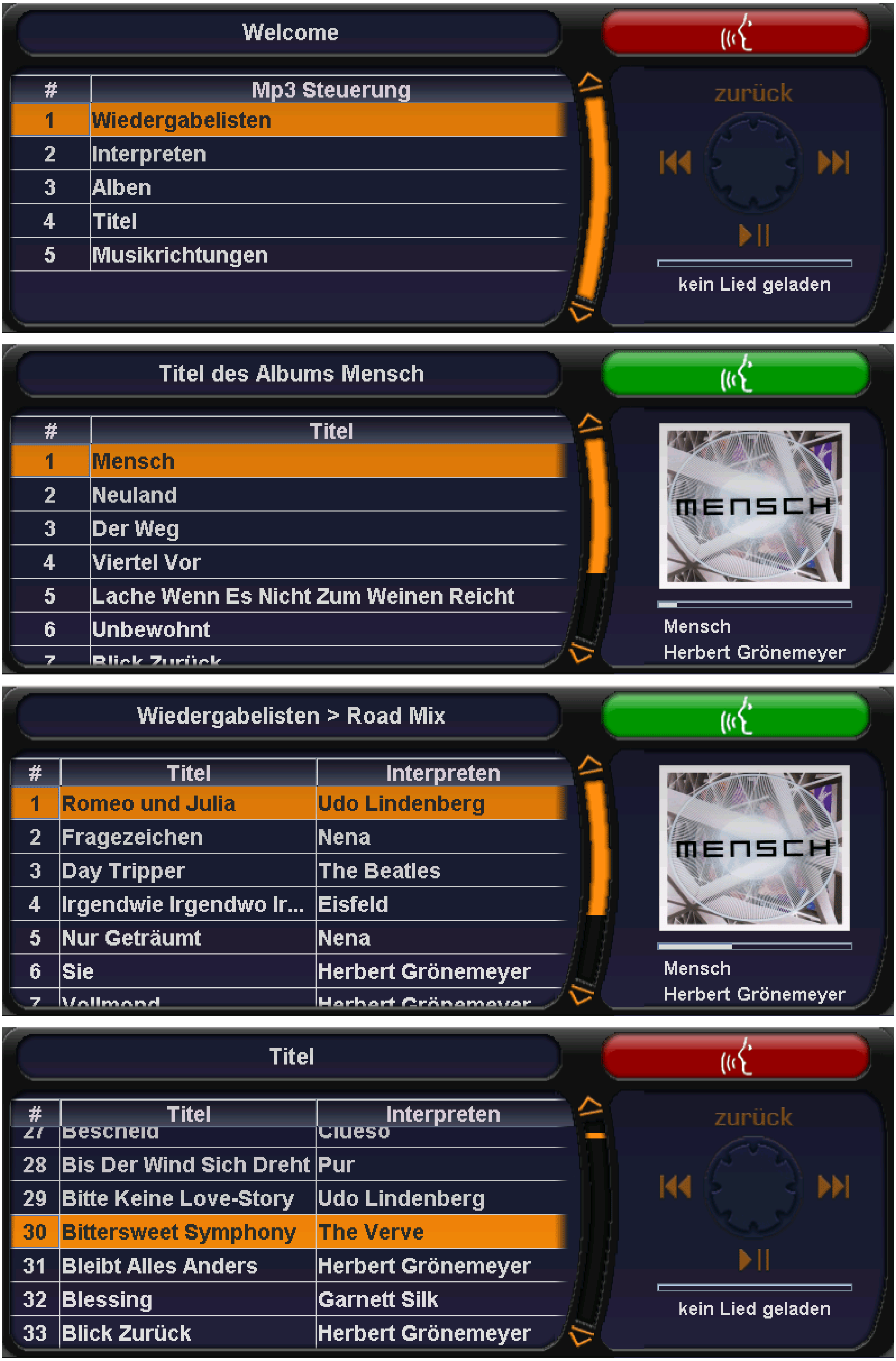


Figure 2: Operating the PTT and iDrive button



Welcome

#	Mp3 Steuerung
1	Wiedergabelisten
2	Interpreten
3	Alben
4	Titel
5	Musikrichtungen

zurück

kein Lied geladen

Titel des Albums Mensch

#	Titel
1	Mensch
2	Neuland
3	Der Weg
4	Viertel Vor
5	Lache Wenn Es Nicht Zum Weinen Reicht
6	Unbewohnt
7	Blick Zurück

Mensch
Herbert Grönemeyer

Wiedergabelisten > Road Mix

#	Titel	Interpreten
1	Romeo und Julia	Udo Lindenberg
2	Fragezeichen	Nena
3	Day Tripper	The Beatles
4	Irgendwie Irgendwo Ir...	Eisfeld
5	Nur Geträumt	Nena
6	Sie	Herbert Grönemeyer
7	Vollmond	Herbert Grönemeyer

Mensch
Herbert Grönemeyer

Titel

#	Titel	Interpreten
27	Bescheid	Clueso
28	Bis Der Wind Sich Dreht	Pur
29	Bitte Keine Love-Story	Udo Lindenberg
30	Bittersweet Symphony	The Verve
31	Bleibt Alles Anders	Herbert Grönemeyer
32	Blessing	Garnett Silk
33	Blick Zurück	Herbert Grönemeyer

zurück

kein Lied geladen

Figure 3: Examples of MP3 (SAMMIE) displays

2.2 Experimental course

The experimental course had to meet following criteria:

- Long enough to allow about 10 tasks
- Short enough to keep the overall session time within 2,5 – 3 hours
- Starting and ending the drive at a point, which was easily accessible to the Subjects (main station of Karlsruhe)
- No hard driving and traffic situations (no sharp curves, not too much traffic)
- Preferring speed limits of ≤100 km/h to keep traffic noise within limits (no motorways)
- Preferring express highways with 4 lanes or country roads with few traffic for long or complex tasks to keep oncoming traffic within limits
- Roads with a certain amount of changing speed limits to study the distraction from driving while operating the SAMMIE systems (´provocation of driving errors´)
- Avoiding approach roads and traffic lights within the task segments as far as possible (no forced interrupts at approach roads, no task completion in standing car)
- Some structuring for setting task begin and end marks

The resulting experimental course is shown in the next table and figure. The distance was 34,5 km for the SAMMIE run, which was shortened to 19 km for the C&C run (s. below). A typical Subject needed about 35 – 40 min to drive the SAMMIE run, and about 20 – 25 min to drive the C&C run. The task segments of the course had two lanes with few or medium traffic or four lanes with medium or dense traffic. The task segments had not more than two approach roads (within task 5) and no traffic lights within the task processing. Mostly, there were speed limits to 70, 80 or 100 km/h, which changed in the majority of segments.

Within the pre-tests the distances for the tasks were chosen to allow the complete performance of a task, when no hard driving or dialogue problems occurred. Since the criteria of having as much tasks as possible was more important than enabling long task performance times, several tasks followed close to each other and/or had a rather limited performance time, particularly tasks 1-5 and 10. As consequence task 2 was started immediately after the ending of task 1, even if the mark (traffic sign “Oststadt”) had not been reached, yet.

The traffic density was low in the course segments of tasks 3, 4, 6 and 7, whereby the mostly two lanes were rather narrow. Altogether, there was a pronounced load either by oncoming traffic on narrow streets or by more traffic at higher speed. Very high loads were not given (e.g. by much traffic on curvy roads or by very high speeds).

In some course segments the Subjects were free to operate the systems autonomously.

Time	km	Task	Traffic signs etc.	Characteristics	Permitted speed
00:00	**0**	**Start**	**Station, parking**		
Express highway (Südtangente)					
01:00	1,2	**Task begin**	**After approach road to express road**	**4 lanes, straight on, much traffic**	**80**
		1. Task "Albums"	several traffic signs 80 etc		
		Task end	traffic sign "Oststadt"		
02:00	2,3	**Task begin**	traffic sign "Oststadt"	**4 lanes,**	**80, 70**
		2. Task "Song Der Weg"	several signs 80 etc., passing several exits	**straight on,**	
		Task end, turning right	end of express highway --> B3	**much traffic**	
Main road B3					
04:00	4,3	**Task begin**	**After approach road to B3**	**2 lanes,**	**80**
		3. Task "Playlist Pur Klassik	traffic sign 80	**wide curves,**	
		Task end	**traffic lights**	**few traffic**	
06:00	6,4	**Task begin**	**After traffic lights**	**2 lanes,**	
		4. Task "Live by Pur"	several traffic signs "free" etc.	**straight on,**	**100, 70**
		Task end	**Hedwigshof**	**few traffic**	
07:00	8,5	**Task begin**	**After Hedwigshof**	**2 lanes, 4 lanes,**	**70**
		5. Task "Swing song"	several traffic signs 70 etc.	**several approach roads,**	
		2xturning to the right	2x approach roads	**yield right of way,**	
			several traffic signs "free" etc.	**medium traffic**	**100, 70**
		Task end	**traffic light, traffic sign "Rastatt"**		
10:00	10,9	Free interaction	**2 -4 lanes, several approach roads, yield right of way, medium traffic**		
Country road L506					
14:00	15,2	**Task begin**	**After approach road to L506**		**80**
		6. Task "99 Luftballons"	several traffic signs 80 etc.	**2 narrow lanes,**	
				straight on,	
			traffic sign 50	**few traffic**	**50**
		Task end	**railway crossing**		
19:00	18,8	**Free interaction**	**2 lanes, 4 lanes, approach road, medium traffic**		
Main road B36					
Country road K3581					
23:00	22,6	**Task begin**	**after approach road**	**2 narrow lanes,**	
		7. Task "Song Yesterday"	several traffic signs 70 etc.	**straight on,**	**70, 50**
			roundabout, tunnel, traffic sign "Light!"	**roundabout, tunnel,**	
		Task end	traffic sign 70	**few traffic**	**70**
26:00	25,2	**Task begin**	**Traffic sign "free"**		**100,80,60**
		8. Task "New playlist"	several traffic signs 80 etc.	**2 narrow lanes,**	**100, 70**
			passing several exits	**wide curves,**	
		Task end	**traffic sign "Karlsruhe"**	**medium traffic**	
		Turning to the right			
28:00	26,5	Free interaction	**4 lanes, several approach roads, yield right of way, medium traffic**		
Express highway (Brauerstr)					
30:00	28,7	**Task begin**	traffic sign "Skidding"		**130,100,70**
		9. Task "Romeo and Julia"	traffic signs 100 etc.	**4 lanes,**	
			passing several exits	**straight on,**	
		Turning right	**Exit "Wolfartsweier"**	**medium traffic**	
Express highway (Südtangente)					
33:00	31,5	**Task begin**	traffic sign 80		**80**
		10. Task "Rock song"	2 tunnels, passing several exits	**4 lanes,**	
		Task end, turning to the right	Exit "Hauptbahnhof"	**straight on,**	
				much traffic	
37:00	34,5		Station, parking		

Table 1: Experimental course with tasks, segments and details

only for Full SAMMIE

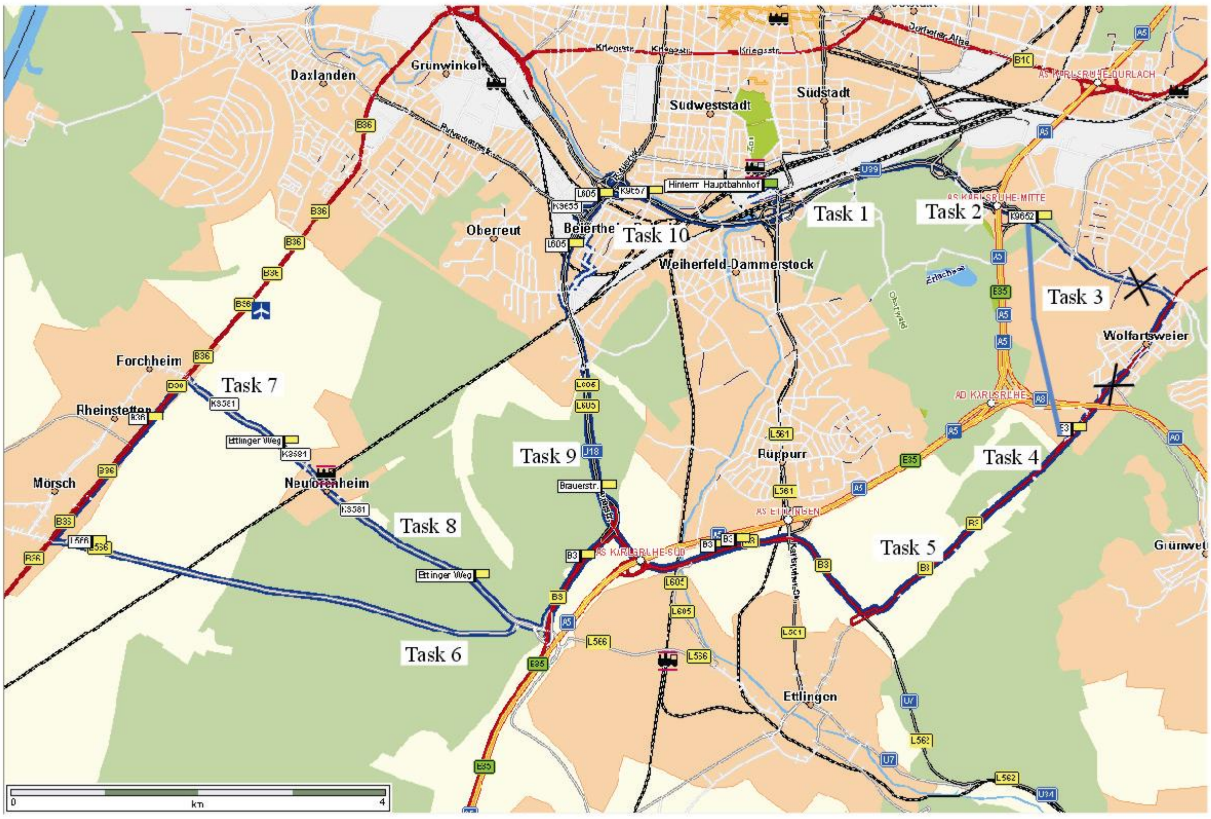


Figure 4: Experimental course as map

2.3 Subjects

A sample of 21 Subjects was recruited (s. Table 2) [3]. Essential requirements for the participation were:

1. Some or much experience with MP3 players or similar software
2. Participation in the baseline evaluation study, if possible
3. Very safe driver
4. Regular driving experience
5. Capable to avoid any strong dialect
6. Involved in former BEF-studies, if possible [4]
7. Knowledge of local roads, if possible

No specific design with other Subject parameters was envisaged, but a certain variance in sex and professional background was aspired (not too much technicians). The age was practically limited to the young and middle age group, because of the conditions 1. and 3.

As the following table shows, there were 10 Subjects, who had some MP3 experience ("1") and 11 Subjects, who had much MP3 experience ("2"). 'Much MP3 experience' means "Having already used an iPod" or "Using regularly an MP3 hardware or software system".

[3] 20 Subjects were originally planned for the evaluation study. The 21. Subject was included as a reserve.

[4] The standard sample of BEF ensures safe driving, reliability and some kind of a sophisticated expressiveness.

Not more than 11 Subjects of the baseline study met the conditions and were at disposal. [5] I.e., 11 Subjects already participated in the baseline study, 6 Subjects participated in other BEF studies, e.g. in the VICO field study. The average age was 36,2 years with a range from 20 to 56.

Relatively many Subjects had a technician background, which means here engineer or software specialist. [6] This has to be handled as a bias within the experiment. [7]

Most Subjects had an actual driving experience of at least 7000 km/year [8] and assessed themselves at least averaged experienced (rating scale 1 – 5, with 5=maximal experience).

No.	**Short name**	**BEF studies**	**Sex**	**Age**	**Technician background**	**MP3 experience**	**Driving experience [km/year]**	**Self-assessment**
1	**Hau**	**Baseline**	**female**	**46**	**no**	**1**	**7500**	**5**
2	**Hol**	**Baseline**	**female**	**36**	**yes**	**1**	**9000**	**4**
3	**Eig**	**Baseline**	**male**	**50**	**yes**	**1**	**25000**	**5**
4	**Opi**	**Baseline**	**male**	**31**	**yes**	**1**	**15000**	**4**
5	Bof	VICO	female	50	**yes**	1	25000	3
6	Hof	VICO	female	47	no	1	10000	5
7	Gön	VICO	male	37	no	1	30000	5
8	Ben	other	male	56	no	1	´--	4
9	Ose	other	male	35	´--	1	15000	4
10	Rau	new	female	23	no	1	10000	3
11	**Ros**	**Baseline**	**female**	**39**	**no**	**2**	**2500**	**3**
12	**Beh**	**Baseline**	**male**	**38**	**yes**	**2**	**20000**	**4**
13	**Dis**	**Baseline**	**male**	**49**	**yes**	**2**	**10000**	**5**
14	**Hat**	**Baseline**	**male**	**41**	**yes**	**2**	**12000**	**4**
15	**Rot**	**Baseline**	**male**	**31**	**no**	**2**	**14000**	**3**
16	**Sau**	**Baseline**	**male**	**21**	**yes**	**2**	**2000**	**3**
17	**Sch**	**Baseline**	**male**	**21**	**yes**	**2**	**8500**	**2**
18	Zsc	VICO	female	36	yes	2	7000	3
19	Bot	new	male	20	yes	2	15000	3
20	Kru	new	male	27	yes	2	20000	3
21	Pla	new	male	27	yes	2	7000	4

Table 2: Subjects of the SAMMIE experiment [9]

[5] Four persons of the baseline study were excluded in advance, because of more than one self induced accident within the last years or other reasons.

[6] Subjects 9 did not specify his profession beyond a general statement "employee".

[7] The motivation of those technicians were very high. They participated in the experiment sometimes even within their working hours. Nearly all of this subgroup had an academical background, partly still studying.

[8] Those persons with a low actual driving experience were included, because they participated already in the baseline study without having had much accidents (11, 16) or were known from other BEF-experiments as safe drivers (8). Subject 8 does not own a car.

[9] The numbering does not correspond chronologically to the session order. But roughly speaking, most of the persons with low numbers performed their session in the first part, most of the persons with high numbers performed their session in the second part of the study.

2.4 Tasks

The basic principles for defining the tasks were [10]

- to use a considerable number of tasks from the baseline study
- to cover the performance of the SAMMIE system
- to include tasks with pure information content
- to consider more demanding functions, i.e. neglecting the simple play back functions
- to choose items that do not require lengthy scrolling in the displayed lists

So, Bosch and BEF together with the partners defined the following tasks for the test (Attachments

).

SAMMIE number, SAMMIE Task	*Baseline number*
1. Ask for the existing albums	**1.4**
2. Play back the song ´Der Weg von Herbert Grönemeyer´	**1.3**
3. Find out the songs on the playlist ´Pur Klassiker´	**3.3**
4. Browse within the albums, search for the album ´Live´ by Pur and play it back	**1.5**
5. Find and play back a Swing song by Michael Buble	--
6. Add the song ´99 Luftballons´ by Nena to the new playlist	**3.5**
7. Find the song Yesterday by the Beatles and play it back	--
8. Create a new playlist	**3.4**
9. Find the artist of Romeo and Julia on the playlist Cool Hits.	--
10. Choose any song of the genre Rock and play it back.	--

Table 3: Tasks used in the SAMMIE evaluation study

All tasks were given in the Full SAMMIE run. The **bold tasks** were transferred from the baseline study. Not more than the grey pronounced tasks were given in the C&C run to keep the whole session within time limits.

All task but no. 8 could be performed by speech input or by iDrive. Task 8 had to be solved exclusively by speech input.

The experimenter presented the tasks in a consistent way by reading them from paper. Each task was repeated once with a different formulation to avoid predefining a single specific formulation and to assist the recollection. [11] Formulation and presentation of the tasks 1-4, 6 and 8 was identical to the baseline study with partly slight differences in the formulations.

The songs and albums, which had to be played, were actually realised acoustically and played back partly. When the Subject did not stop it, then the respective song or the next song of the album continued to play until the next (or even to the next but one) task.

[10] The grouping of tasks into scenarios as in the baseline study was abandoned, as well as the categorizing into different difficulty levels.

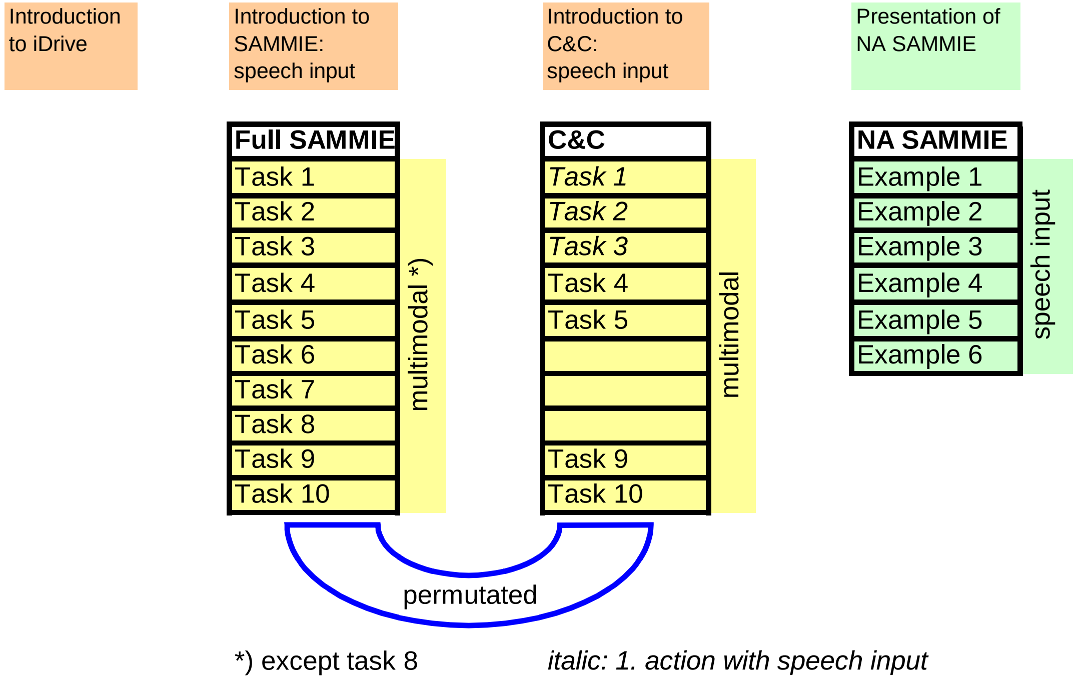
[11] The presentation of the tasks was a critical aspect. The alternative of presenting them visually was excluded for reasons of visual distraction.

2.5 Experimental design

The study was conceived as critical experiment. I.e. hypotheses were defined on the basis of the baseline study and other deliberations (s. chapter 4.5). Moreover, additional results were expected concerning the multimodality and efficiency of the SAMMIE system.

The main variable was the multimodal interaction system. The Full SAMMIE system was the main system. The C&C system was used as a reference system for direct comparison with the SAMMIE system. The Non-Adaptive (NA) SAMMIE system was presented by the experimenter at the end of the session to get a subjective judgment and a comparison to the Full SAMMIE system.

As far as possible, the results should include a system comparison between Full SAMMIE and C&C on the one hand and Full SAMMIE and the results of the baseline evaluation on the other hand. The respective first action of task 1-3 in C&C mode had to be done by speech input to ensure, that the Subjects used speech input at least a few times.



*) except task 8

italic: 1. action with speech input

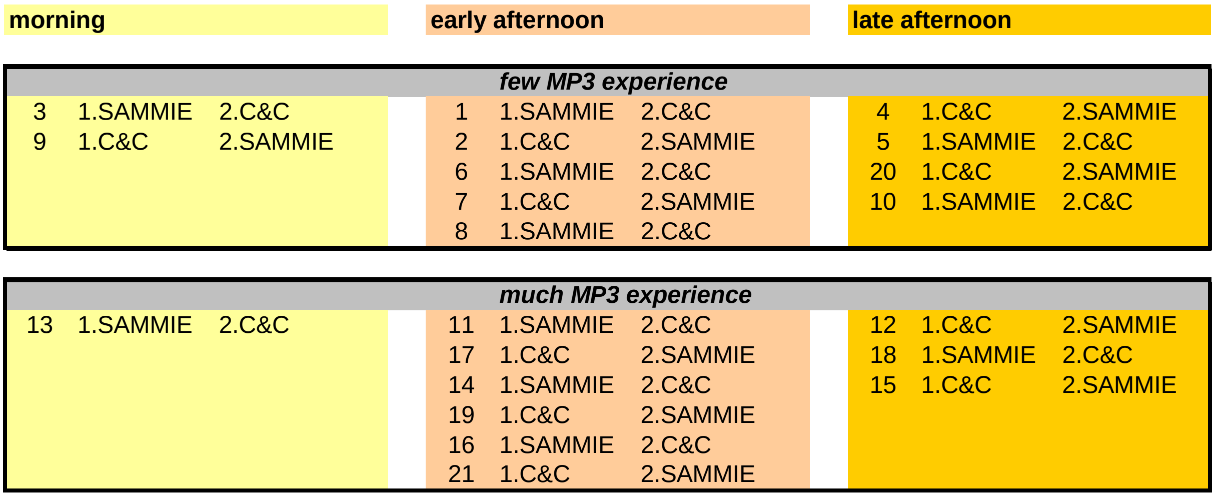
Figure 5: Experimental design in terms of systems and tasks

The Full SAMMIE run ("SAMMIE") and the C&C reference run were balanced across Subjects, to get a fair comparison in respect to traffic situation, order and learning effects. So, about half of the Subjects began with SAMMIE and continued with C&C, while the other Subjects began with C&C and continued with SAMMIE.

For the same reasons a balance between low and much MP3 experience was included. Additionally, a balance between session day times was introduced, since traffic differed considerably over day time. E.g. a similar number of Subjects with few MP3 experience started with SAMMIE as with C&C at the early and the late afternoon. [12]

So, following experimental design was resulting:

[12] To maintain the balance with MP3 experienced Subjects was difficult because of dating problems.



morning			early afternoon			late afternoon		
few MP3 experience								
3	1.SAMMIE	2.C&C	1	1.SAMMIE	2.C&C	4	1.C&C	2.SAMMIE
9	1.C&C	2.SAMMIE	2	1.C&C	2.SAMMIE	5	1.SAMMIE	2.C&C
			6	1.SAMMIE	2.C&C	20	1.C&C	2.SAMMIE
			7	1.C&C	2.SAMMIE	10	1.SAMMIE	2.C&C
			8	1.SAMMIE	2.C&C			
much MP3 experience								
13	1.SAMMIE	2.C&C	11	1.SAMMIE	2.C&C	12	1.C&C	2.SAMMIE
			17	1.C&C	2.SAMMIE	18	1.SAMMIE	2.C&C
			14	1.SAMMIE	2.C&C	15	1.C&C	2.SAMMIE
			19	1.C&C	2.SAMMIE			
			16	1.SAMMIE	2.C&C			
			21	1.C&C	2.SAMMIE			

Figure 6: Experimental design in terms of Subjects (numbers on the left side of each box)

2.6 Experimental realisation

Following Figure 7 illustrates the experimental realisation. The preparation concerned the setting-up of all devices including the experimental vehicle, the SAMMIE system, the video recorders, synchronizing all clocks etc.

The Subject was successively introduced through the explanation of main car functions, several video clips with typical speech and iDrive examples and written instructions, which explained the experiment on the whole and the SAMMIE and C&C system in detail (s. attachment 8.2). [13]

The introduction to the experiment and SAMMIE system comprised:

- Objectives and experimental realisation: multimodality, sequence of activities
- Functions of speech input and microphone: buttons, microphone opening/closing functions, reformulation after misunderstandings, signals, possibility for human communication
- iDrive: movements, functions
- Experimental design: tasks, input modalities, runs
- MP3 display: microphone icon, lists, cursor

The different dialogue and speaking styles for the two systems were explained explicitly.

The training run with pure driving without any additional tasks took about 5 min, where the Subject was getting accustomed to the specific features and driving behaviour of the experimental car (pedals, blinking,, darkened windscreen etc.). The Subject was advised to try some simple manoeuvres like braking.

Before each run equivalent training video clips were shown with typical functions: playing back a specific song, adding a song to a playlist, requiring an information about an album. These functions were shown with Full SAMMIE - iDrive, Full SAMMIE - speech input and C&C system - speech input in a timely manner. As Figure 5 and Figure 7 show, the training video for the iDrive was given once before the first run. The sample of the training video illustrated a few possible formulations and dialogue sequences of the operation including rejections.

After the instructions to the SAMMIE system the Subject trained with the system of the next run with an unstructured sequence and contents of the exercises. Basic functions like searching for albums/songs, playing back songs, including songs to a playlist were included.

The trainings, the two test runs and the completion of the intermediate questionnaires were conducted one after the other with short pauses in between. The experimenter gave the tasks at

[13] The invitation letter already contained an overview of the experiment.

the specific marks on the course. The Subject signalised the finishing of a task, i.e. that he did not intend to continue task processing in any way. [14] If a task was not completed within the given segment, it was broken off at the equivalent mark. The Subject was free, however, to stop earlier, if he would do so in real live. He was asked for his mental load on a 5-point-scale (score between 1 = not stressed at all and 5 = very much stressed).

The supervisor showed the way and supervised the driving with a possibility to intervene verbally. During the test segments he identified the driving errors and signalised them by gestures to the experimenter. After each run the experimenter and the supervisor evaluated the driving performance on standardized scales independently from each other.

After a run the Subject filled in the equivalent intermediate questionnaire (s. attachment 8.3). The final questionnaire (s. attachment 8.4) was handed out after the session to be filled in soon at home. The subjects were paid by 40 Euros for participation.

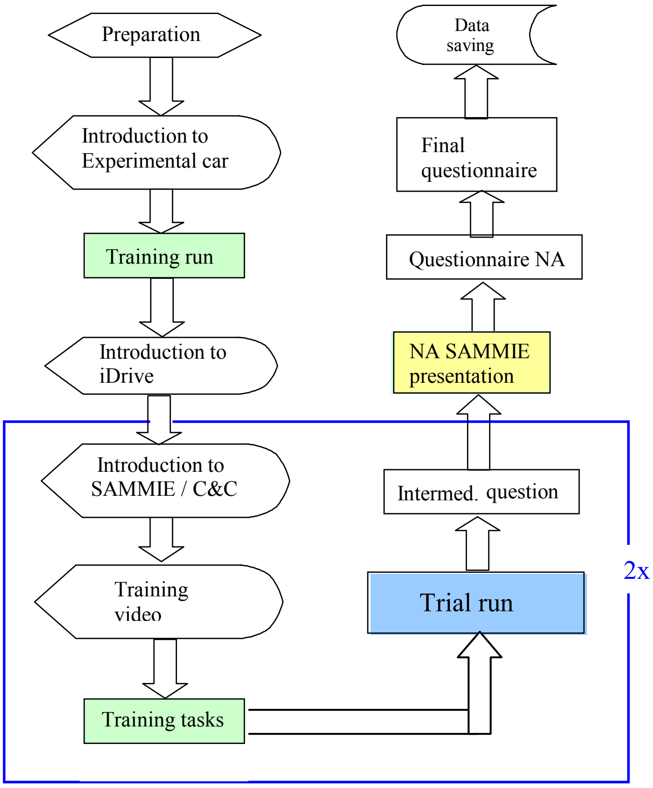


Figure 7: Experimental realisation

The Non-Adaptive (NA) SAMMIE system was presented at the end of the session. To this end six video clips were shown, where the same task was firstly presented in the non-adaptive version then in the adaptive version. The tasks represented functions of the systems where the different features could be illustrated: Personal addressing, differentiation of optical and acoustical presentation, presentation of albums without/with artists, usefulness of confirmation, user guidance, adaptation to user's vocabulary; s. attachment 8.5)

Three pre-tests were carried out at Bosch and three pre-tests in BEF to test the envisaged method for the main evaluation sessions. The tasks and the experimental design were tested in respect to feasibility and duration.

Within the C&C run there were some erroneous settings of the system mode. I.e. with Subjects 2 and 15 there was non-adaptive SAMMIE variant instead of the C&C variant set, and with Subjects 14 (partly) and 21 the Full SAMMIE was set instead of C&C. These data were excluded from the objective and partly from the subjective results.

[14] This sign was necessary to differentiate between the objective and subjective TCR. When it was obvious for the experimenter, that the task was finished, she nevertheless demanded the sign.

3 OBJECTIVE RESULTS

3.1 Preferred modality

One of the main objectives of the evaluation study was to investigate the multimodal interaction between the user and the SAMMIE system. Do users prefer one modality to another or do they change between modalities during their dialogue? Since the Subject usually had the free choice between modalities this question could be answered clearly.[15]

The following Figure 8 shows the overall preferred modality for the SAMMIE, C&C and baseline system. Here, the modality of the respective first action was considered, i.e. the input mode with which the Subjects started to perform the task. The SAMMIE and baseline data include all tasks of the respective study, the C&C data include only those tasks without any constraints as to modality (4, 5, 9, 10). The baseline data comprise the free run without any constraints as to modality as well ("free run", see TALK deliverable D6.3, 17.02.2006, chapter 4.1).

At the beginning of a task, there was a very clear preference for speech input with all systems. At the beginning speech input was used 2,5 - 5 times more frequently than the iDrive. One of the most important reasons for this result was less distraction from driving (visually and manually), as the statements of the Subjects revealed (s. chapter 4.2). Moreover, for many Subjects especially for the technicians and young Subjects speech input seemed to be the more interesting mode, which they could compare with earlier systems (baseline etc.).

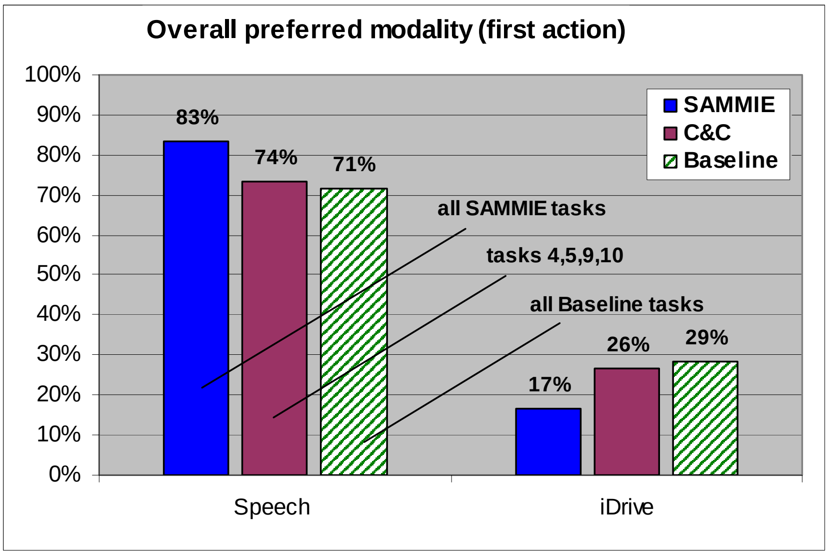


Figure 8: Overall preferred modality of the first action, averaged over tasks and Subjects

The following Figure 9 shows the overall preferred modality for the SAMMIE and C&C system, considering the complete tasks. [16] The SAMMIE data include all tasks, the C&C data include only those tasks, which were given without any constraints as to the modality. The left side

[15] Task 8, which was given only in SAMMIE mode, could be performed exclusively by speech input. The respective first action of task 1-3 in C&C mode had to be done by speech input to ensure, that the Subjects used speech input at least a few times.

[16] Preference was measured on the basis of turns and the most effective modality. E.g. when a Subject operated 4x successfully by speech and 2x successfully by iDrive, the preference was set to "speech". When a Subject started with 3 more or less unsuccessful speech inputs and ended up with 2 successful iDrive inputs, the preference was set to "iDrive".

shows the results of the tasks, which are averaged over the Subjects and the selected tasks. The right side shows the results of the free interactions (SAMMIE: two periods, C&C: one period).

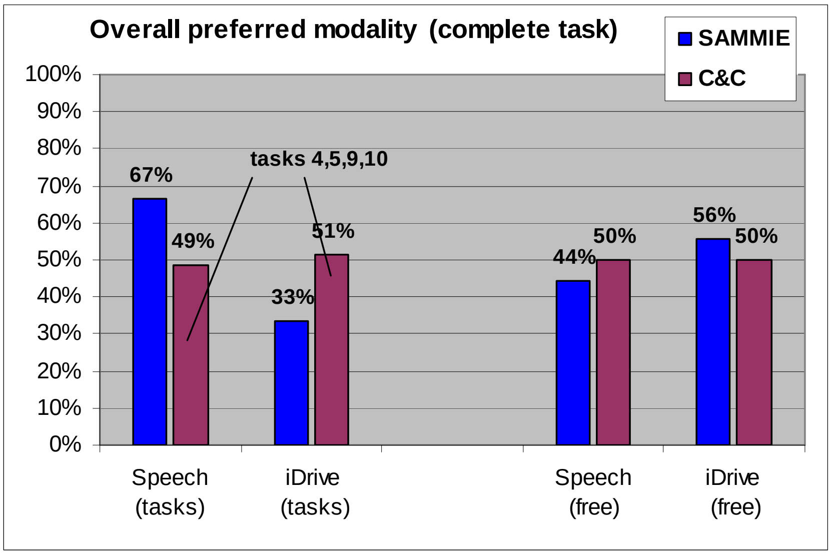


Figure 9: Overall preferred modality in selected tasks, averaged over tasks and Subjects

When compared to the previous figure, there is a pronounced reduction in speech preference within the ongoing interactions during a task. The rejections and false reactions of the systems led to changes to iDrive mode, where the Subjects were sure to get the tasks done. Sometimes, a long cumbersome speech interaction was followed by a short successful iDrive interaction. The obviously fretful Subjects changed to iDrive eventually.

For the tasks in the SAMMIE mode there was still a considerable preference for speech input. Most Subjects preferred in most tasks to interact by speech than manually by iDrive, even with the experience of rejections and false reactions. They took advantage of the possibility to get their MP3 item quickly often within one or a few actions, e.g. one phrase/sentence including all parameters.

For the tasks in C&C mode, however, there was a balance between the preferred modalities during the ongoing interactions. Speech and iDrive were preferred similarly often by Subjects. The C&C mode required the user to follow the menu in the same manner as with the iDrive mode. So, it was no basic difference in the effort between modalities except the additional drawback of rejections and false system reactions with speech input.

Even more interesting is the result, that iDrive was preferred somewhat more frequently in the SAMMIE mode during the periods of free interaction, though it was clearly less preferred during the interactions to fulfil a given task. One possible explanation could be, that the first bad experiences with system reactions onto speech input induced partly a shift to iDrive during free interaction periods. In addition users probably were able to explore the system more easily and systematically by browsing the hierarchical menu structure using the well-known haptic-visual modality.

A Wilcoxon Matched Pair test revealed, that the difference of preferred modality between systems for the given tasks is significant (Wilcoxon Matched Pairs: T=0, T´=28, p<0,05, tasks 1-5, 9-10 included) [17]. I.e. speech input was preferred statistically more frequently with the SAMMIE system than with the C&C system. Vice versa, iDrive input was preferred statistically less frequently with the SAMMIE system than with the C&C system.

[17] The nonparametric Wilcoxon Matched Pairs Test is comparing two variables. It assumes that the variables were measured on a scale that allows the rank ordering of observations based on each variable (i.e. ordinal scale) and that allows rank ordering of the differences. This test can be almost as powerful as the t-test.

The next Figure 10 shows the frequency of tasks and free interaction periods, which were processed consequently in one modality throughout the equivalent interactions. Speech input was exclusively used relatively often with the SAMMIE system, much more frequently than iDrive. Within the free interaction periods, however, speech and iDrive were used similarly frequently, particularly with SAMMIE (s. explanation above).

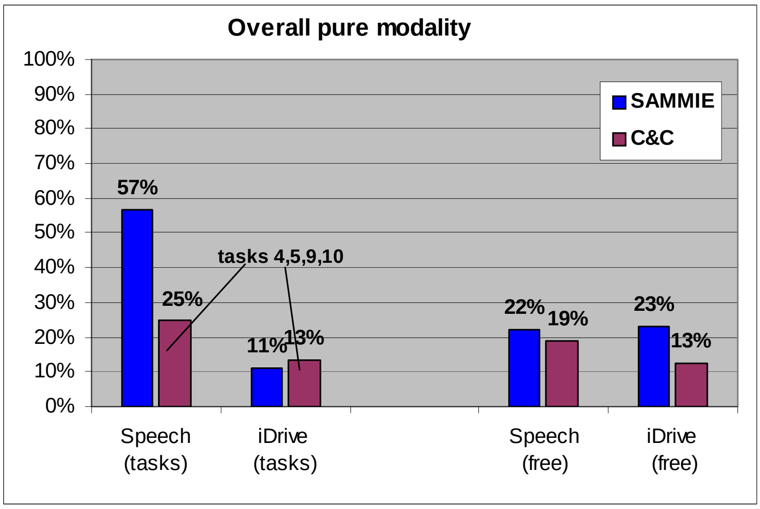


Figure 10: Overall pure modality, averaged over tasks and Subjects

The next Figure 11 represents the frequencies of modality changes during the task processing (all changes during task processing were counted). Subjects changed from speech to iDrive more than twice as frequently than vice versa from iDrive to speech. By far the most frequent reason for a change from speech to manual input were repeated rejections or false system reactions onto speech input.

One of the reasons for a change from manual operation to speech input was e.g. in task 7: Scrolling to the song '99 Luftballons' manually and then copying to the playlist verbally. [18] Another reason was the unsuccessful manual search for a song/album/playlist in tasks 6, 7, 9 or 10 which led to a change to speech input.

Another result is, that much more changes (both directions) occurred with the C&C system. Here, a change between modalities was easier, because both modalities were menu-based. (In the baseline study there was a somewhat different mode of calculation.)

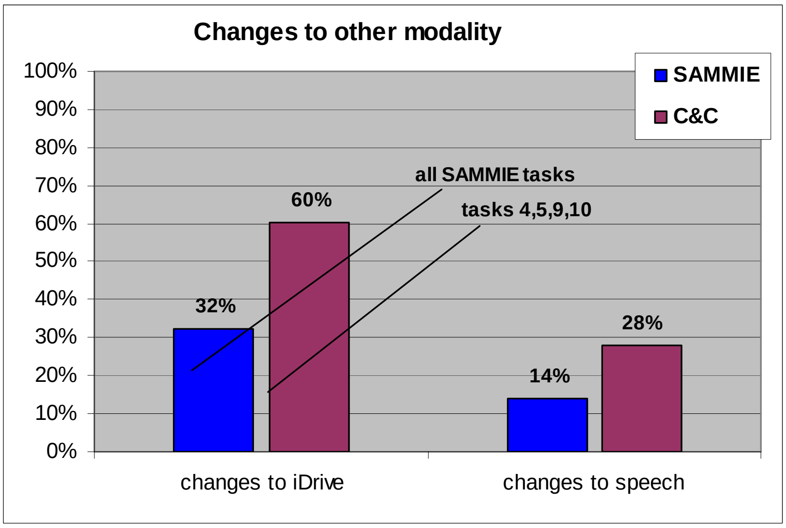
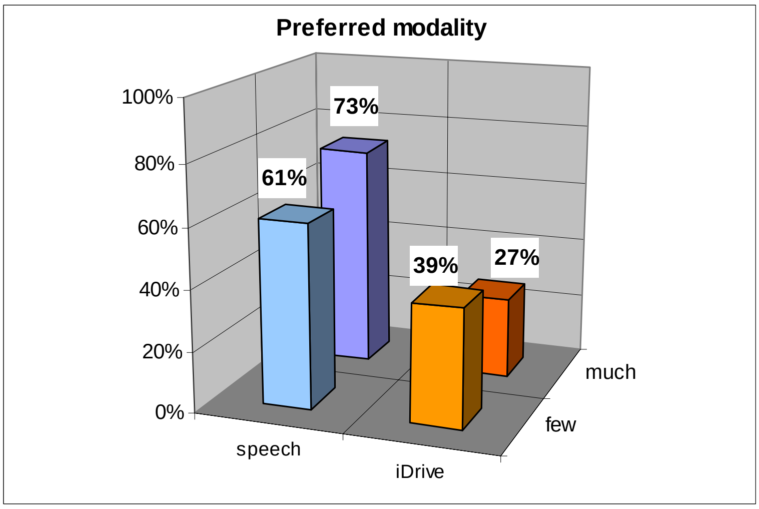


Figure 11: Modality changes during tasks, averaged across tasks and Subjects

The next Figure 12 demonstrates the preferred modality for the levels of MP3 experience, averaged over systems and Subjects of the MP3 level groups. Experienced Subjects preferred clearly more speech than the less experienced Subjects and vice versa for iDrive. The result of

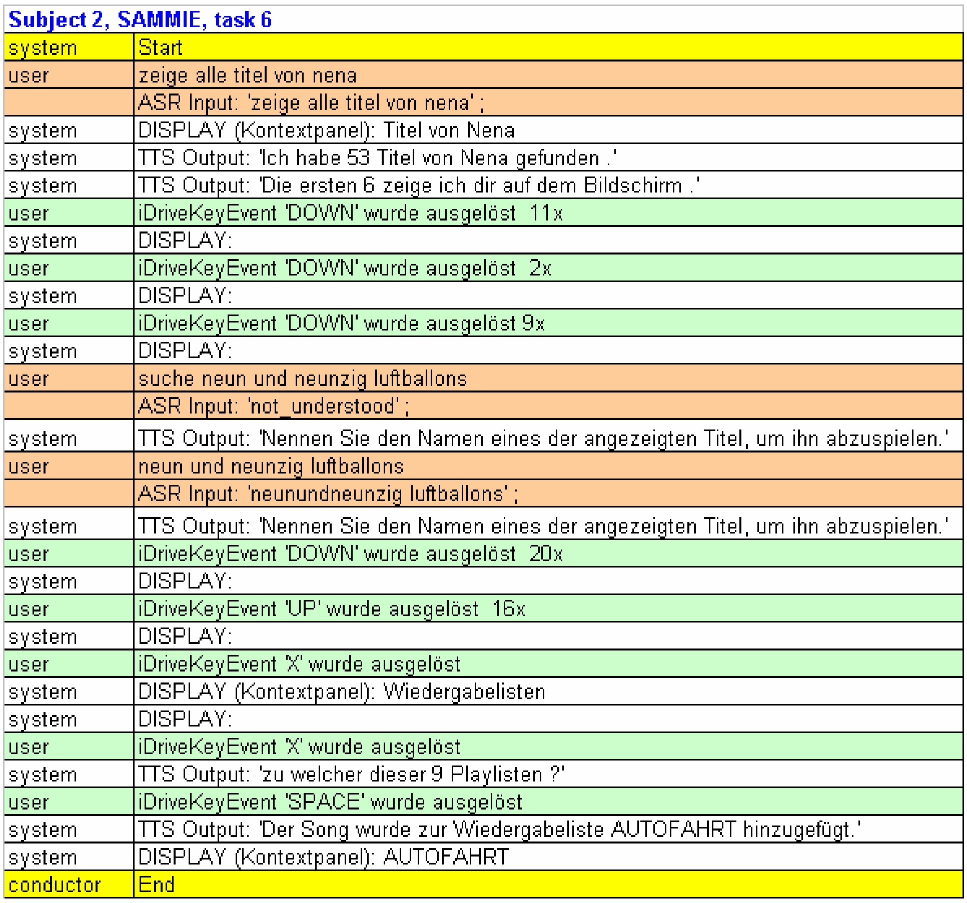
[18] Though having learned the manual (but not obvious) copying into the playlist, s. chapter 2.6.

the statistical χ^2-Test shows no significant difference of the preferred modality between experience levels, neither for speech nor for iDrive use (much / low experienced: χ^2=2,62, f=1, p=0,11). [19]



The following dialogue shows an example for several changes between modalities, though there were no system errors. The Subject changed three times the modality, when she did not succeed with the preceding one. (Subject 2, SAMMIE, task 6: 3 speech turns, 8 iDrive turns, t=1:37 min, 1 false reaction, 1 rejection, 5 driving errors, TCR=1)

Figure 12: Preferred modality for different MP3 experience, averaged over systems and Subjects of the groups



Subject 2, SAMMIE, task 6	
system	Start
user	zeige alle titel von nena
	ASR Input: 'zeige alle titel von nena' ;
system	DISPLAY (Kontextpanel): Titel von Nena
system	TTS Output: 'Ich habe 53 Titel von Nena gefunden .'
system	TTS Output: 'Die ersten 6 zeige ich dir auf dem Bildschirm .'
user	iDriveKeyEvent 'DOWN' wurde ausgelöst 11x
system	DISPLAY:
user	iDriveKeyEvent 'DOWN' wurde ausgelöst 2x
system	DISPLAY:
user	iDriveKeyEvent 'DOWN' wurde ausgelöst 9x
system	DISPLAY:
user	suche neun und neunzig luftballons
	ASR Input: 'not_understood' ;
system	TTS Output: 'Nennen Sie den Namen eines der angezeigten Titel, um ihn abzuspielen.'
user	neun und neunzig luftballons
	ASR Input: 'neunundneunzig luftballons' ;
system	TTS Output: 'Nennen Sie den Namen eines der angezeigten Titel, um ihn abzuspielen.'
user	iDriveKeyEvent 'DOWN' wurde ausgelöst 20x
system	DISPLAY:
user	iDriveKeyEvent 'UP' wurde ausgelöst 16x
system	DISPLAY:
user	iDriveKeyEvent 'X' wurde ausgelöst
system	DISPLAY (Kontextpanel): Wiedergabelisten
system	DISPLAY:
user	iDriveKeyEvent 'X' wurde ausgelöst
system	TTS Output: 'zu welcher dieser 9 Playlisten ?'
user	iDriveKeyEvent 'SPACE' wurde ausgelöst
system	TTS Output: 'Der Song wurde zur Wiedergabeliste AUTOFAHRT hinzugefügt.'
system	DISPLAY (Kontextpanel): AUTOFAHRT
conductor	End

Table 4: Example for a repeated modality change (Subject 2, SAMMIE, task 6)

[19] The Chi-Square Test is a nonparametric test and compares the observed and expected frequencies in each category to test that all categories contain the same proportion of values or not. It assumes ordinal or nominal levels of measurement.

3.2 Task Completion Rate

The Task Completion Rate (TCR) is defined as number of accomplished tasks in relation to the number of the given tasks. The objective TCR represents the correctly accomplished tasks. The subjective TCR represents those tasks, where the Subjects thought to have accomplished the tasks correctly, usually with a wrong parameter or without playing back a song/album, which was already displayed correctly, e.g. in task 4.

After a first failure to accomplish a task, the Subjects were permitted to repeat the task until the end of the course segment was reached. [20] The Subjects themselves had the possibility to stop the processing of the task, whenever they would have done so in real live (which was made rather seldom use of, see below).

The experimental concept included the possibility of experimenter's help (s. below). These "helps" were given particularly when the Subject forgot a parameter. In relatively rare cases the experimenter gave an additional support, when he had the impression, that the Subject did not understand the task. E.g. several Subjects conceived the task 3 with playlist "Pur Klassiker" as a playlist or album of Pur, called "Klassiker". In those cases the experimenter pointed to the obvious misunderstanding. [21]

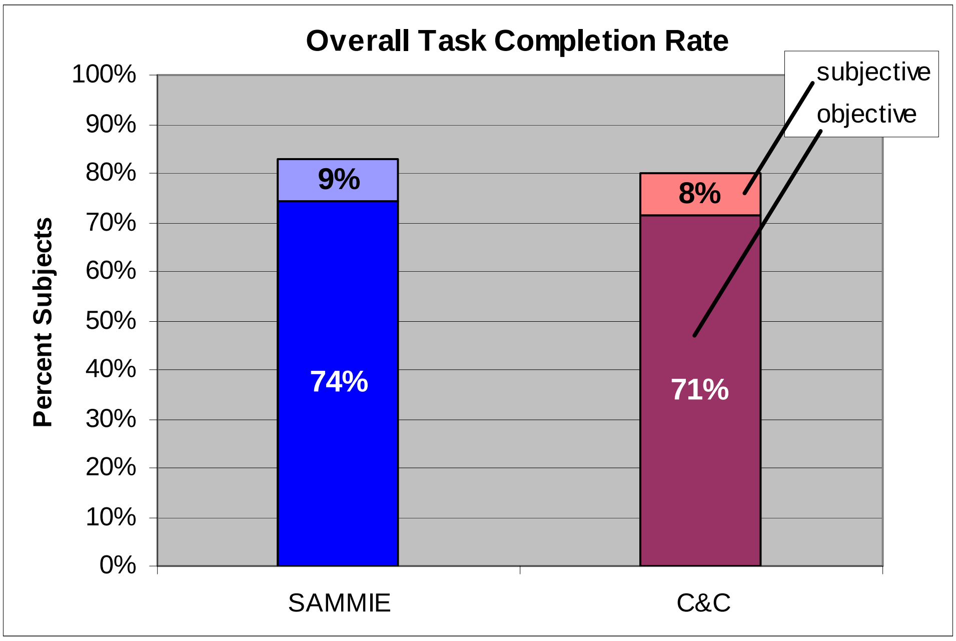


Figure 13: Overall Task Completion Rate for the systems, averaged over tasks 1-5, 9-10 and Subjects

In general it is not possible to draw a fair comparison to the baseline study concerning task completion rate because of several reasons:

- The experimental setup / environment was quite different (lab vs. car)
- The driving task was different (driving simulation vs. real driving conditions)
- The experimental design had to be changed: In the baseline study the subjects were given 5 attempts to accomplish a task without a distinct time limit. In the in-car evaluation tasks had to be completed within pre-defined course segments, which in turn resulted in tighter time constraints with usually less attempts to finish the task.

[20] To remind: In the baseline study a maximum of 5 attempts was permitted.

[21] If the Subject performed the task after an experimenter's help, this was counted as TCR=1, but was separately noted.

- A different selection of tasks was used: The very simple playback tasks (pause song, continue song etc.) were dropped in the present study.

The previous Figure 13 shows the overall objective and subjective TCR (i.e. perceived TCR), averaged over those tasks and Subjects. The bars include all those tasks, which were given in both runs (tasks 1-5, 9-10). The perceived TCR of SAMMIE and C&C is 83% and 79%.

The TCR results of the present study were on a level of about 80%. This has to be interpreted as a general high level, especially when considering the partly tight time conditions (The average time at disposal was about 1:30 min). The tasks with SAMMIE were completed somewhat more frequently than the tasks with C&C. [22]

A Wilcoxon Matched Pair test revealed, that the difference of the perceived TCR between systems for the given tasks is not significant (Wilcoxon Matched Pairs: n=7, T=10, T′=18, p=0,5, tasks 1-5, 9-10 included).

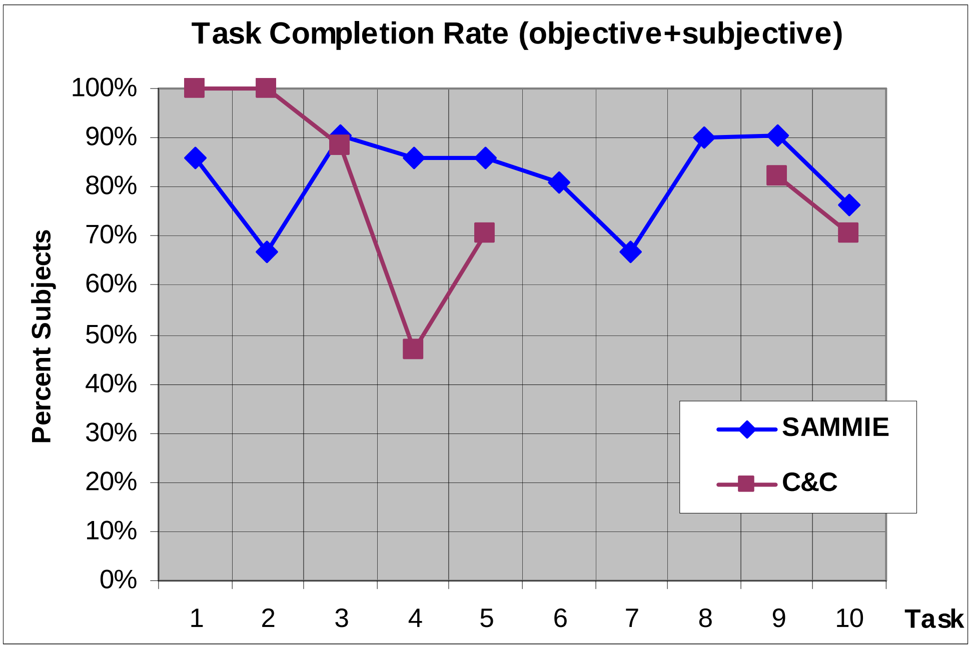


Figure 14: Task Completion Rate for the systems and tasks, averaged over Subjects

The previous Figure 14 shows the TCR for the systems of the present and baseline study for the individual tasks, averaged over Subjects. In tasks 1 and 2 the C&C-TCR was better than the SAMMIE-TCR. A further analysis shows, that most of the Subjects with failures in these tasks started with SAMMIE and were not well experienced with MP3 systems. As a consequence understanding, dialogue and system problems were confounded. (These problems were less in the later tasks for these Subjects.)

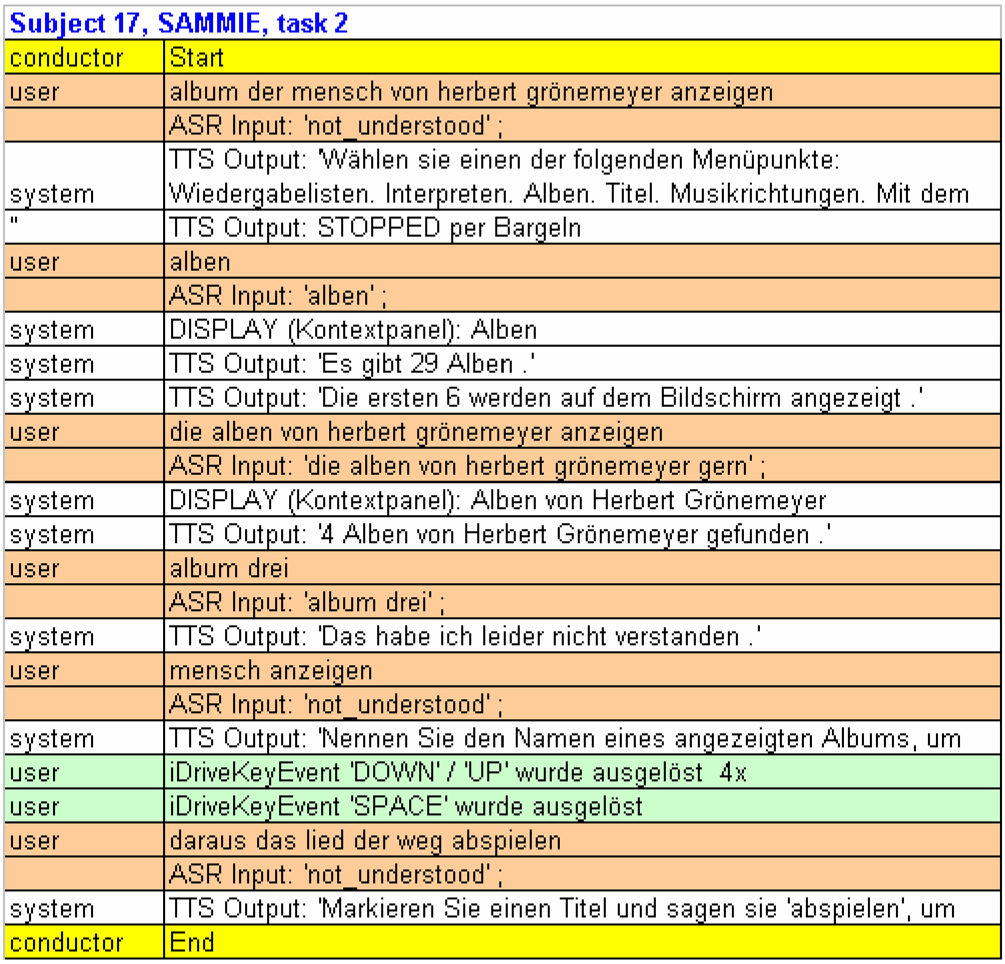
Particularly in tasks 4 and 5 the SAMMIE-TCR was clearly superior to the C&C-TCR. Those tasks belonged to the complex tasks with three information items (e.g. task 4: album, artist and play back). With optimal performance of the SAMMIE system not more than one speech input should have been sufficient. In practice two actions were at least necessary, if the system reaction was correct (e.g. task 4: a) “Spiele mir das Album Live von Pur”, b) “von Pur”).

The performance increase from 40% for the baseline to 81% for the SAMMIE system was very striking for task 6, where “99 Luftballons …” was often not recognized in the baseline study.

[22] If all tasks 1-10 with SAMMIE would have been counted, an equal TCR of 82% would result, i.e. objective TCR = 74% and subjective TCR = 8%.

The completion of task 7 suffered from the fact, that the word “Beatles” was often not understood and the verbal specification of the four elements (artist, album, song, play back) within a relative short driving segment led to relatively many rejections.

Following example illustrates a missed task completion for task 2. (Subject 17, SAMMIE, task 2: 6 speech turns, 2 iDrive turns, t=1:35 min, 4 rejections, 2 driving errors, TCR=0).



Subject 17, SAMMIE, task 2

conductor	Start
user	album der mensch von herbert grönemeyer anzeigen
	ASR Input: 'not_understood' ;
system	TTS Output: 'Wählen sie einen der folgenden Menüpunkte: Wiedergabelisten. Interpreten. Alben. Titel. Musikrichtungen. Mit dem
"	TTS Output: STOPPED per Bargeln
user	alben
	ASR Input: 'alben' ;
system	DISPLAY (Kontextpanel): Alben
system	TTS Output: 'Es gibt 29 Alben .'
system	TTS Output: 'Die ersten 6 werden auf dem Bildschirm angezeigt .'
user	die alben von herbert grönemeyer anzeigen
	ASR Input: 'die alben von herbert grönemeyer gern' ;
system	DISPLAY (Kontextpanel): Alben von Herbert Grönemeyer
system	TTS Output: '4 Alben von Herbert Grönemeyer gefunden .'
user	album drei
	ASR Input: 'album drei' ;
system	TTS Output: 'Das habe ich leider nicht verstanden .'
user	mensch anzeigen
	ASR Input: 'not_understood' ;
system	TTS Output: 'Nennen Sie den Namen eines angezeigten Albums, um
user	iDriveKeyEvent 'DOWN' / 'UP' wurde ausgelöst 4x
user	iDriveKeyEvent 'SPACE' wurde ausgelöst
user	daraus das lied der weg abspielen
	ASR Input: 'not_understood' ;
system	TTS Output: 'Markieren Sie einen Titel und sagen sie 'abspielen', um
conductor	End

Table 5: Example for a missed task completion (Subject 17, SAMMIE, task 2)

The following Figure 15 shows the perceived TCR for the systems and the MP3 experience levels, averaged over tasks and Subjects of the subgroups. One result here is the not very distinct difference between Subjects with different MP3 experiences. Even persons with few knowledge concerning MP3 systems and structure could operate the systems to some degree.

While there is no pronounced difference between systems for few MP3 experience, there is an obvious difference for much MP3 experience: Experienced Subjects achieved a somewhat higher TCR with SAMMIE than with C&C. A further analysis shows, that those Subjects relied more on speech input and accomplished tasks more frequently with fewer turns and somewhat faster (s. Chapter 3.3). This group (with much MP3 experience) had a mean age of 32 years, while the other group (few experience) was about 41 years on the average. It can be speculated, that this age difference could have been an additional factor in respect to taking advantage of the less familiar interaction mode of natural speech input.

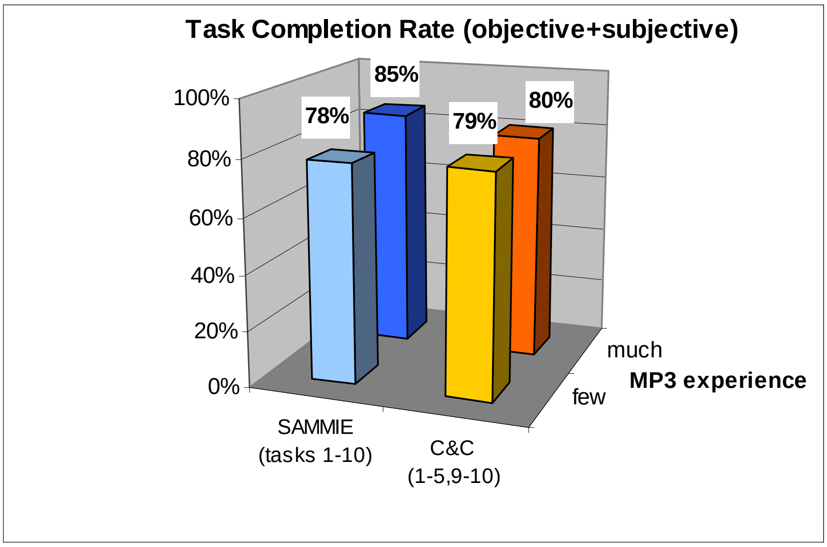


Figure 15: Task Completion Rate for the systems and Subjects´ MP3 experience, averaged over tasks and Subjects of the subgroups

The following Figure 16 shows the abortion of tasks by the Subjects themselves and the experimenter´s helps for the two systems, averaged over tasks and Subjects (“parameter”: Subject forgot a parameter; “else”: additional support, e.g. not understood the task). Since the experimenter´s helps often led to a successful performed task, the TCR data (s. above) have to be interpreted in the context with the help data.

For the SAMMY run there were relatively many helps in tasks 4, 5 and 7. The reasons for that were the task complexity (3-4 elements), the ambiguity (two albums ´Live´) and the strange pronunciation of ´Michael Buble´. For the C&C run there were relatively many helps in tasks 2, 3, 5 and 9. In task 2 the identical word ´Mensch´ as album name and as song name was somewhat irritating. In task 3 the playlist name ´Pur Klassiker´ was misleading for several Subjects (see above).

Without these helps a lower TCR would have been yielded. Particularly the more complex tasks would have been solved less frequently within the given course segment than displayed in Figure 14.

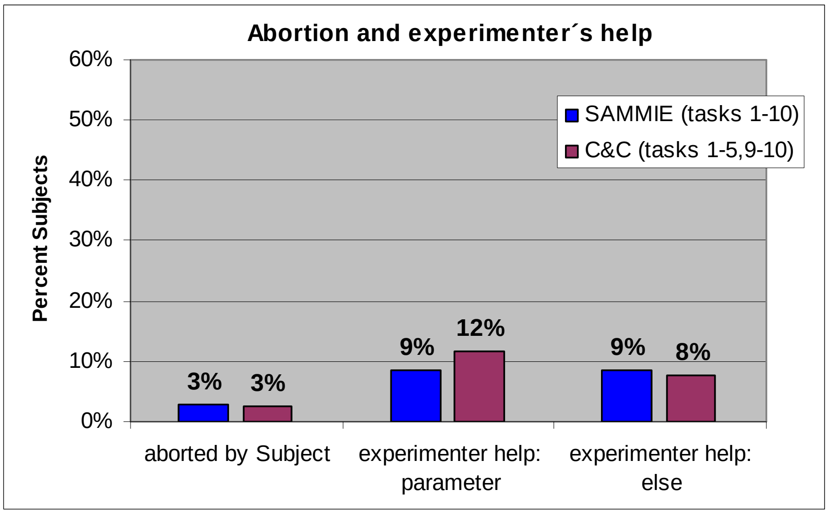


Figure 16: Subjects´ abortion and experimenter´s helps for the systems, averaged over tasks and Subjects

3.3 Number of turns

A "turn" is defined as a pair of a user's input and the corresponding system output. With speech input in the SAMMIE mode a single utterance was theoretically enough to perform a task, if the user included all parameters in one expression and if the system reacted correctly. So, the minimal number of turns with speech input was one, if the operation was done exclusively by speech input. If the dialogue was not optimal (e.g. due to misrecognitions) or the system needed additional information, more than one turn was necessary. E.g. in task 4 an additional choice between two artists was necessary.

With speech input in the C&C mode as much turns as menu presentations were necessary. For most of the tasks the minimal of number of turns was three – four, if the operation was exclusively verbally. For tasks 1 and 3 less turns were sufficient (one and two).

Basically, one action with iDrive was counted as a single input, if one system output followed. E.g. pushing the iDrive controller down, forward or backward together with the corresponding system output was counted as a single turn. For the turnings an action sequence was counted as one turn, when it was followed by one system output in the Log-file. So, a quick turning of the iDrive controller over several raster points and the equivalent system response in the Log-file was ´one turn´. (Thus, the mental user's model of what was one action was more or less modelled). So, with iDrive the minimal number of turns depended very much on the speed of scrolling and was not defined. E.g. the lower limit of number of turns for the rather complex tasks 3 – 7 was usually five to seven.

Here, only the tasks with full subjective or objective accomplishment are considered, i.e. the tasks with TCR=0 are neglected. So, the long taking unsuccessful tasks with a long series of turns are not included in the following figures.

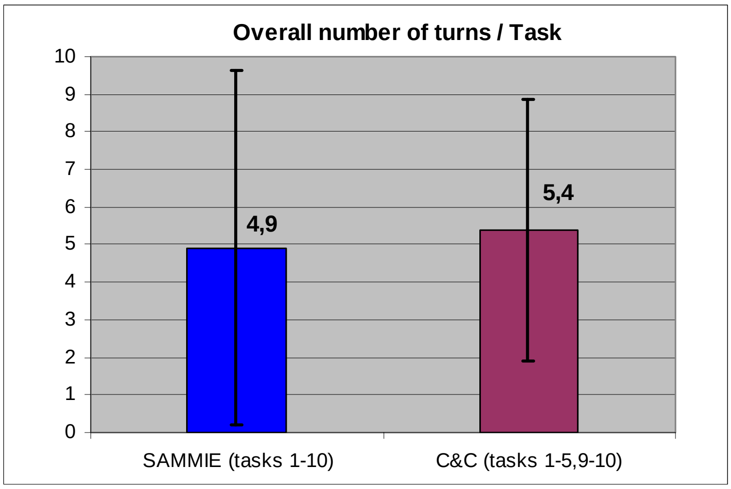


Figure 17: **Overall number of turns, speech and iDrive turns added, averaged over tasks and Subjects. (mean, standard deviation)**

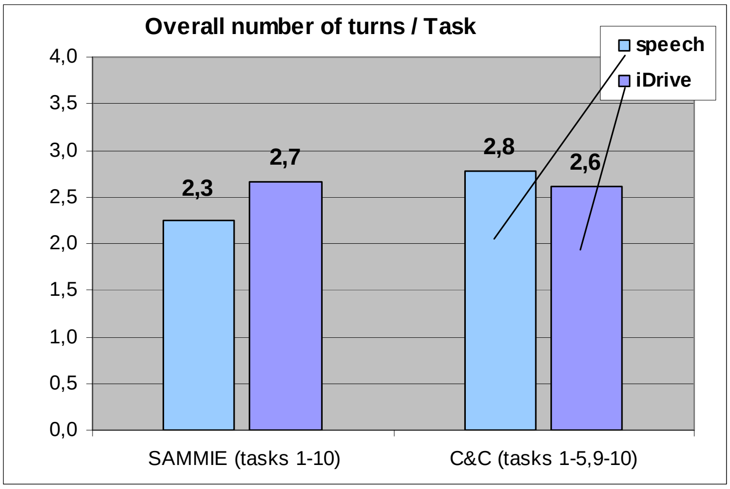


Figure 18: Overall number of turns averaged over tasks and Subjects, speech and iDrive turns separately displayed

The previous Figures show the overall number of turns for the two systems in the present study, averaged over the successfully performed tasks and Subjects. [23] In Figure 17 all turns of each task were counted, i.e. the speech turns and the iDrive turns were totalised (with rounding errors). In Figure 18 the speech and iDrive turns were counted separately.

With the SAMMIE and the C&C system about 5 turns were necessary on the average to complete a task. Considering the complexity of most of the tasks, this seems to be an acceptable level. The difference of number of turns between SAMMIE and C&C in Figure 17 is significant (Mann-Whitney U-Test: $p<0,05$; $n_1=171$, $n_2=95$; $U=6629$) [24]

With the SAMMIE system, however, there were not more than 0,5 turns less than with the C&C system. This seems to be a marginal difference, because with SAMMIE mostly one turn would have been theoretically sufficient. But there were several factors, which affected the number of turns:

- Subjects frequently did not choose the direct and shortest possible dialogue, but partitioned the task in several steps (e.g. firstly calling up the albums or playlists, then specifying them).
- Subjects had to repeat their input after rejections and false reactions by the system, which was more frequently with the SAMMIE system (s. chapter 3.5).

There is a rough balance between the speech and iDrive turns for SAMMIE as well as for C&C.

There is a tremendous standard deviation for both, the number of turns with SAMMIE as well with C&C system, which shows the enormous inter-individual differences. [25]

The next Figure 19 shows the number of turns for the systems and the MP3 experience levels, averaged over tasks and Subjects within subgroups. The speech and iDrive turn data are summed up. As for TCR results (see above) there is no very strong difference between MP3 experiences. Similarly, a Wilcoxon Matched Pair test reveals, that there is no significant difference between experience groups, neither for SAMMIE nor for C&C (e.g. SAMMIE: Wilcoxon Matched Pairs: n=10, T=22, p=0,58).

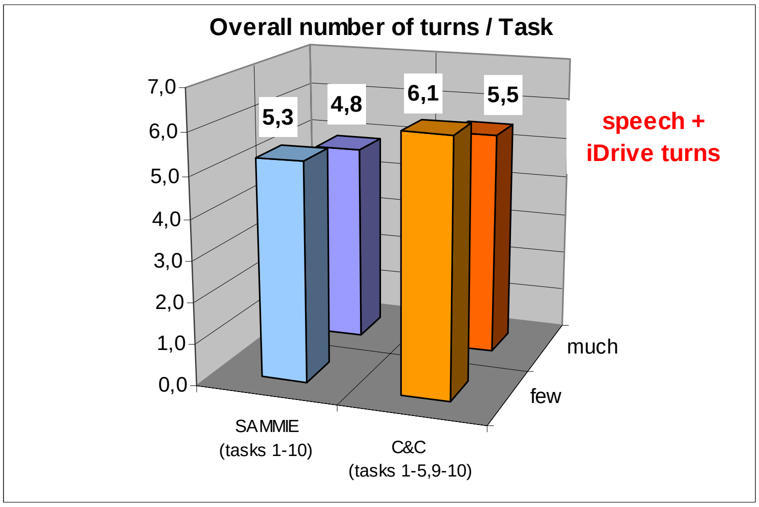


Figure 19: Overall number of turns / Task as a function of system and MP3 experience, speech and iDrive turns added, averaged over tasks and Subjects within subgroups

[23] Since another principle of counting turns was applied in the baseline study, a comparison with the baseline study is not possible.

[24] The Mann-Whitney U-Test for independent samples was used for these comparisons. All single tasks of all 21 (17) Subjects were considered here.

[25] The calculation of the standard deviation requires a normal distribution, which is not given here. But as a rough measure for the data variance it is used here nevertheless. It was calculated by considering directly all tasks from all Subjects.

As the next diagrams in Figure 20 reveal, however, Subjects with few MP3 experience relied much more on iDrive operation, while Subjects with much MP3 experience operated more with speech. This younger group took more advantage of the usually less familiar interaction mode of natural speech input. By this, they achieved a higher TCR with SAMMIE than with C&C. The older group with less MP3 experience relied more on the better known manual operation with a direct connection of input device to the display.

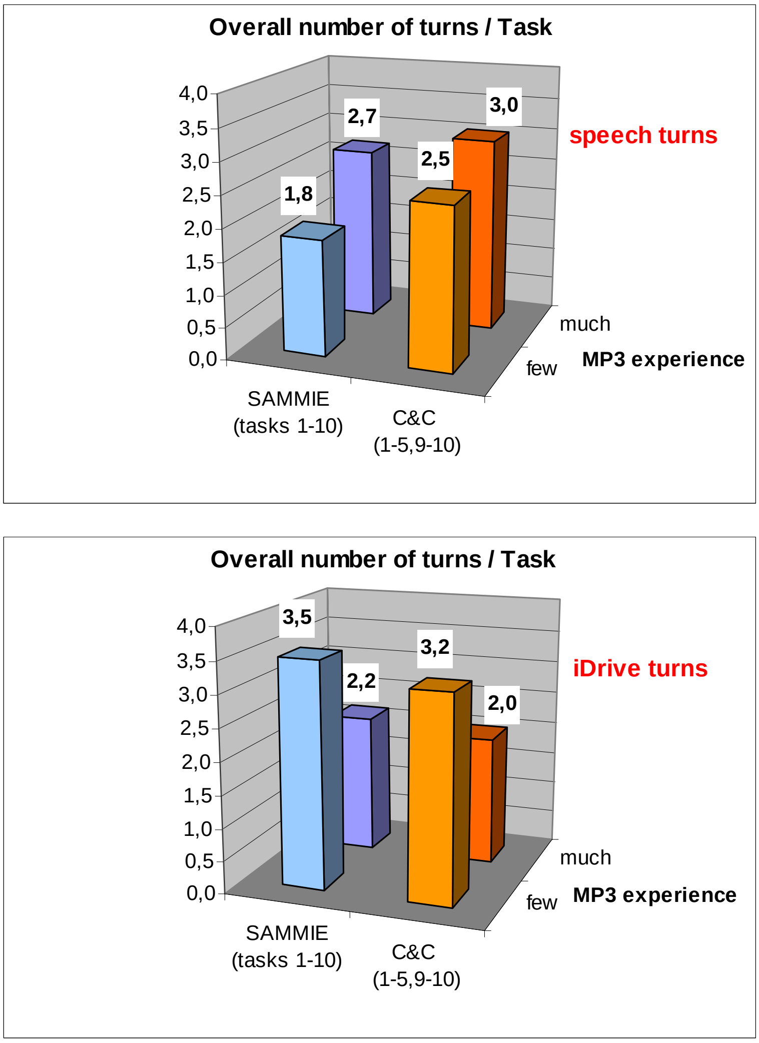
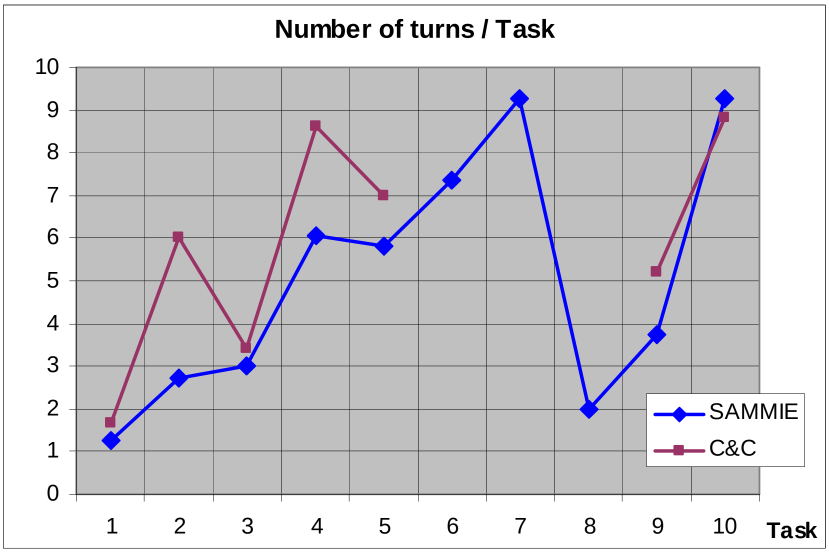


Figure 20: Overall number of turns / Task as a function of system and MP3 experience, speech and iDrive turns separated, averaged over tasks and Subjects within subgroups



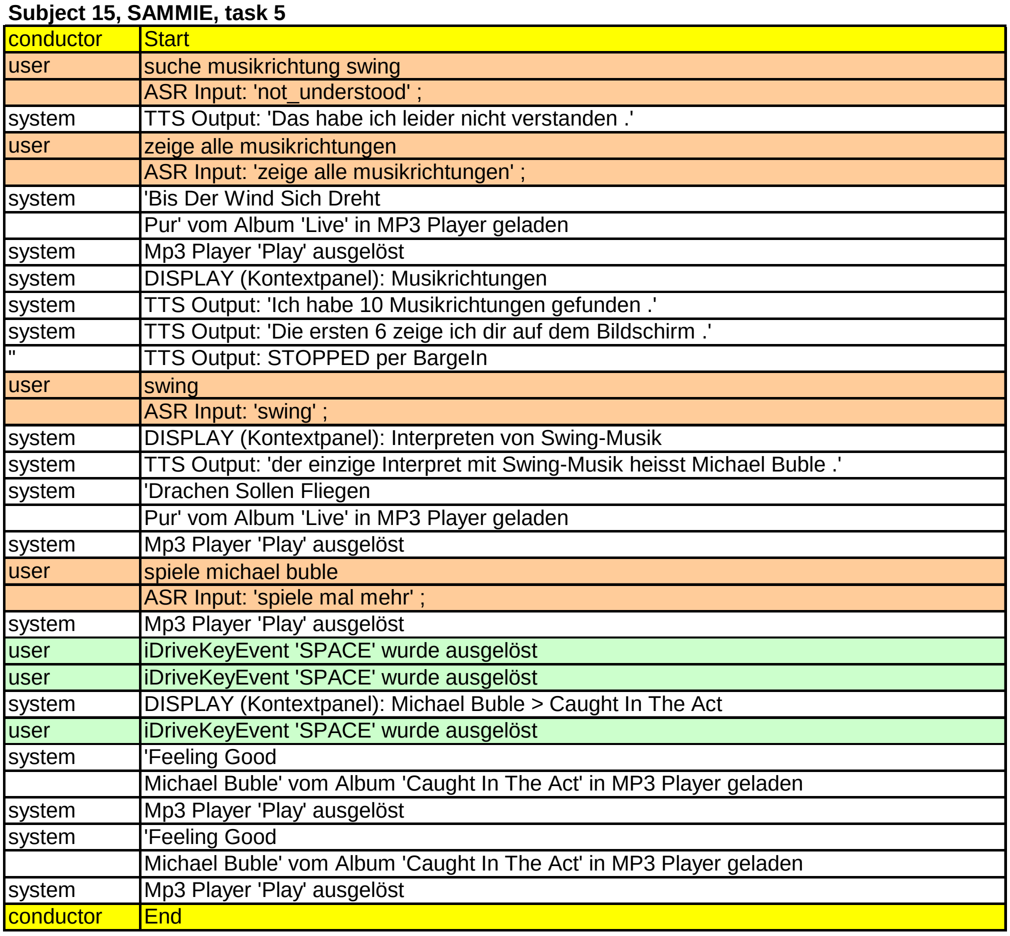
The previous Figure 21 shows the number of turns for the systems and single tasks averaged over Subjects. There was a tremendous difference of number of turns between tasks. Much more turns were necessary to perform tasks with more parameters (tasks 4, 6, 7) or/and where the system performance was lower than else (tasks 6, 7, 10). A specific situation with recollection and pronunciation problems arose in task 5 (‘Michael Buble’).

Figure 21: Number of turns for the systems and tasks, averaged over Subjects

For most tasks there were more turns with C&C than with SAMMIE. This holds true particularly for the rather complex tasks with several parameters.

In all tasks more turns occurred than the minimum, which was very pronounced with the SAMMIE system. In tasks 1 and 8, however, there was just a slight difference between minimal and actual number of turns.

Following example shows a typical number of turns for task 5. (Subject 15, SAMMIE, task 5: 4 speech turns, 3 iDrive turns, t=1:17 min, 2 false reaction, 1 rejection, 3 driving errors, TCR=1).



Subject 15, SAMMIE, task 5

conductor	Start
user	suche musikrichtung swing
	ASR Input: 'not_understood' ;
system	TTS Output: 'Das habe ich leider nicht verstanden .'
user	zeige alle musikrichtungen
	ASR Input: 'zeige alle musikrichtungen' ;
system	'Bis Der Wind Sich Dreht
	Pur' vom Album 'Live' in MP3 Player geladen
system	Mp3 Player 'Play' ausgelöst
system	DISPLAY (Kontextpanel): Musikrichtungen
system	TTS Output: 'Ich habe 10 Musikrichtungen gefunden .'
system	TTS Output: 'Die ersten 6 zeige ich dir auf dem Bildschirm .'
"	TTS Output: STOPPED per BargeIn
user	swing
	ASR Input: 'swing' ;
system	DISPLAY (Kontextpanel): Interpreten von Swing-Musik
system	TTS Output: 'der einzige Interpret mit Swing-Musik heisst Michael Buble .'
system	'Drachen Sollen Fliegen
	Pur' vom Album 'Live' in MP3 Player geladen
system	Mp3 Player 'Play' ausgelöst
user	spiele michael buble
	ASR Input: 'spiele mal mehr' ;
system	Mp3 Player 'Play' ausgelöst
user	iDriveKeyEvent 'SPACE' wurde ausgelöst
user	iDriveKeyEvent 'SPACE' wurde ausgelöst
system	DISPLAY (Kontextpanel): Michael Buble > Caught In The Act
user	iDriveKeyEvent 'SPACE' wurde ausgelöst
system	'Feeling Good
	Michael Buble' vom Album 'Caught In The Act' in MP3 Player geladen
system	Mp3 Player 'Play' ausgelöst
system	'Feeling Good
	Michael Buble' vom Album 'Caught In The Act' in MP3 Player geladen
system	Mp3 Player 'Play' ausgelöst
conductor	End

Table 6: Example for a typical number of turns (Subject 15, SAMMIE, task 5)

3.4 Task duration

The duration of a task was measured between the end of the experimenter's task announcement and the confirmation of the Subject, that he finished the task. [26] Here, only those tasks with full subjective or objective accomplishment are considered, i.e. the tasks with TCR=0 are neglected. So, the longer unsuccessful tasks are not included in the following figures.

Figure 22 shows the overall task duration for the systems in the present and the baseline study, averaged over Subjects and selected tasks. For the SAMMIE run and the baseline study (free run) only the identical tasks were considered, which were performed in both studies (tasks 1-4, 6, 8).

[26] In the baseline study a task was considered as being finished at the last system output. Here, however, the confirmation of the Subject represented the ending of a task (s. chapter 2.6).

The average task with SAMMIE and C&C took about 40 – 50 s. The minimal task durations were about 10 s – 12 s. The parallel tasks in the baseline study, however, took clearly longer. [27]

None of the pairs are significant (Mann-Whitney U-Test: SAMMIE/C&C: n_1=10, n_2=7, U=34, p=0,92; SAMMIE/Baseline: n_1=10, n_2=6, U=19, p=0,23) [28].

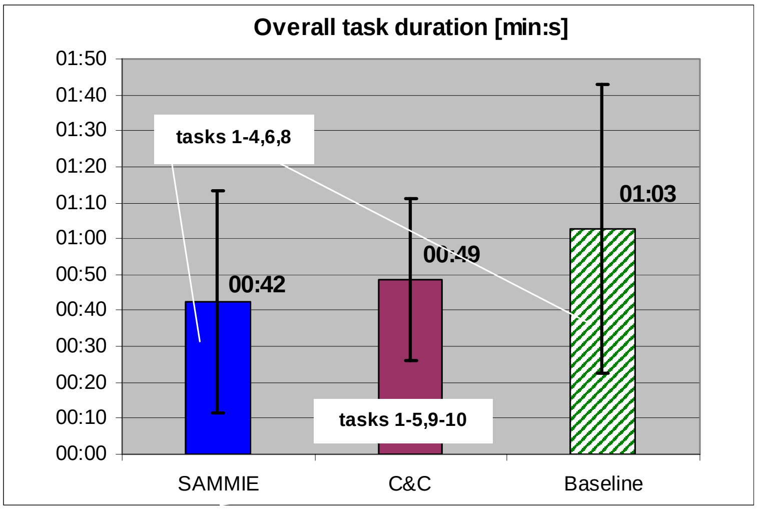


Figure 22: Overall task duration for the systems in the present and the baseline study, averaged over selected tasks and Subjects (means and standard deviations)

The next Figure 23 shows the number of turns for the systems and the MP3 experience levels, averaged over tasks and Subjects within subgroups. As for TCR and number of turns results (see above) there was no very strong difference between MP3 experiences. But MP3 experienced Subjects were somewhat faster with SAMMIE, which reflects the number of turns (see above).

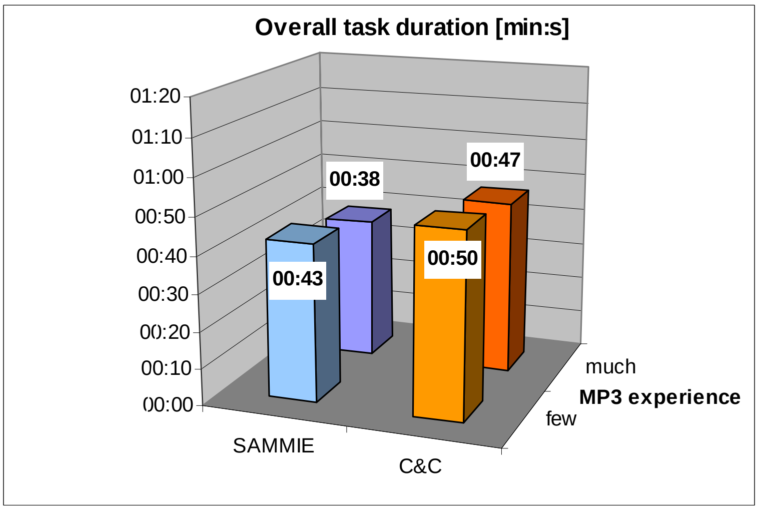


Figure 23: Overall task duration as a function of system and MP3 experience, averaged over tasks and Subjects within subgroups

[27] If the tasks 1-5, 9-10 would be considered an average SAMMIE task duration of 00:43 s would result (to compare with the C&C data of 00:49 s). If all tasks are considered an average SAMMIE task duration of 00:48 s would result.

[28] A comparison of task duration SAMMIE/baseline on the basis of all single tasks of all Subjects with Mann-Whitney U-Test would have been presumably attained significance, but was too costly. A t-test is revealing significance. But task duration is not normally distributed, which prohibits the application of this test.

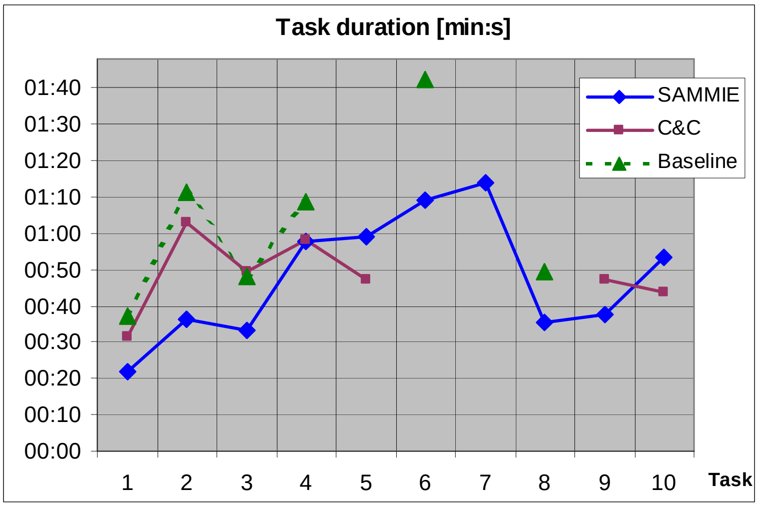


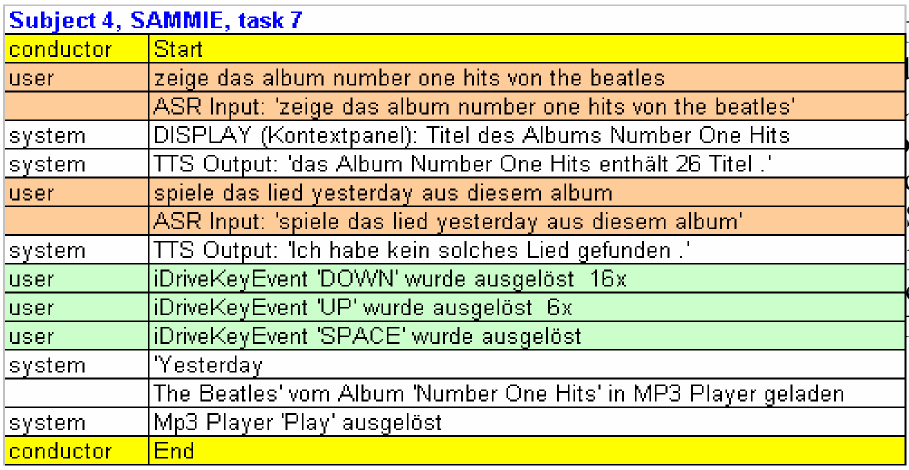
Figure 24: Task duration as a function of system, averaged over Subjects

The previous Figure 24 shows the task duration for the single tasks averaged over Subjects. The number of turns contributed mostly to the task duration. The task duration reflects very well the number of turns (s. Figure 21), if broken down to the data of the single tasks of the individual Subjects (Pearson correlation coefficient r=0,7, highly significant p<0,001).

The task processing is faster with SAMMIE than with C&C and the baseline study in most of the tasks. The inverse result in task 5 is attributed to the recollection and pronunciation problem of ´Michael Buble´, where the menu driven dialogue with C&C relieved the problem. Very striking is the much shorter task duration with SAMMIE in task 6 as compared to the baseline study, which is reflecting the relation in TCR (s. Figure 14).

When any problems with any ambiguity (e.g. two Live albums), recollection and pronunciation (´Michael Buble´) or system performance (´99 Luftballons´, ´Beatles´, ´Rock´) occurred, the task duration generally increased. The similarly complex task 2 (´Der Weg von Herbert Grönemeyer´) led to a good system performance and included more easily to remember parameters, which resulted in a short task duration.

A general impression was, that the task duration was not a too critical factor, if task processing progressed. The given domain of MP3 infotainment seemed to be a ´play-ground´ for several Subjects and the driving task was not dependent on any MP3 results. Particularly in the free interaction periods the Subjects browsed the MP3 system for a considerable period of time.



Subject 4, SAMMIE, task 7

conductor	Start
user	zeige das album number one hits von the beatles
	ASR Input: 'zeige das album number one hits von the beatles'
system	DISPLAY (Kontextpanel): Titel des Albums Number One Hits
system	TTS Output: 'das Album Number One Hits enthält 26 Titel .'
user	spiele das lied yesterday aus diesem album
	ASR Input: 'spiele das lied yesterday aus diesem album'
system	TTS Output: 'Ich habe kein solches Lied gefunden .'
user	iDriveKeyEvent 'DOWN' wurde ausgelöst 16x
user	iDriveKeyEvent 'UP' wurde ausgelöst 6x
user	iDriveKeyEvent 'SPACE' wurde ausgelöst
system	'Yesterday
	The Beatles' vom Album 'Number One Hits' in MP3 Player geladen
system	Mp3 Player 'Play' ausgelöst
conductor	End

ask 7 needed the longest task luration. Following example is typical task duration of 1:16 s f task 7, with a comparatively ow number of rejections Subject 4, SAMMIE, task 7: 2 peech turns, 3 iDrive turns, 1 ejection, 4 driving errors, CR=1).

Table 7: Example for a typical task duration (Subject 4, SAMMIE, task 7)

3.5 System errors

There were different system errors, which can be classified in false system reactions and rejections. ´False reactions´ were generally all incorrect reactions of the system, perceived from the user's point of view. ´Rejections´ were system reactions like "I am afraid I did not understand" or a reference to the help system.

The following Figure 25 and Figure 26 show the number of false reactions and rejections per task for the present and the baseline systems, averaged over tasks and Subjects. The SAMMIE bar includes all those tasks, which were given in the baseline study, too (tasks 1-4, 6, 8). The baseline bar represents the TCR data averaged over the tasks of the free run, i.e. with free modality choice [29]. The C&C includes all tasks, which were given in the C&C run (tasks 1-5, 9-10).

There were as many false reactions with the SAMMIE as with the baseline system for the selected tasks. On the average nearly each second (of these rather complex) task was affected by a false reaction of the system, which irritated the user usually more than a rejection. If considering all tasks, then a mean of even 0,46 false reactions / task resulted for the SAMMIE system.

If considering all those tasks, which were given in the C&C run (1-5, 9-10) then a mean of 0,42 false reactions / task resulted for the SAMMIE system. There were considerably fewer false reactions / task of 0,08 with the C&C system. The difference of false reactions SAMMIE - C&C is significant (Wilcoxon Matched Pairs: n=7, T´=0, T=28, p<0,05).

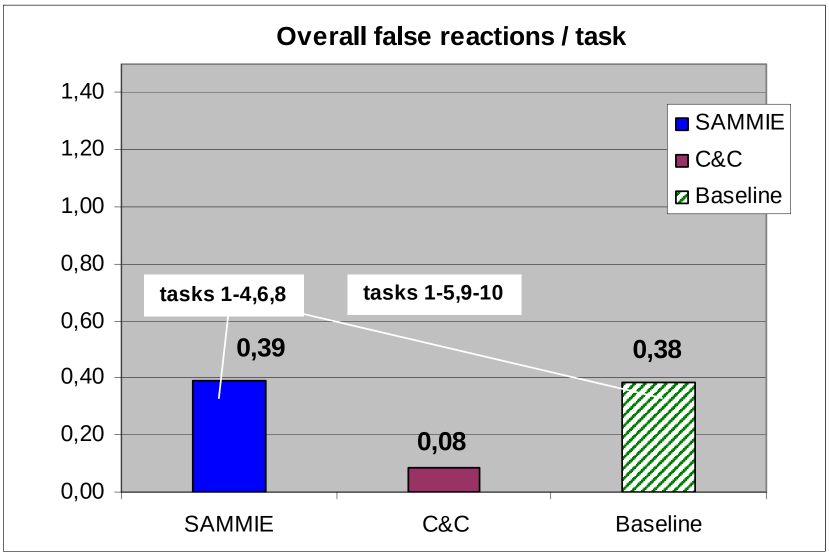


Figure 25: False reactions for the systems, averaged over Subjects and the selected tasks

As the following Figure 26 illustrates, there was about one rejection / task, but fewer rejections / task with the SAMMIE system as compared to the C&C and baseline system. The difference of rejections SAMMIE - C&C is barely missing significance (Mann-Whitney U-Test: SAMMIE/C&C: n_1=209, n_2=119, U=10988, p=0,058, considering all single tasks of all Subjects). If considering all tasks, then a mean of even 0,96 rejections / task resulted for the SAMMIE system.

[29] Corresponds to the green data of tasks 1.4, 1.3, 3.3, 1.5, 3.5 and 3.4 in Figures 20 and 21 of the Final report "Evaluation of the TALK baseline system", BEF, 31.01.2006.

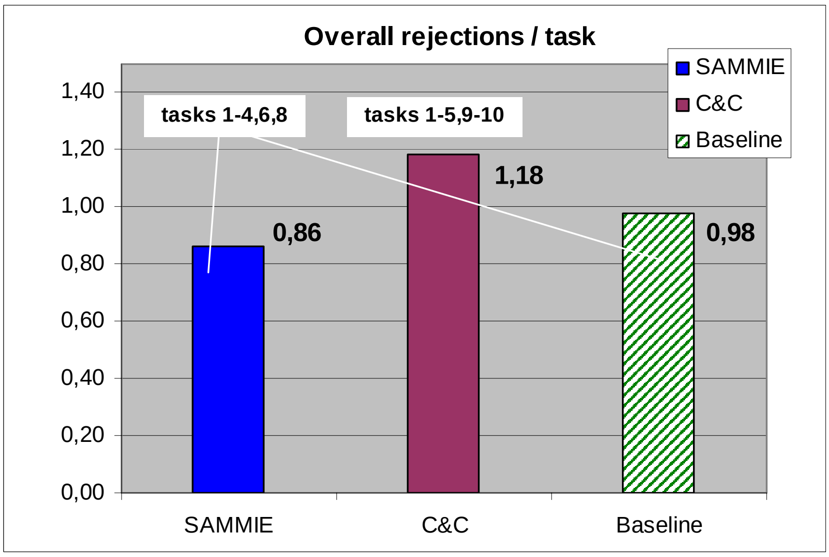


Figure 26: Rejections for the systems, averaged over Subjects and selected tasks

Figure 27 and Figure 28 show the false reactions and rejections per task for the systems and single tasks, averaged over Subjects. There were no false reactions with the C&C system in tasks 1, 9 and 10 (for the relevant Subjects). One explanation is, that in task 1 “Alben” was recognized very well and in task 10 mostly iDrive was used.

The false reactions / task do not well reproduce the number of turns / task (s. Figure 21). I.e. the specific items, formulations and dialogue context seem to be more important for the false reactions than the number of turns. But the rejections reproduce the number of turns relatively well, i.e. more turns resulted in a higher probability of rejections.

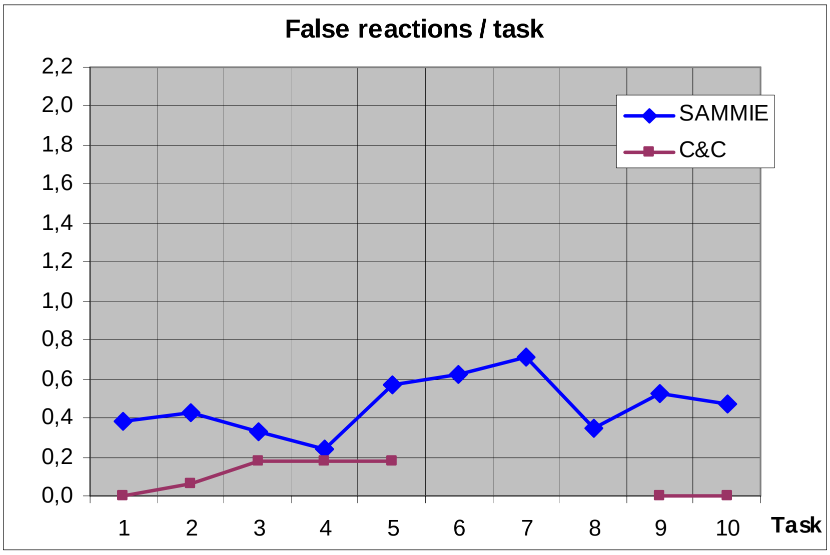


Figure 27: False system reactions per task for the systems and tasks, averaged over Subjects

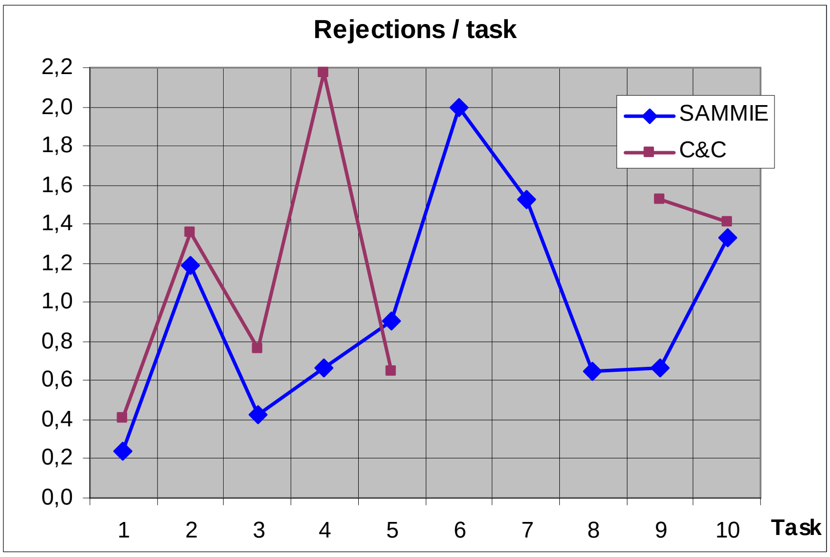
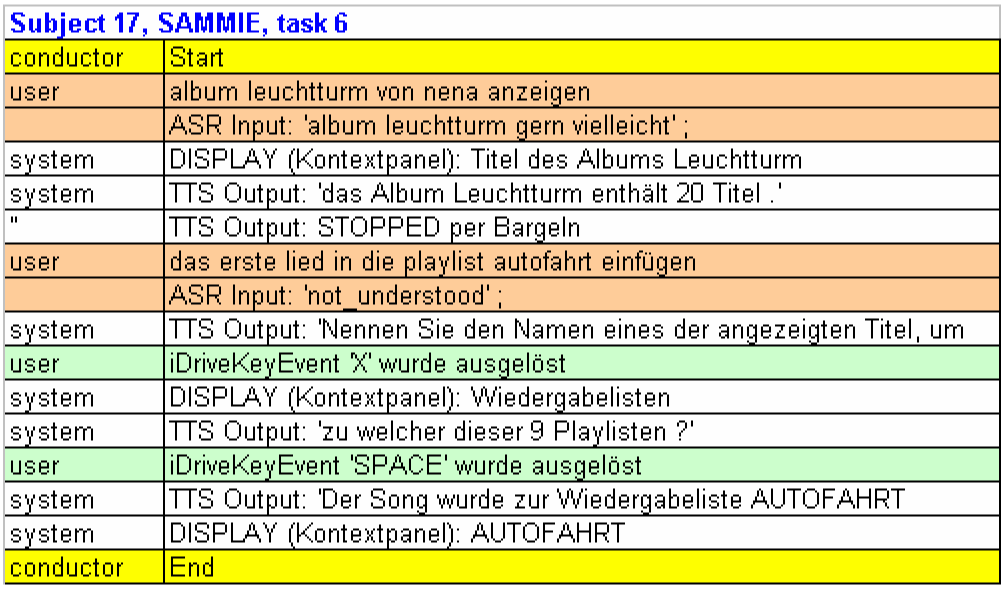


Figure 28: Rejections per task for the systems and tasks, averaged over Subjects

Concerning the rejections, the critical tasks were mainly task 4 (C&C) and 6 (SAMMIE). The number of rejections for each task was relatively balanced across Subjects, i.e. there were no strong outliers.

Following example shows task 6 with one rejection, with the task completion not before the following iDrive turns (Subject 17, SAMMIE, task 6: 2 speech turns, 2 iDrive turns, t=0:32 min, 3 driving errors, TCR=1). The example is below the mean rejection rate.



Subject 17, SAMMIE, task 6	
conductor	Start
user	album leuchtturm von nena anzeigen
	ASR Input: 'album leuchtturm gern vielleicht' ;
system	DISPLAY (Kontextpanel): Titel des Albums Leuchtturm
system	TTS Output: 'das Album Leuchtturm enthält 20 Titel .'
"	TTS Output: STOPPED per Bargeln
user	das erste lied in die playlist autofahrt einfügen
	ASR Input: 'not_understood' ;
system	TTS Output: 'Nennen Sie den Namen eines der angezeigten Titel, um
user	iDriveKeyEvent 'X' wurde ausgelöst
system	DISPLAY (Kontextpanel): Wiedergabelisten
system	TTS Output: 'zu welcher dieser 9 Playlisten ?'
user	iDriveKeyEvent 'SPACE' wurde ausgelöst
system	TTS Output: 'Der Song wurde zur Wiedergabeliste AUTOFAHRT
system	DISPLAY (Kontextpanel): AUTOFAHRT
conductor	End

Table 8: Example for rejections with SAMMIE (Subject 17, task 6)

3.6 Driving quality

The driving quality was measured by recording the driving errors online during the runs and by scoring the overall driving quality. Following driving error categories were considered:[30]

No	Category:	Driving errors:
1	**Dangerous situation**	Intervention of driving instructor, etc.
2	**Speed too low**	Speed ≤ 20 km/h below limit Speed too low with respect to traffic situation
3	**Speed too high**	Speed ≥ 10 km/h above limit Speed too high with respect to traffic situation
4	**Distances too low**	Longitudinal distance too low (“1/4 of tachometer”) Lateral distance too low (0,5 – 1,5 m depending on situation and StVO)
5	**Keeping lane inexactly**	Lane departure False lane used
6	**Insufficient observation**	Bad observation of traffic ahead, behind or beside Blind area disregarded, etc.
7	**Inappropriate braking**	Hard braking Late braking
8	**Other driving errors**	Wrong gear No blinking, etc.

Driving errors were counted only within task processing. To enable a comparison of driving errors between Subjects and tasks, they were normalized to one minute.

A lane departure error was defined as exceeding the middle or edge line of the lane with the edge of the car. Lane departures with a duration of more than about 7-8 s were counted repeatedly. [31]

After each run the experimenter and the supervisor assessed the driving quality on five 5-point scale. These were:

A. very safe / very unsafe [32]
B. defensive / aggressive
C. adapted / not adapted
D. rule conformity / no rule conformity
E. concentrated / not concentrated

The assessments by the two persons were independent from each other and were averaged afterwards. By this, a certain level of objectivity was achieved.[33] This categorization system

[30] This categorization is a result of preceding tests in BEF and is similar to other projects (e.g. INVENT), where it was gradually developed.

[31] This critical time of lane departure was assessed subjectively by the supervisor, depending on the driving situation.

[32] The German word “sicher” could be interpreted in terms of “confident” or in terms of “safe”. Both interpretation were used, depending on the Subject. The sovereign drivers were confident, but not necessarily safe drivers.

[33] There was no specific training for the subjective assessment of driving quality. While categories C. and D. could be reduced to some objective criteria, the other categories had to do with the experience of the evaluators and their personal driving behaviour. By far the most categorizations of the two evaluators were identical or differed not

differentiated e.g. between sovereign but risky drivers (B, C, D low; E high; left = high, right = low) and slow jerky drivers (A, C low; B, D, E high).

The following Figure 29 shows the overall driving errors per minute for the systems, averaged over error categories, tasks and Subjects. There was no pronounced difference of the mean number of driving errors between systems. The use of both systems seem to be coupled to some lack of driving quality. Since no reference run without any tasks was performed, no statement, however, can be made about the effect of multimodal operation on driving safety in general.

The standard deviations were remarkable high, since there was a very large interindividual range of driving errors: With some Subjects there were not more than occasional driving errors, while others crossed the lane edges continuously during task processing.[34]

The difference of driving errors between systems are not significant (Mann-Whitney U-Test: SAMMIE/C&C: n_1=119, n_2=209, U=12382, p=0,95).

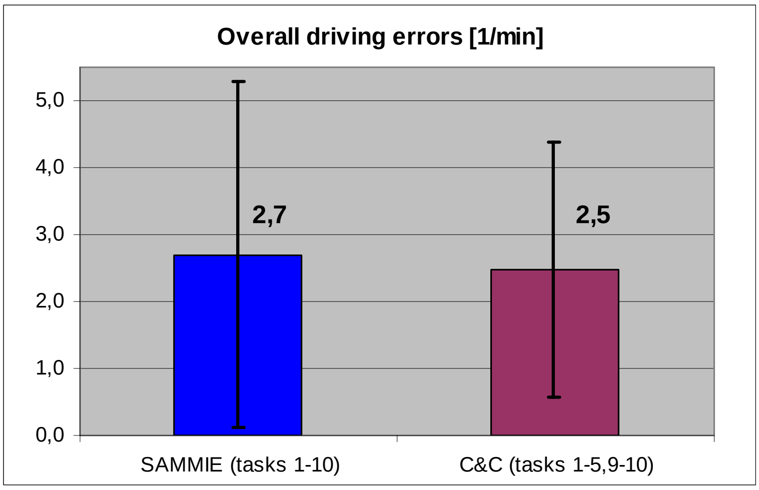


Figure 29: Overall driving errors per minute for the systems, averaged over error categories, tasks and Subjects (means, standard deviations)

The next Figure 30 shows the driving errors for the individual error categories. There is no obvious difference of the individual driving errors between the systems. Any distraction from driving is equivalent for both systems in all measured categories.

Lane departures and low speeds were the most frequent errors. 1,2 lane departure errors per minute seems to be relatively high and can be attributed to the visual distraction when observing the display. I.e. the display was presumably as frequently observed with SAMMIE as with C&C, though a more speech based dialogue would have been possible. (As could be observed during the test there were rather few glances onto iDrive.) [35]

There were clearly more speed too low as speed too high errors. The operation of the systems needed some visual attention, which was compensated by reducing the speed. The experimental car was relatively often overtaken, even on the two-lanes roads.

There were very few dangerous situations. Since a relatively broad definition of "dangerous" was introduced, these were mainly situations, where the supervisor warned (which he did for

more than one point on the scales. Larger differences were discussed after the score specification, so that a certain degree of adaptation to each other can not be excluded.

[34] The lower standard deviation with C&C should not be interpreted because of less tasks and Subjects considered with C&C.

[35] Though not comparable to the present results, a result from the baseline study should be mentioned: There were 2,0 lane departure errors per minute in the free run of the baseline study on the average.

safety reasons very early.) Just one situation occurred, where an accident was certainly prevented by the supervisor's warning.

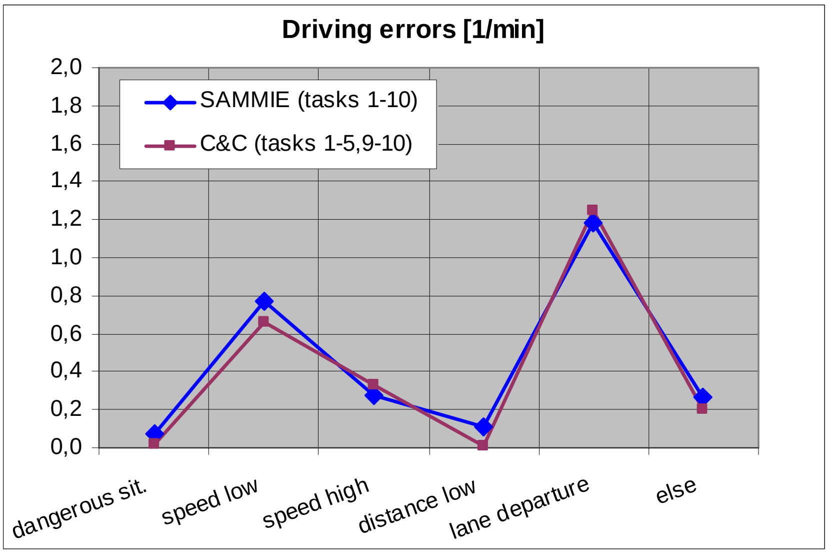


Figure 30: Driving errors per minute for the systems and error categories, averaged over tasks and Subjects (means, standard deviations)

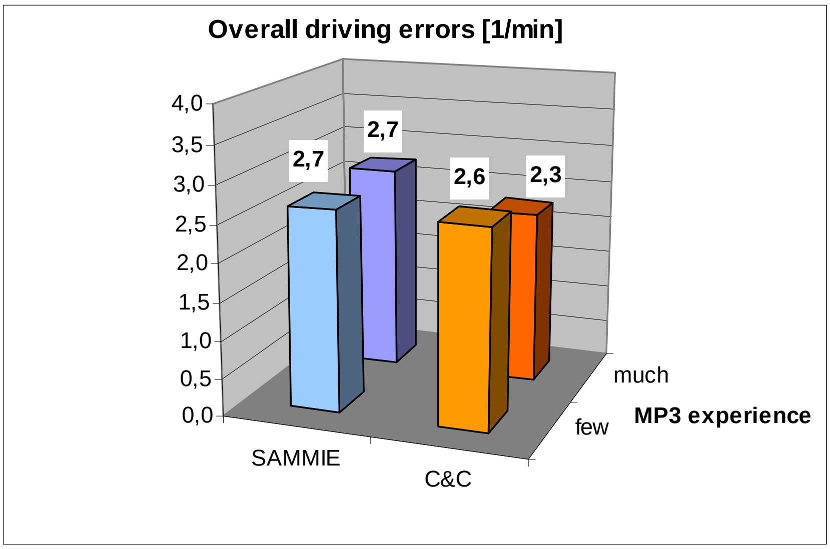


Figure 31: Overall driving errors as a function of system and MP3 experience, averaged over tasks and Subjects within subgroups

The previous Figure 31 shows the overall driving errors for the systems and the different MP3 experience levels, averaged over tasks and Subjects within subgroups. As for TCR, number of turns results and task duration (see above) there was no very strong difference between MP3 experience levels. This can be a hint onto the possible fact, that driving errors depends much more on the individual driving performance than on the operation of the multimodal systems. A better mastering of MP3 systems does not necessarily lead to a better driving.

The following Figure 32 and Figure 33 show the driving quality scores for the systems, evaluated subjectively by the experimenter and supervisor (s. above), averaged over tasks and Subjects, in the first figure additionally over scales. There is no pronounced difference between driving quality scores for the systems. I.e. the subjectively judged driving quality of the Subjects was nearly equal with both systems, which confirms the objective driving quality results.

As could be observed, some Subjects drove very cautiously and relatively slowly during the complete session, more or less independent from system and tasks / no tasks. They wanted to perform well and did not "play" with the MP3 system and the car. Often they relied somewhat more on manual input by iDrive.

Some other Subjects (mostly the younger ones) drove in a superior style, played with the MP3 system and the car and operated often with speech input. Those individual differences effected the driving quality more than the respective interactive system.

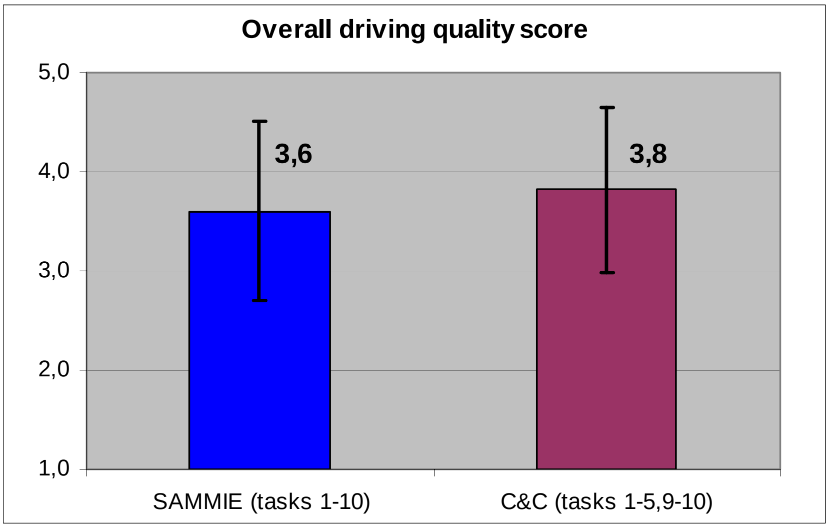


Figure 32: Overall driving quality score for the system, averaged over scales, tasks and Subjects (means, standard deviations)

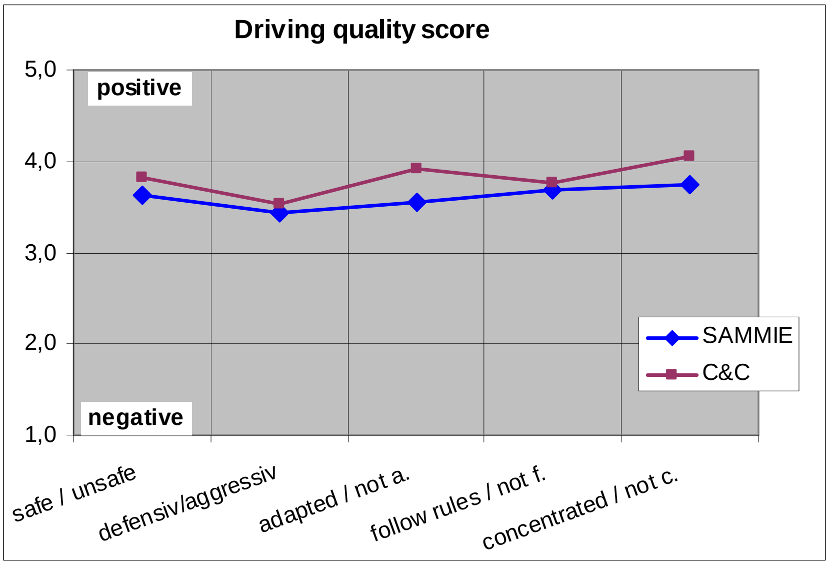


Figure 33: Driving quality score for the system and quality scales, averaged over tasks and Subjects

3.7 Mental load

After each task the Subjects had to specify their mental load ("*Beanspruchung*") on a 5-point-scale ("1" = no mental load, "5" = strong mental load). [36] It represents an overall score for the load given by driving *and* MP3 task [37].

The following Figure 34 shows the overall mental load score for both systems, averaged over tasks and Subjects. In Figure 35 the scores are displayed additionally for the different tasks.

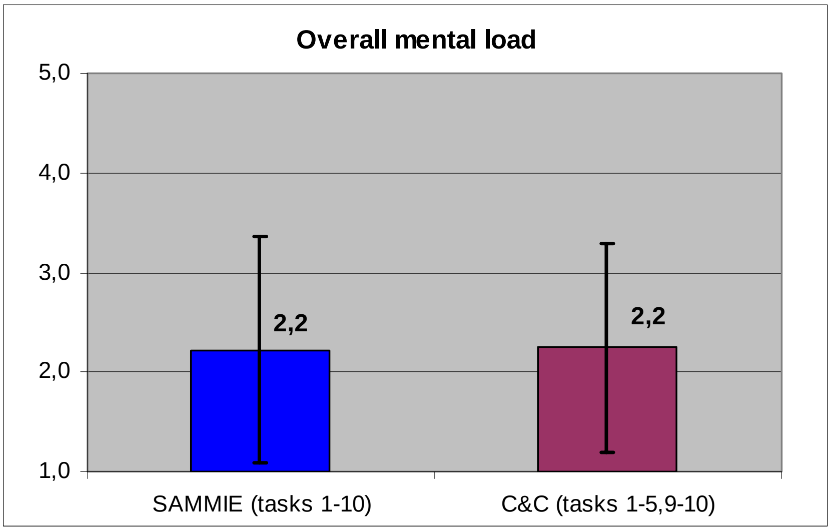


Figure 34: Overall mental load score for the systems, averaged over tasks and Subjects (mean, standard deviation)

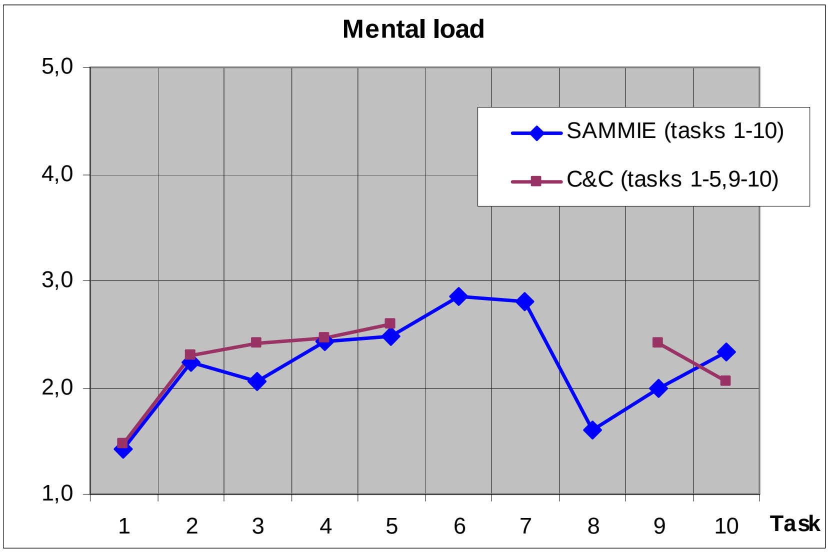


Figure 35: Mental load score for the systems and tasks, averaged over Subjects

[36] During driving "no mental load" is not possible. The lowest level was meant as "Minimal mental load, not more than by driving without additional tasks". This was explained to the Subjects.

[37] Since it was asked immediately after a task, there were no recollection effects and it can be assumed to be a reliable and consistent score.

The mental load was on a generally low level of about two, which can be translated into "strain somewhat above minimum". There was no difference of mental load between systems. Asked about the reasons for scores ≥ 3 the Subjects explained with

- Operating MP3 system within a demanding traffic situation
- dialogue and speech recognition problems
- searching in lists

The mixed demand of driving and operation is presumably an essential factor, not depending on the system. The processing of tasks with a good progress and without serious driving or operation problems were generally not assessed to be demanding. Here, a score of 1 was very often specified.

The iDrive functionality was identical in both modes. With the SAMMIE system the thinking about the formulation or reformulation after rejections was felt to be straining by many Subjects (s. chapter 4.2). Additionally, there were clearly more rejections and false reactions with SAMMIE (s. chapter 3.5). With the C&C system the Subject was more bound to the menu and had to do more turns. These factors seem to be more or less compensatively as to the subjectively felt mental load.

There were low mental load scores in tasks 1 and 8, both of which were usually done fast and with 1 – 2 turns, task 1 either verbally or manually, task 8 exclusively verbally.

The mental load curve over the tasks resembles mostly that of task duration (but task 8), to a somewhat less degree to that of number of turns and rejections. The task duration includes the effort concerning the turns as well as the driving situation and reflecting pauses.

The highest scores were given in tasks 6 and 7, each with four elements (e.g. task 6: artist, album, song, playlist) on a relatively narrow two-lanes road. Both tasks needed most turns (beside task 10), took longest and led to the most rejections and false reactions, accompanied by one of the highest driving error scores.

3.8 Speech Recognition Performance

The evaluation of the baseline system already showed the important influence of speech recognition performance on the evaluation of the system. Improvement compared to the baseline system could however only be achieved by revising the grammar and tuning some recognition parameters as still the same speech recognition engine (Nuance 8.5) was used. The natural language grammar for the SAMMIE system was revised and restructured using the data collected during the evaluation of the baseline system. Additionally a 2nd grammar for the Command&Control (C&C) like system was developed.

When comparing the results we have to keep in mind that the evaluation of the SAMMIE and the C&C system was carried out in the running car with a far-talk microphone while the baseline system was evaluated in the lab with a headset. Thus the different acoustic environment has a prominent influence on the recognition performance.

The figures Figure 36 and Figure 37 below give an overview of the speech recognition performance metrics for the SAMMIE system and for the C&C system. They show the most relevant error rates of the speech recognizer with mean values over all tasks and test subjects and the corresponding standard deviation interval.

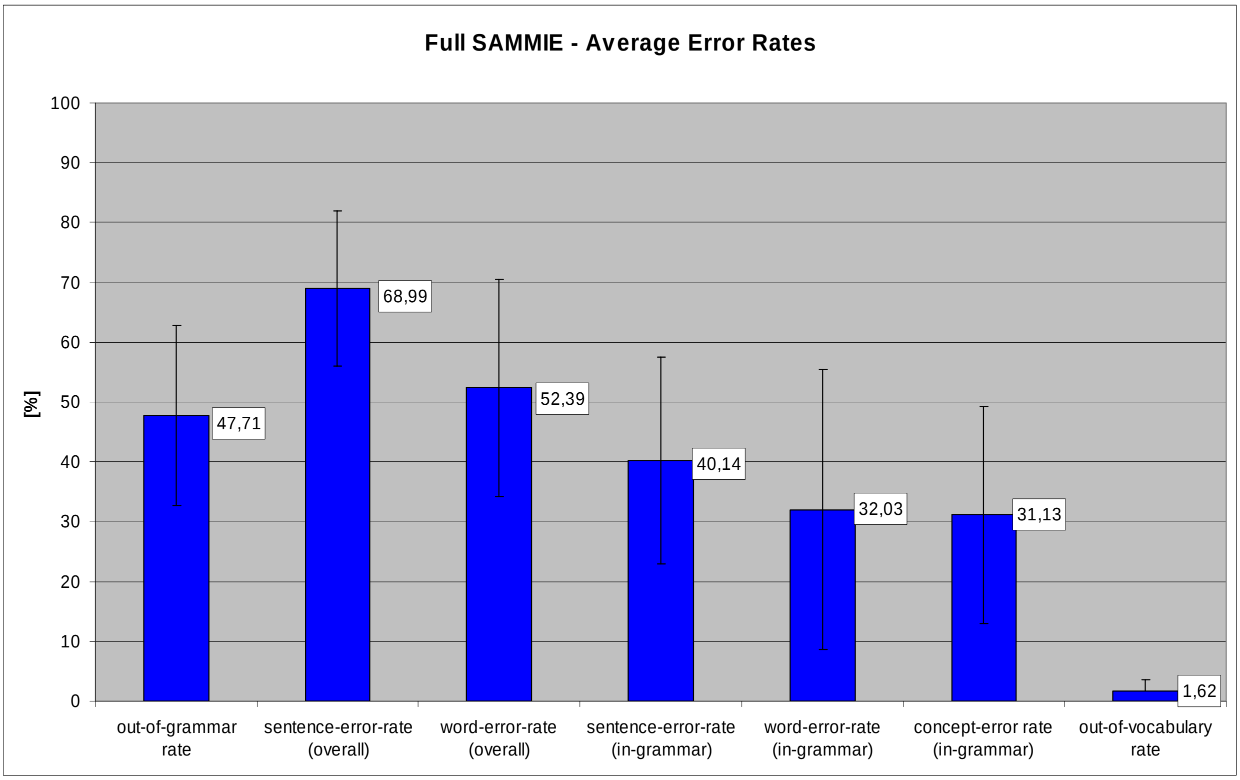


Figure 36: Speech recognition error rates for the Full SAMMIE system, averaged over subjects and tasks (means, standard deviations)

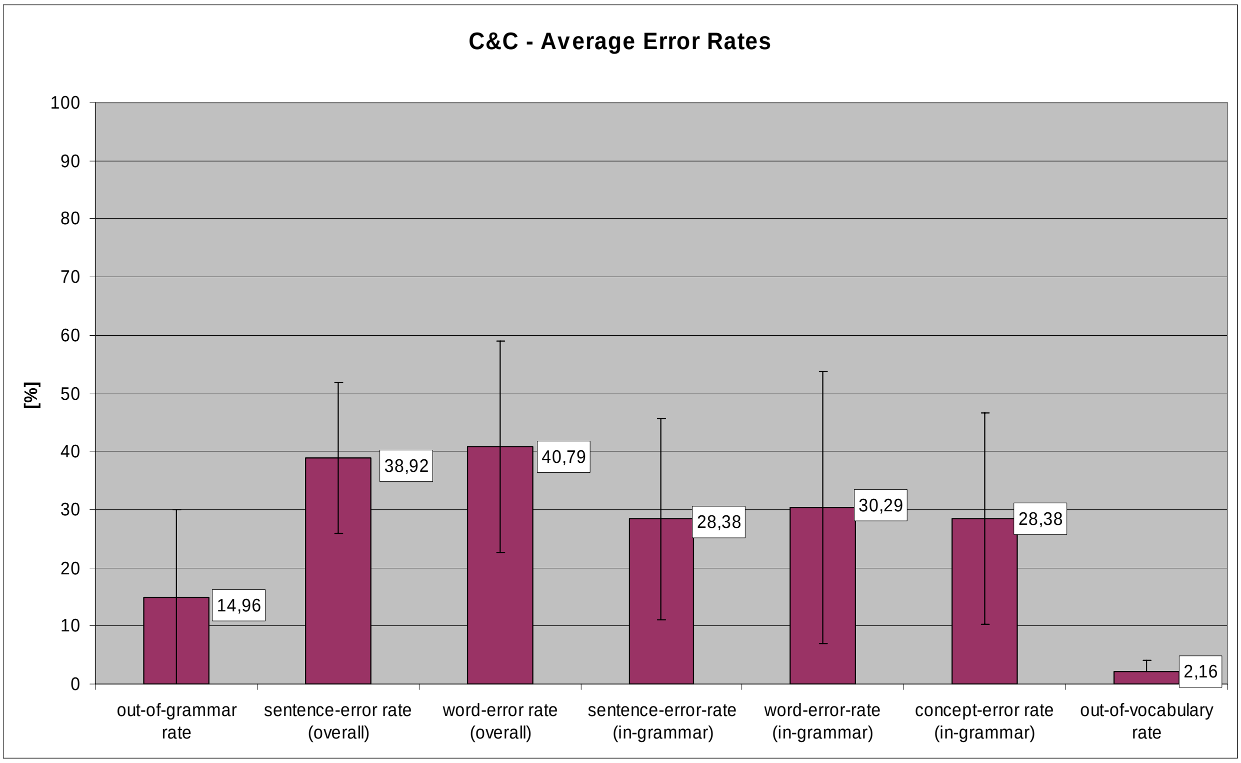


Figure 37: Speech recognition error rates for the C&C system, averaged over subjects and tasks (means, standard deviations)

The figure for the Full SAMMIE system shows high error rates for all test data as well as in-grammar data. Given the reasonably low out-of-vocabulary rate (1,6%) the high out-of-grammar rate (47,7%) is somewhat surprising. Still the grammar seems to contain almost all the necessary words but obviously does not cover sufficiently the variety of phrases used by the subjects, which were encouraged to use natural language.
However for some cases – although not all words were recognized correctly – the semantic concept of the user utterance could still be preserved. This can be seen from the difference between sentence error rate (SER, 40.1%) and concept error rate (CER, 31.1%) for in-grammar data. One may assume that this is also true for a significant number of out-of-grammar utterances[38].

Figure 37 shows a different picture for the Command&Control system: Although error rates for in-grammar data are high as well, the overall sentence and word error rate (WER) are in a more acceptable range due to a quite low out-of-grammar rate. Possible misunderstandings could be reduced by advising the subjects to use displayed items as commands (“what you see is what you can speak”).

Figure 38 depicts a comparison of the average error rates for the Full SAMMIE, the Command&Control and the Baseline system, the latter evaluated in November 2005 (see deliverable D6.3 [2]).

[38] There is no tool support to compute the concept-error-rate for all data, i.e. including out-of-grammar utterances.

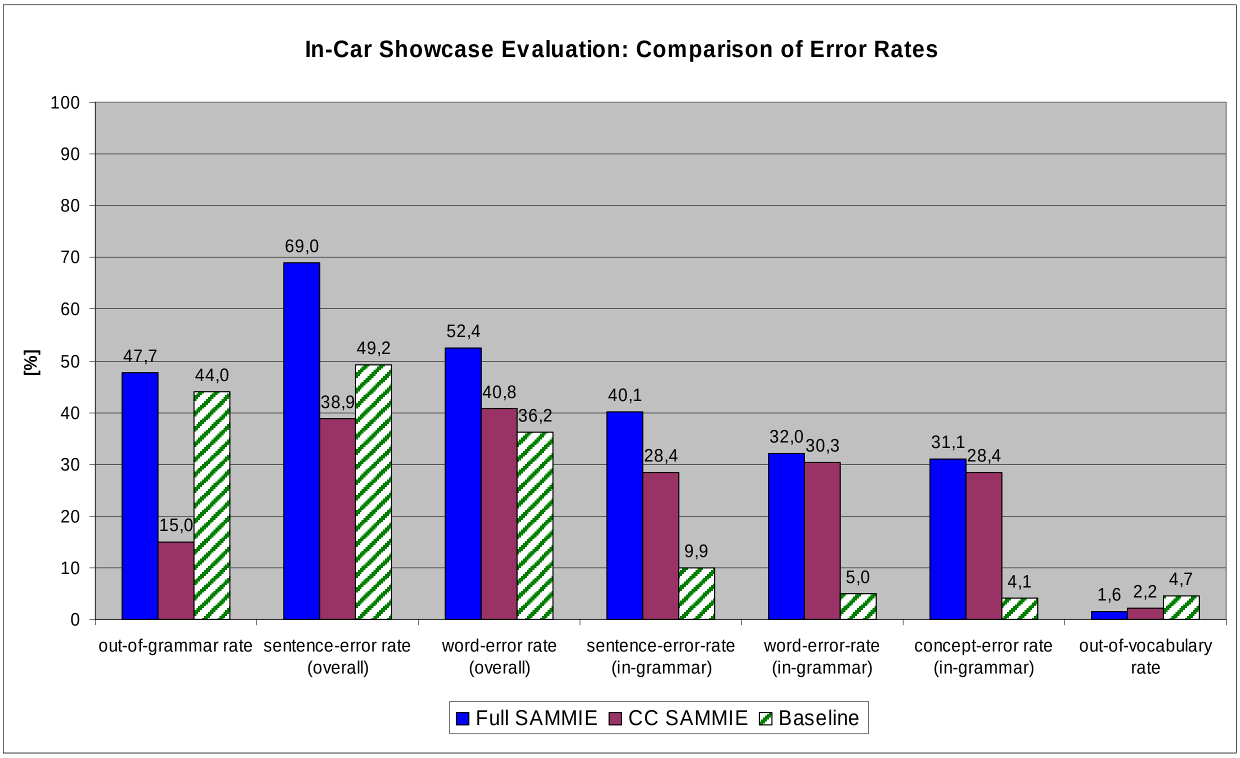


Figure 38: Speech recognition error rates for all evaluated systems

There are two eye-catching differences between the systems:
First of all there is a big difference in WER and SER for Full SAMMIE and C&C when compared to the Baseline system, especially when referring to the in-grammar utterances. The obvious reason for the degrading speech recognition performance is the noisy car environment and the usage of a far-talk microphone compared to the Baseline lab environment with the subjects wearing a headset.
Secondly we see a big difference in out-of-grammar rates between the C&C and the Full SAMMIE system. As already pointed out the subjects were advised to use only specific command words and displayed items respectively while operating the C&C system. The Full SAMMIE however claims to enable natural language input so the subjects could use their own wording with only little indications by the experimenter. However the results show that this freedom is obviously not sufficiently supported by the coverage of the grammar.
On the other hand there is an additional effect which qualifies the high WER and SER for the systems that enable natural language input: The concept error rate is in general significantly lower than the sentence error rate, i.e. the semantic information issued to the dialogue manager often is correct even if some words have not been recognized. Here this can only be proven for the in-grammar data but it can be assumed for the out-of-grammar utterances as well. For the C&C there is no difference between sentence error rate and concept error rate due to the short commands, which are either right or completely wrong.

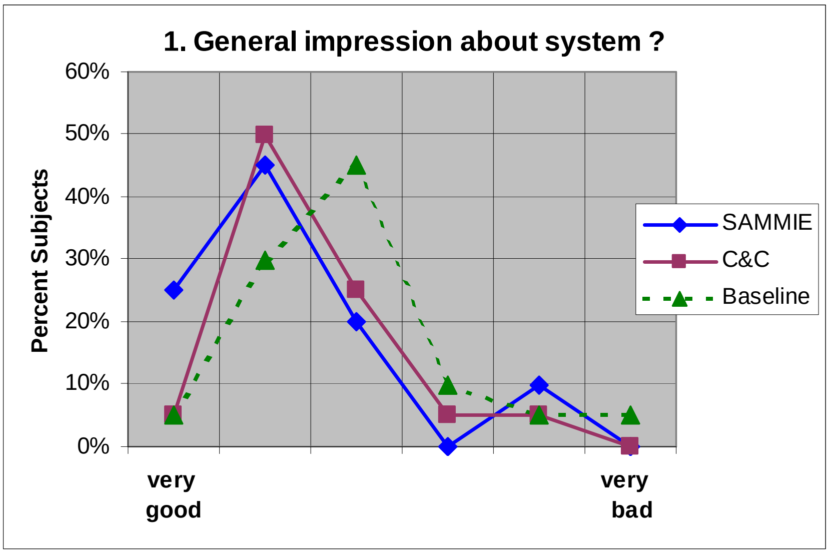
4 SUBJECTIVE RESULTS

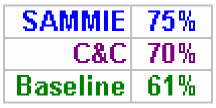
4.1 Intermediate questionnaires

The Subjects filled out intermediate questionnaires after both runs (s. Attachment 3) and a final questionnaire with their own subjective evaluation at home (s. Attachment 4). They were urged to do it at the same or at the following day for reasons of recollection. Most of the questions were identical to the final baseline study questionnaire for reasons of comparison [39] and included mainly 6-point rating scales. [40]

The following Figure 39 shows the frequency of the answers to the first question of each of the intermediate questionnaires concerning the general impression about the interaction systems. [41] With the present systems by far most of the Subjects tended to a positive rating (Ratings on the left side: SAMMIE: 90%, C&C: 80%. Summarized and normalized scores: SAMMIE: 75%, C&C: 70%). With the baseline system, however, there was a maximum nearby the centre of the scale. I.e., there is a clear improvement concerning the subjective overall impression from the baseline to the SAMMIE systems.

SAMMIE and C&C are different mainly at the highest score. I.e. the general impression about SAMMIE was judged to be very good by 25% of the Subjects, only by 5% about C&C.





SAMMIE	75%
C&C	70%
Baseline	61%

[42]

Figure 39: Answers to the question “1. How is your general impression about the entire operating system?”

[39] The questions in the baseline study, however, was exclusively as a final questionnaire.

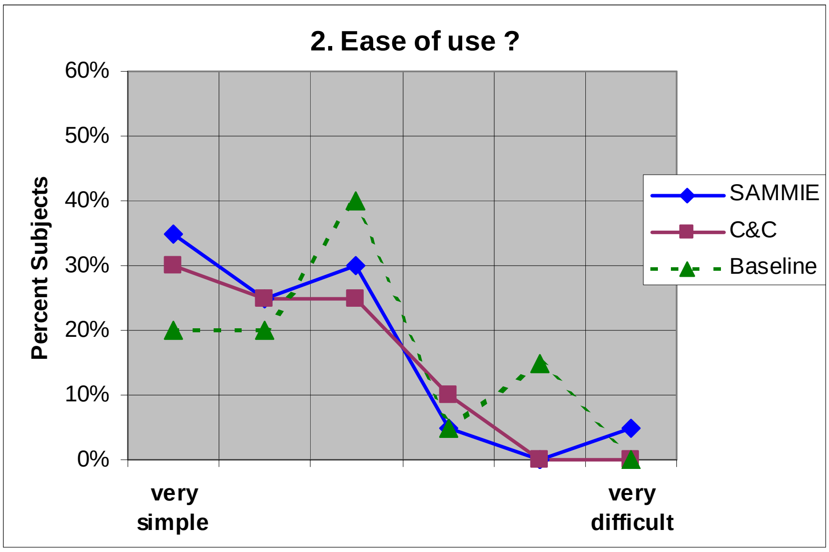
[40] There is a discussion in the literature about scales with even and uneven scales. The present even scale urges the Subjects to give their opinion with some rating tendency (avoiding the tendency to the scale centre on uneven scales).

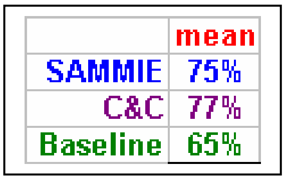
[41] Subject 1 is excluded from the data of the intermediate questionnaire (SAMMIE and C&C), because the questions were allocated differently to intermediate and final questionnaire after her session. Subjects 14 and 21 are excluded from the C&C data of the intermediate questionnaire, because they (partly) operated with the wrong system in the C&C trial. Subjects 2 and 15 were included, though having had the non-adaptive system, because the system outputs of the NA system and the C&C were identical. So, 20 Subjects were considered with SAMMIE and 18 Subjects with C&C.

[42] The overall data at the right side of the following figures represent the summarized results which are normalized to 0% -100%. This was done by weighting the answer categories from 1 to 6 and then scaling the range from 0% to 100%. E.g. if all Subjects would have marked ´very good´ an overall score of 100% would have been resulted. If all Subjects would have marked ´very bad´ an overall score of 0% would have been resulted.

The next Figure 40 is concerning the ease of use and shows, that the use of speech operation was easier with the present systems as compared to the baseline system (free run). With the present systems by far most of the Subjects tended to a positive rating (Ratings on the left side: SAMMIE: 90%, C&C: 80%). There was a relatively slight difference between SAMMIE and C&C.

For the present systems the answers were spread over the positive part of the scales. I.e. the systems were felt to be easy, but the degree of ease of use was judged interindividually differently. There is a significant correlation between the answers to question 2 and the respective rejection rates of the systems (Pearson correlation coefficient r=-0,41, p<0,05). [43]





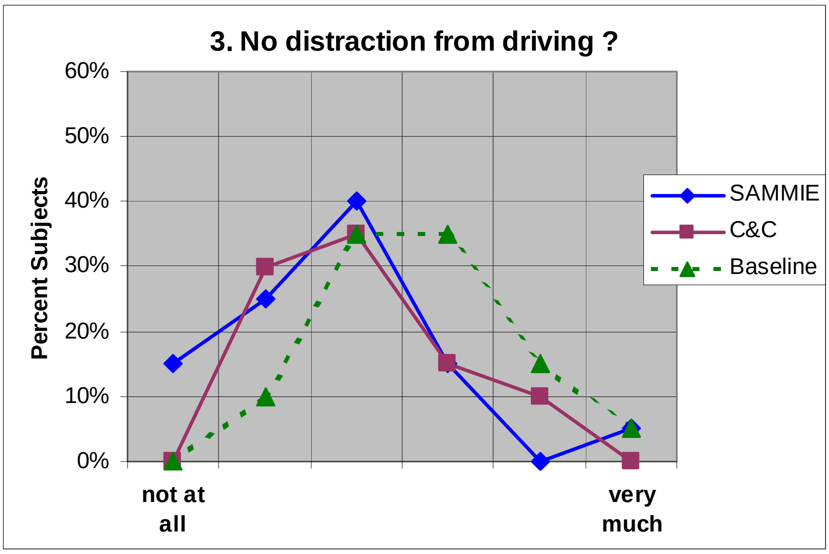
	mean
SAMMIE	75%
C&C	77%
Baseline	65%

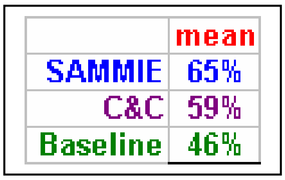
Figure 40: Answers to the questions "2. How easy was the system operation for you?"

Question 3 concerned the problem of distraction from driving, s. following Figure 41. Corresponding to the preceding figures there is a similar curve between the present systems, apart from the highest ranking of "not at all". A certain distracting effect was felt by most of the Subjects, with C&C more than with SAMMIE, as the maximums are near the scale centre. (The Pearson correlation coefficient of r=0,21 between the answers to this question and the individual driving errors is, however, not significant.)

The SAMMIE and C&C systems were assessed to be less distracting than the baseline system (free run). But again, when comparing with the baseline evaluation results one should consider that the experimental setup was quite different, because the baseline evaluation used a driving simulation.

[43] It is conceivable, that there are still other correlations, e.g. to the number of turns, task duration etc., which is beyond the scope of this report.

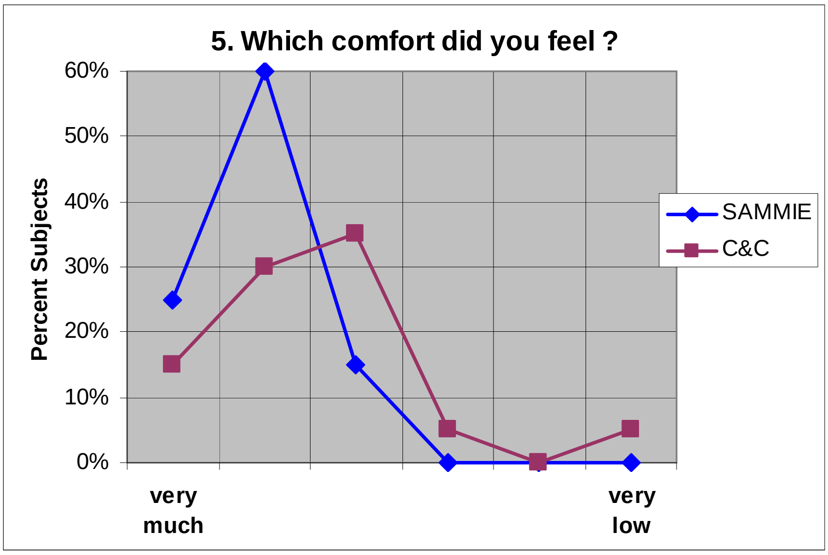


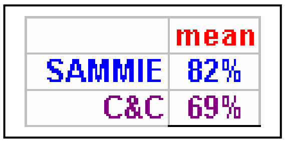


	mean
SAMMIE	65%
C&C	59%
Baseline	46%

Figure 41: Answers to the question "3. To which degree were you distracted from driving during operation?"

As the next Figure 42 about the felt comfort reveals, there was a clear advantage of the SAMMIE system over the C&C system. [44] All Subjects gave a positive answer with the SAMMIE system, by far most of them in the upper two categories. This distinct vote for SAMMIE as to comfort should be attributed to the Subjects´ experiences of one-input-tasks with SAMMIE speech input. [45] This subjective result is one of the most distinct ones concerning the comparison between the systems.





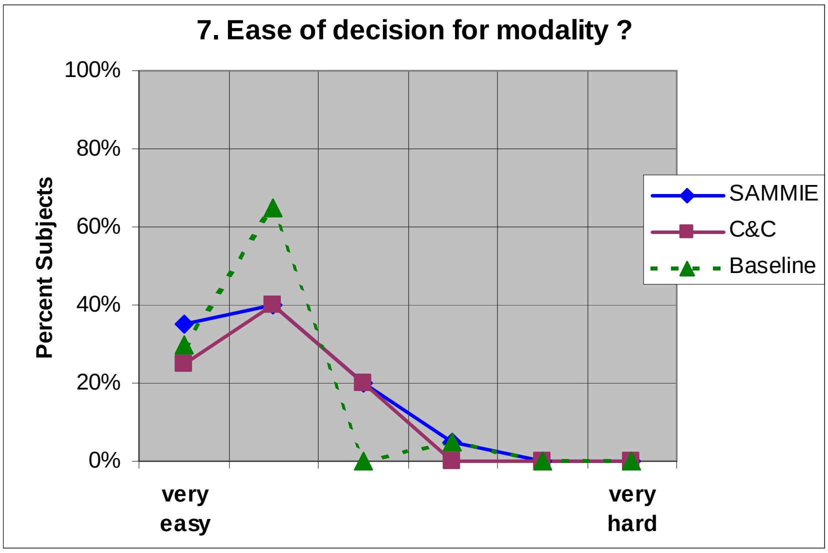
	mean
SAMMIE	82%
C&C	69%

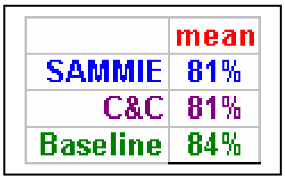
Figure 42: Answers to the question "5. Which comfort did you feel?"

As the next Figures (Figure 43, Figure 44) show, the decision and the change between modalities was easy or very easy for most of the Subjects. This is an important result for the concept of multimodality, since a change between modalities at pleasure is easily possible. Interesting is, that the decision was easier in the baseline study, which can possibly be interpreted with the steadily open microphone.

[44] If no baseline data are shown in the figures, the equivalent question was not put in the baseline study.

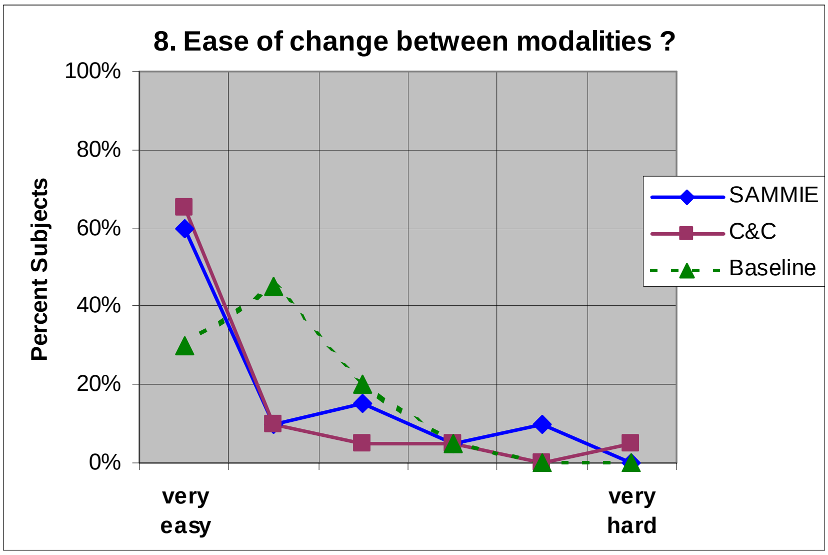
[45] Even if not all Subjects profited by this, they became acquainted with it within the video clip introduction.

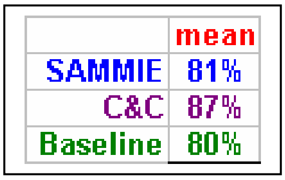




	mean
SAMMIE	81%
C&C	81%
Baseline	84%

Figure 43: Answers to the question “7. How easy was the decision for speech or manual input for you?”

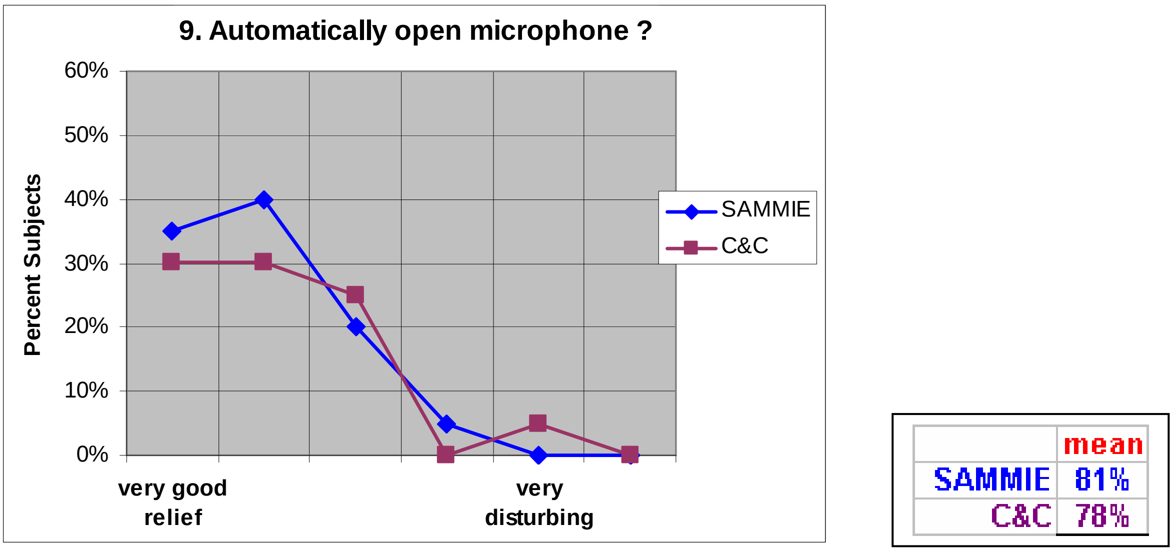




	mean
SAMMIE	81%
C&C	87%
Baseline	80%

Figure 44: Answers to the question “8. How easy was the change between speech and manual input for you?”

The next Figure 45 concerning the automatically open microphone demonstrates, that not all Subjects agreed completely with the autonomous opening of the microphone. There were many situations, where the Subjects continued an interaction with iDrive or talked to the passengers and the microphone opened autonomously. In those cases the system tried to understand the human communication which was irritating and intervened with the meanwhile progressed interaction.

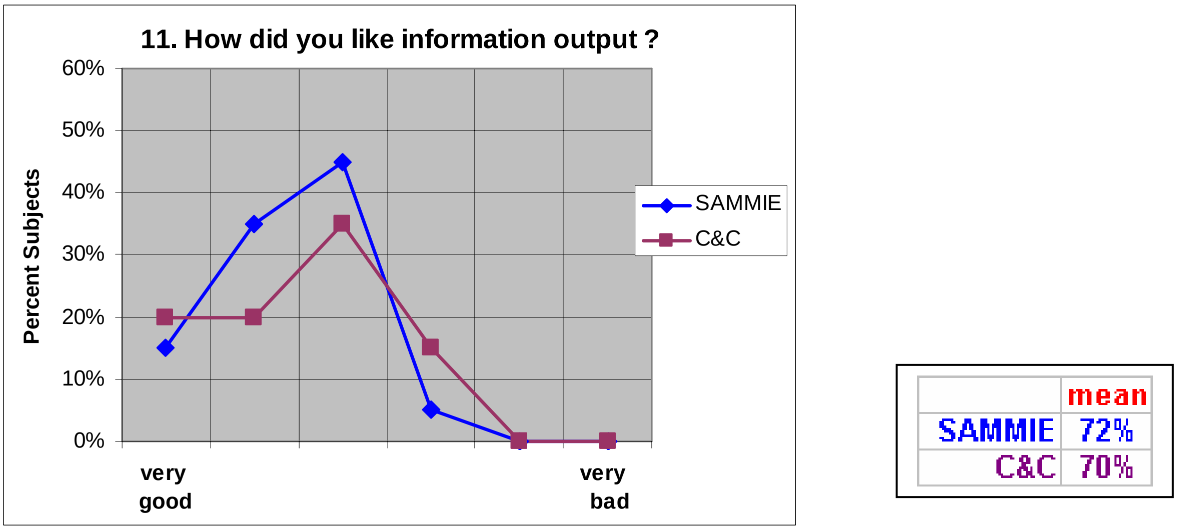


	mean
SAMMIE	81%
C&C	78%

Figure 45: Answers to the question "9. How do you judge the microphone characteristics, i.e. automatically open microphone?"

The next three figures were dedicated to the system output in general. [46] Concerning the Figure 46 and Figure 47 with questions about the attitude to information output and the support by the system the maximum is at the 3. scale category. I.e. there is some reservation as to these criteria of the system. This had sometimes to do with the extent of speech output and the restricted context sensitivity of the helps. The system outputs did not resolve a user disorientation in each case.

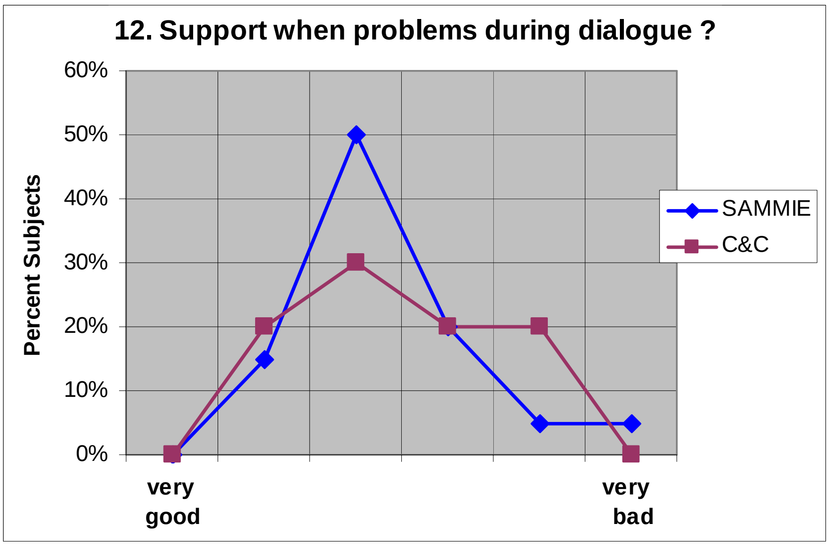
The question in respect to information distribution between speech and display presentation (Figure 48) was answered more positively, but still with even some negative judgements. There was a tendency to the opinion, that there were sometimes too much spoken outputs, e.g. the hint to the help system.

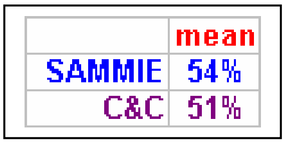


	mean
SAMMIE	72%
C&C	70%

Figure 46: Answers to the question "11. How did you like the system output (optically, acoustically)?"

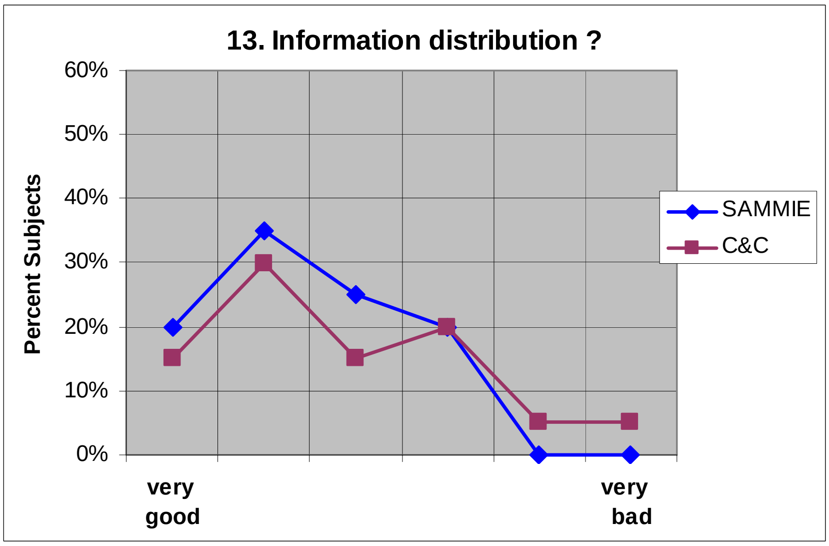
[46] The equivalent questions in the baseline study were put several weeks after the study with a considerable recollection problem, so that the baseline data are not included here.

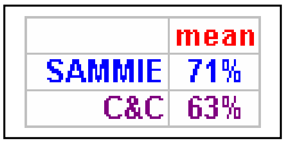




	mean
SAMMIE	54%
C&C	51%

Figure 47: Answers to the question "12. Were you supported in case of dialogue problems?"

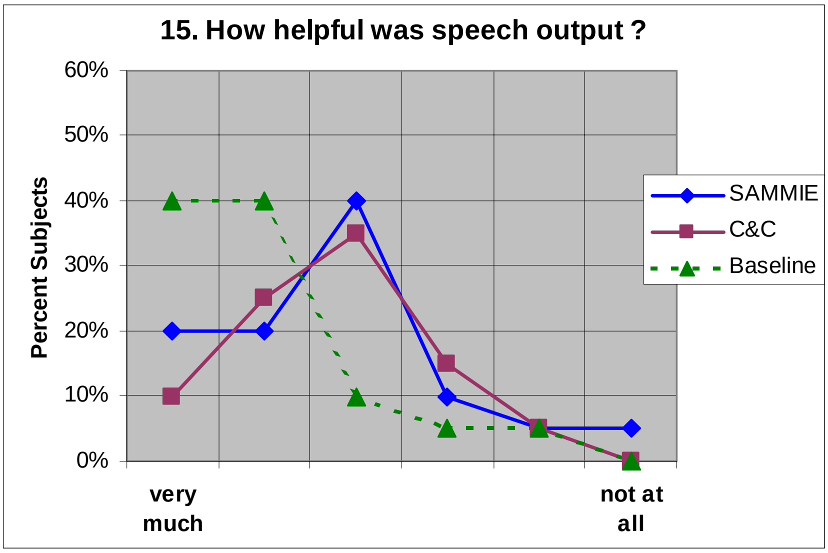


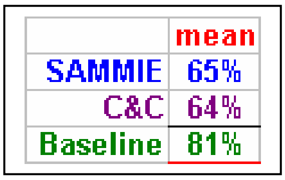


	mean
SAMMIE	71%
C&C	63%

Figure 48: Answers to the question "13. How do you judge the distribution of the information to speech output and display?"

The next figures concern the subjective evaluation of speech output. (Figure 49 - Figure 53). There is a general trend towards a positive judgement, but often clearly below maximum. Speech output was judged to be less helpful in the present study than in the baseline study (Figure 49)! This can be interpreted in terms of the actual good display presentations (see below) with too much speech information now or vice versa in the baseline system. Another explanation could be, that the simulated driving task was more demanding than the real driving, so that Subjects were more dependent on speech output.

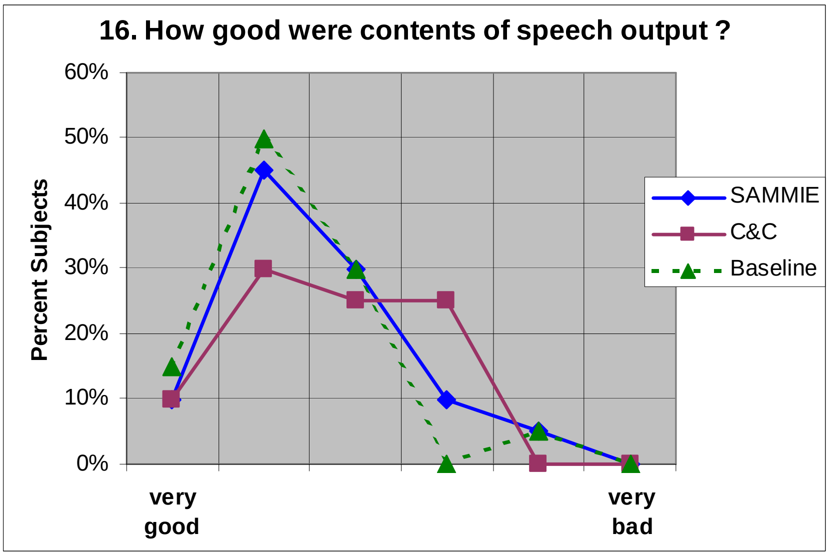


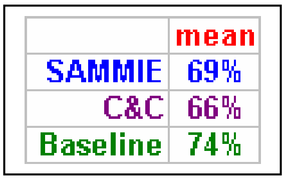


	mean
SAMMIE	65%
C&C	64%
Baseline	81%

Figure 49: Answers to the question "15. How helpful were the speech outputs for you?"

The contents of speech output was assessed worse with C&C system than with SAMMIE system (Figure 50). This can be associated with the verbal listings of items, which was not accepted by a part of the Subject sample.





	mean
SAMMIE	69%
C&C	66%
Baseline	74%

Figure 50: Answers to the question "16. How did you judge the contents of speech output?"

The extent of speech output was often judged to be relatively good (Figure 51). [47] There was a slight tendency, that Subjects felt SAMMIE speech output to be somewhat too extensive, possibly because of more rejections and hints to the help system. [48]

The formulation of speech output was regarded as rather positive, better with the SAMMIE system than with the C&C system (Figure 52). This can possibly be attributed to the general trend to judge the C&C speech output worse than the SAMMIE speech output ("Halo-effect", i.e. the generalization of the judgement in respect to one aspect to the judgement of others).

[47] The Subjects missed here a central scale category = ´OK´.

[48] When a dialogue seemed to be in a deadlock, the system offered a help with the announcement: "Wählen sie einen der folgenden Menüpunkte: Wiedergabelisten. Interpreten. Alben. Titel. Musikrichtungen. Mit dem Kommando 'Hilfe' erhalten sie jederzeit nützliche Informationen zur Bedienung des Systems."'

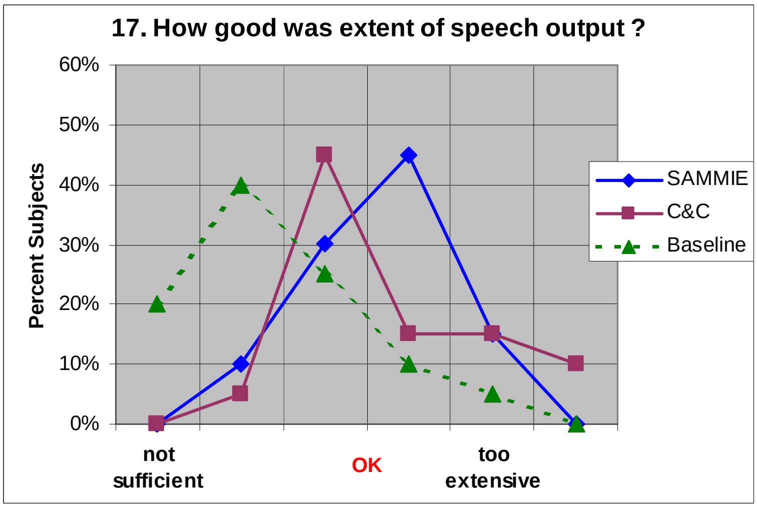
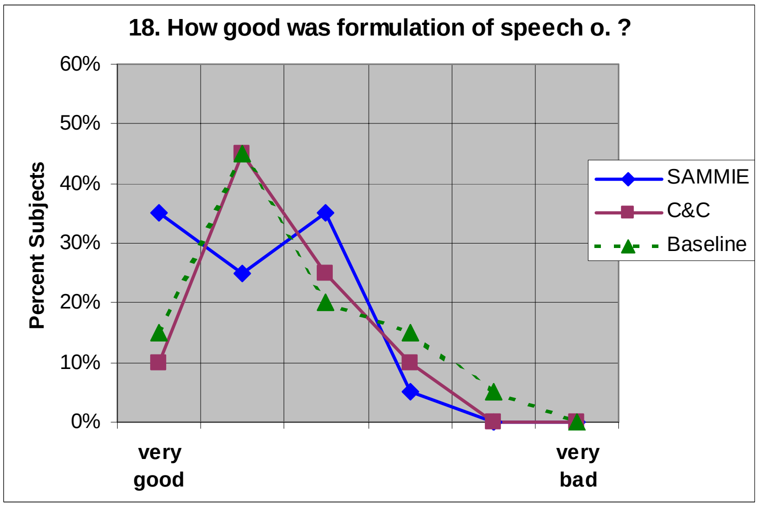
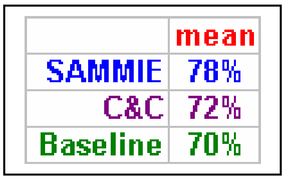


Figure 51: Answers to the question "17. How do you judge the extent of speech output?"

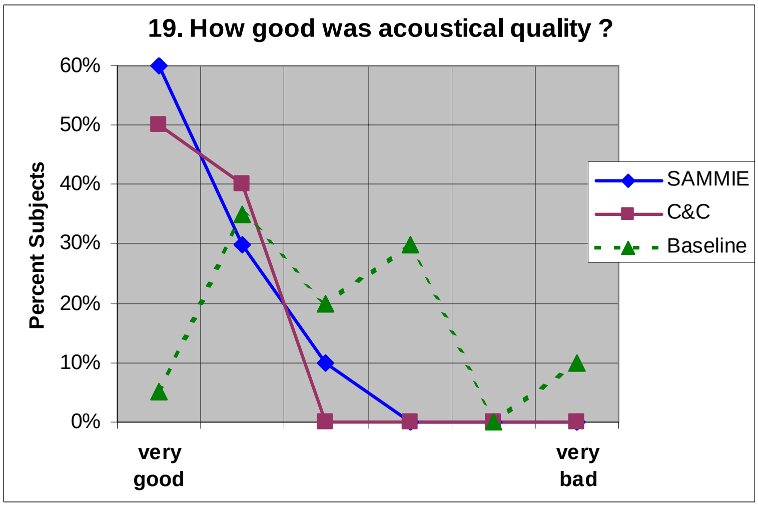


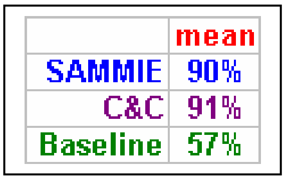


	mean
SAMMIE	78%
C&C	72%
Baseline	70%

Figure 52: Answers to the question "18. How do you judge the formulations of speech output?"

Very high scores got the present system in respect to the acoustical quality (Figure 53). This was stated several times spontaneously during the runs, too.

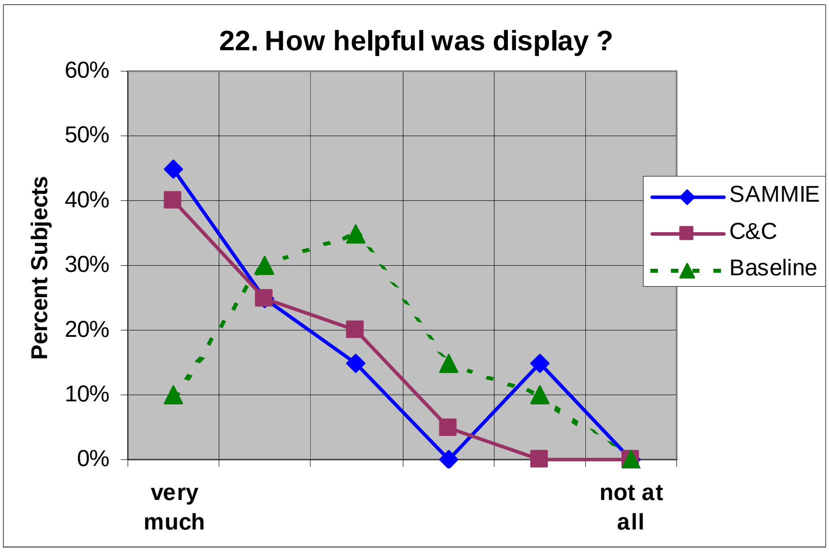


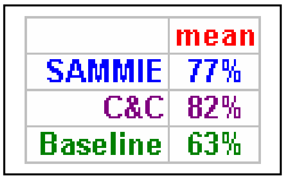


	mean
SAMMIE	90%
C&C	91%
Baseline	57%

Figure 53: Answers to the question "19. How good was the acoustical quality of speech output?"

The next figures concern the subjective evaluation of the display (Figure 54 - Figure 57). There is a general trend towards a positive judgement, but often clearly below maximum. The display was judged to be more helpful in the present study than in the baseline study (Figure 54), which is contrary to the judgements as to speech output. As the informal interview revealed, the display was felt to be clear and easy to survey, which was not the case in the baseline study.

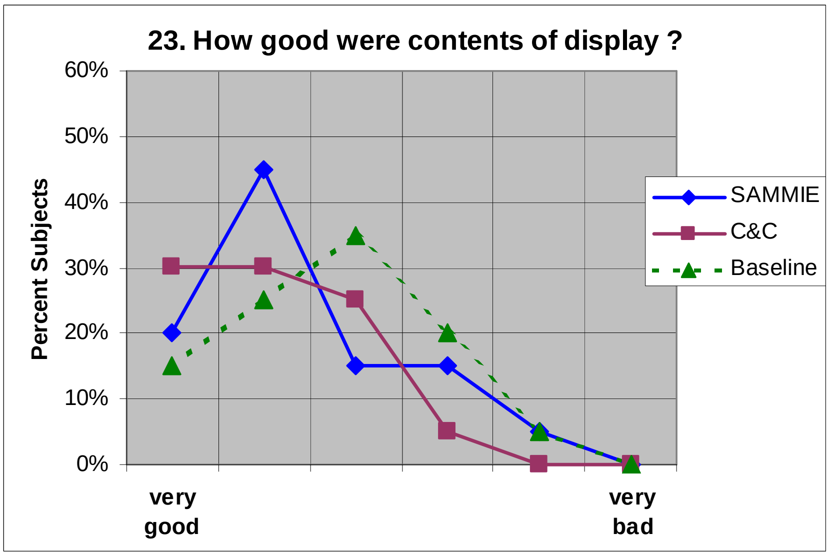


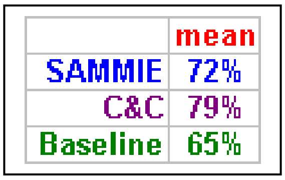


	mean
SAMMIE	77%
C&C	82%
Baseline	63%

Figure 54: Answers to the question "22. How helpful was the display for you?"

Similarly, the contents of the display was judged to be good, particularly better than in the baseline system (Figure 55).





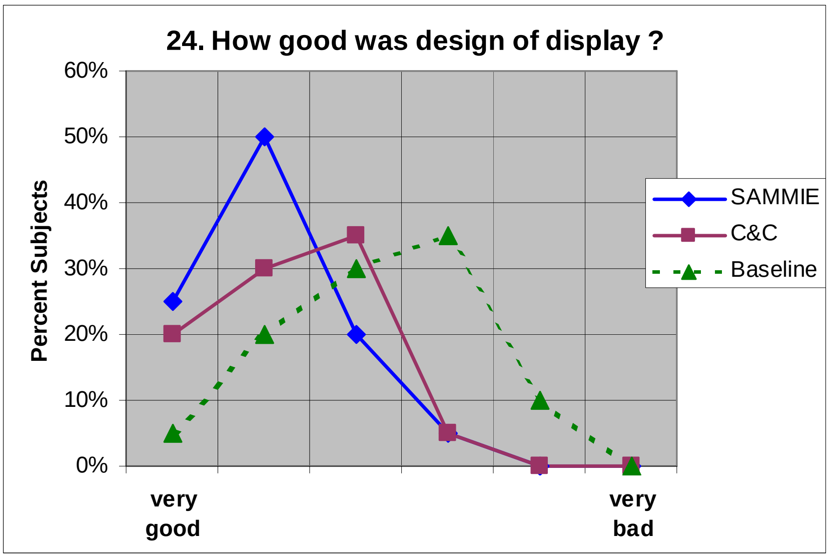
	mean
SAMMIE	72%
C&C	79%
Baseline	65%

Figure 55: Answers to the question "23. How did you judge the contents of the display?"

Concerning the design of the display, there were some reservations in respect to the C&C system (Figure 56). The Subjects missed here the specifications in the headings.

The extent of the display was regarded as OK (Figure 57) [49]

[49] The Subjects missed here a central scale category = OK



	mean
SAMMIE	79%
C&C	74%
Baseline	55%

Figure 56: Answers to the question "24. How did you judge the design of the display?"

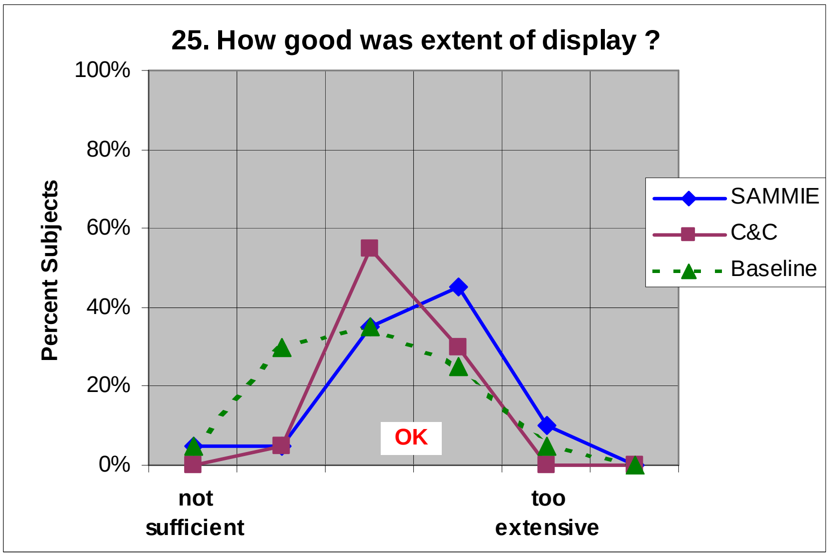


Figure 57: Answers to the question "25. How do you judge the extent of the display?"

The next two figures show the answers to the statements concerning the dialogue (cf. Communicator evaluations [4]), separated for the SAMMIE and the C&C system (Figure 58, Figure 59). The answers were spread over the first four categories, with no clear preference for one of the systems.

The best scores got the statement concerning the understanding of what the system said. It is not clear, however, if the statement was conceived as acoustical or content-related understanding.

A relatively bad judgment refers to the statement, that it was easy to get the information which the user wanted, particularly with the C&C system. Actually, the Subjects were often disoriented about the present system state, e.g. though they asked for a specific album, they still were in the general album menu level because of misunderstandings. This holds true more for the C&C system.

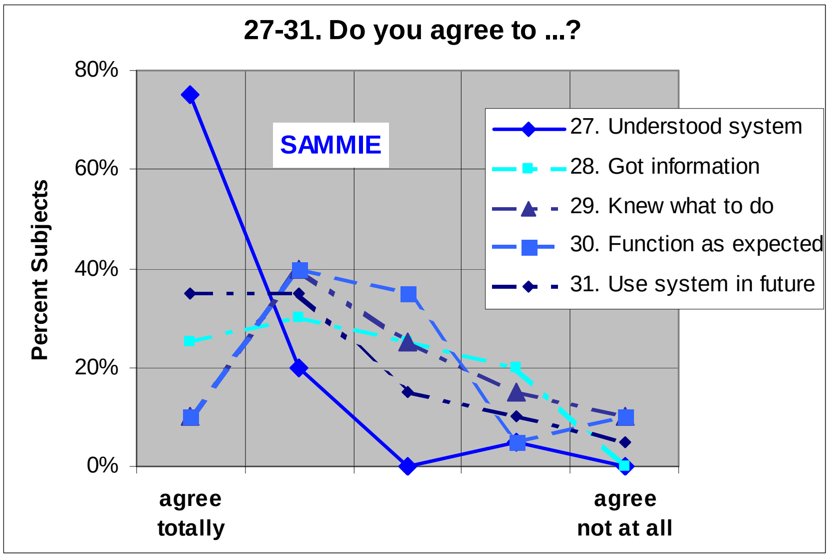


Figure 58: Answers to the statements of questions "27-31. Do you agree to the statements…?" for SAMMIE

	mean
27. Understood system	91%
28. Got information	65%
29. Knew what to do	56%
30. Function as expected	59%
31. Use system in future	71%

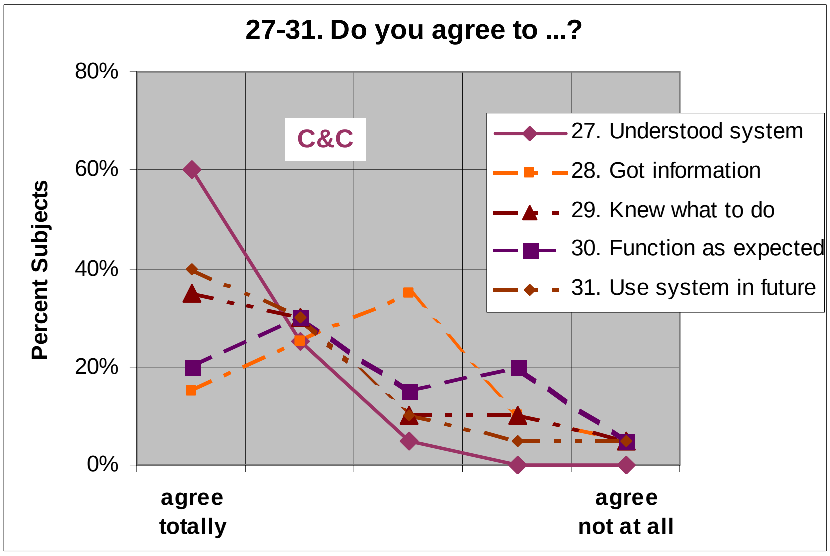


Figure 59: Answers to the statements of questions "27-31. Do you agree to the statements…?" for C&C

	mean
27. Understood system	90%
28. Got information	60%
29. Knew what to do	72%
30. Function as expected	61%
31. Use system in future	76%

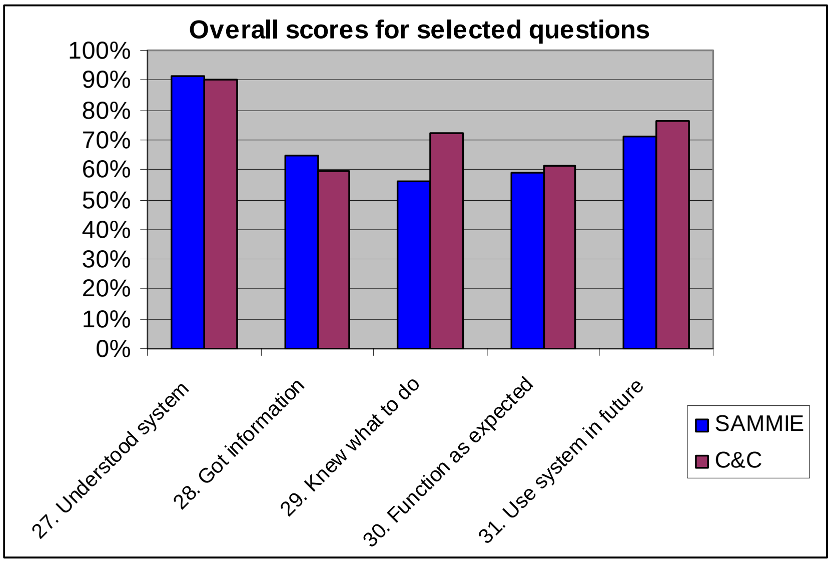


Figure 60: Overall scores for selected questions

The last Figure represent the overall scores of the selected questions concerning the dialogue.

In the intermediate questionnaires there were some open questions concerning general remarks to the speech output, the display presentation and the SAMMIE operation system (question 20, 26, 32).

In the SAMMIE questionnaire there were single comments as to system output like:

"Speech output has become more melodious and therefore more comprehensible as compared to the baseline system"

"I could not take advantage of the speech output, since I was understood seldom"

"It should not say, what has been done, but that something has been done"

"Display was very simple and clear"

"In long lists the selection of a letter would be good"

"Priorities should be set to 'Which album and song is currently playing' "

"Button for submenu, e.g. create playlist"

"Accommodation of the eyes is a problem"

Concerning the open evaluation of the SAMMIE system on the whole five Subjects stated that the SAMMIE system has become faster or better than the baseline system. There were statements like:

"System understands better and works faster than the TALK Baseline system"

"System has been very much improved"

"Complete tasks in one sentence are appropriate to reduce distraction"

"It is a pleasure to work with the system, though series maturity has not been reached, yet"

In the C&C questionnaire there were single comments as to system output like:

"I was not always informed, when I was not understood and what I can choose"

"Speech output detains"

"After a change from speech input to iDrive speech output should be stopped"

"Speaking speed something too slow"

"Lists should be specified by speech output" / "Lists should not be specified by speech output"

"Time delay until a song has been found. Egg-timer icon!"

"Not enough information on display"

"Picture of the album is superfluous"

"Details like song duration, number of list elements etc. are missing"

Concerning the open evaluation of the C&C system on the whole two Subjects found it better than the SAMMIE system. There were statements like:

"iDrive device positioned too far backward" (which was confirmed by several female Subjects informally)

" I expect a natural spontaneous speech input"

"Speech understanding problems with low voice and high surrounding noise, e.g. in tunnel"

"More reliable but less fun"

"One song chosen, then all other songs were played back, which I did not want"

"Microphone icon should be near tachometer"

During the sessions Subjects uttered spontaneously or they were asked by the experimenter about their behaviour.

During the SAMMIE runs there were statements like:

“Thinking about formulation is strenuous”

“A complete sentence takes longer than a command”

“Understanding problems in the tunnel”

“Angry when not understood”

“Safe control of car or system”

“Annoying if another song instead of ´that song´ is included into the playlist”

“I looked for ´The Beatles´ at ´B´”

“I relied on acoustical dialogue, but recollecting is difficult”

“I would like to adjust volume and bass by speech input”

“A permanent open microphone when driving alone”

“Music should become lower when PTT is activated”

“A shuffle mode would be good”

During the C&C runs there were statements like:

“Distraction by system errors”

“Mental load by distraction”

“Display ´Kein Lied geladen´ is irritating”

“I turned off the music to prevent disturbance of the speech input”

“Not clear if I have to wait for end of speech output to proceed manually”

Concerning iDrive there were statements like:

“Position too far backwards”

“Faster. At the beginning speech input because of being new device”

“Better structured input and overview”

“I did not think to proceed in the list by turning”

“Delay time with iDrive is irritating”

4.2 Final questionnaire

The Subjects completed the final questionnaire at home, i.e. after having got known both systems. It contained several questions in respect to a general view of the multimodal interaction during driving.

Question 1 "Which input modality would you prefer in the long run?" was asked, because it was assumed, that the learning effect was still pending during the sessions. As the next Figure 61 illustrates, there was a slight preference for the C&C system. This is an unexpected result, because the C&C system was meant as a reference system for the SAMMIE system. This could be attributed to the better speech recognition performance of the C&C system. Possibly, the better orientation along the menu with C&C is another reason for it.

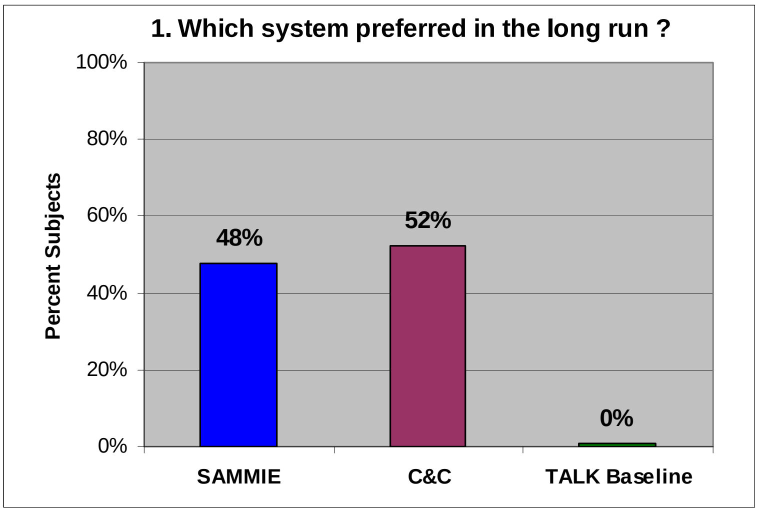
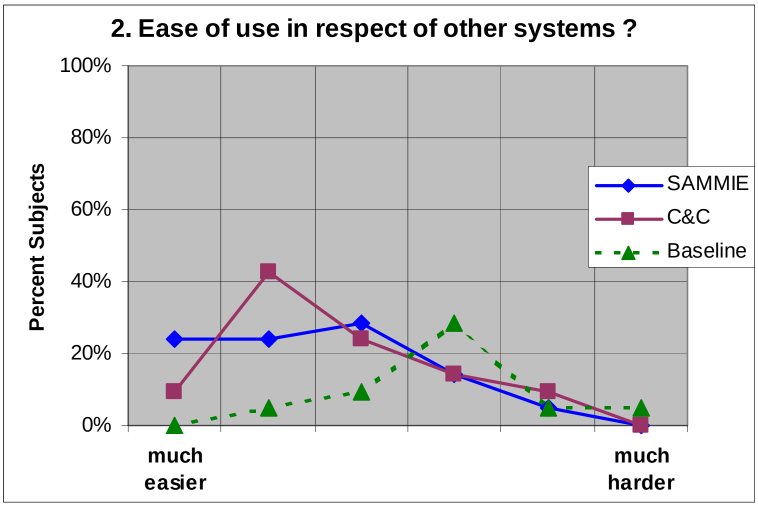
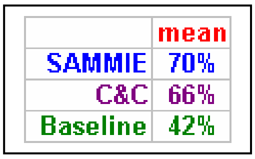


Figure 61: Answers to the question "1. Which system would you use in the long run?"

The ease of use of the present systems were judged much better than the baseline system (Figure 62). While in Figure 40 with a similar question original data of the baseline study were used, the data here were collected with knowledge of both systems and with new data as to the baseline study.





	mean
SAMMIE	70%
C&C	66%
Baseline	42%

Figure 62: Answers to the question "2. How easy were the systems to operate in respect to other systems?"

In questions 3 - 7 the Subjects were asked about the advantages and disadvantages of the input modalities. There were five options, whereby several options could be checked.

The following Figure 63 represents the answer frequencies concerning advantages and disadvantages of the natural speech input with SAMMIE. The safety aspects dominated as in the baseline system (´no averting glances´). The possibility to formulate freely and the new technology were pronounced much more frequently than in the baseline study.

´Looking for formulations´ was pronounced nearly as often as in the baseline study. That can be interpreted either by a still missing acceptance of the still restricted formulation freedom or by the instruction to formulate in whole sentences. There was a considerable decrease of number of Subjects, who objected to the longer inputs.

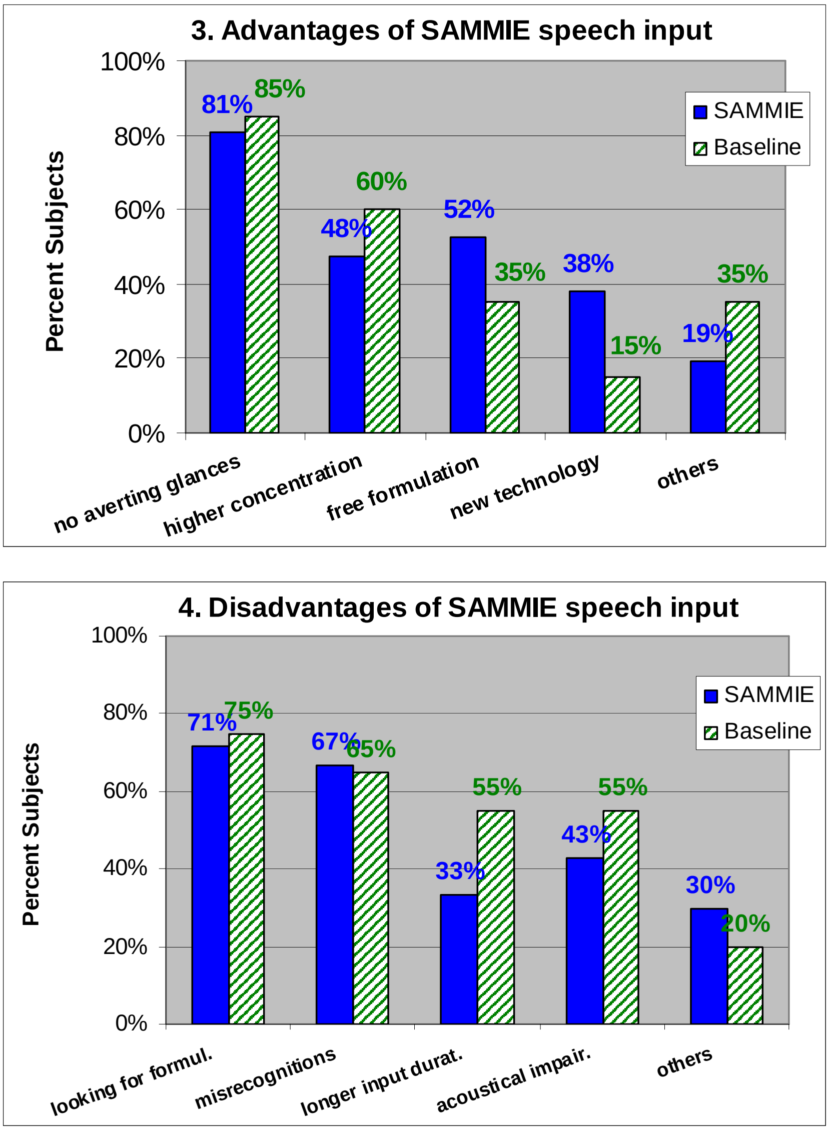
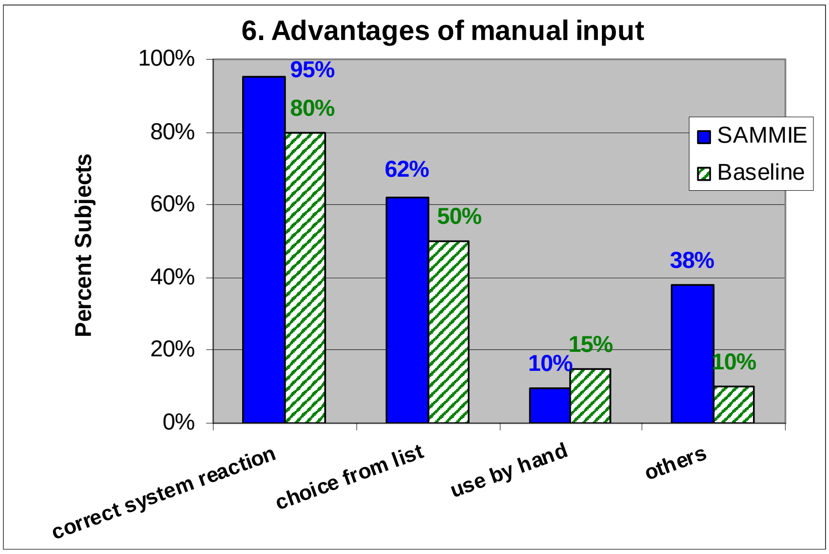


Figure 63: Answers to the questions “3./4. Which of the following aspects represented advantages / disadvantages of speech input for you in relation to the manual input?” “

The following Figure 64 represents the answer frequencies concerning advantages and disadvantages of manual input. Again, the option ´correct system reaction´ was pronounced most frequently, even more frequently than in the baseline study. The (easy) choice from a list was pronounced next. As an additional advantage the faster operation was noted by 5 Subjects.

The main disadvantages were the safety aspects like ´eyes off road´, ´hands off steering wheel´ and ´searching by hand´. Since the first option was not included in the baseline study, the related options (´searching by hand´, ´searching by eyes´) were presumably pronounced more frequently than in the present study.

As other disadvantages of the manual input similar aspects were noted, like searching iDrive button and display cursor, position of the iDrive button too far backward.



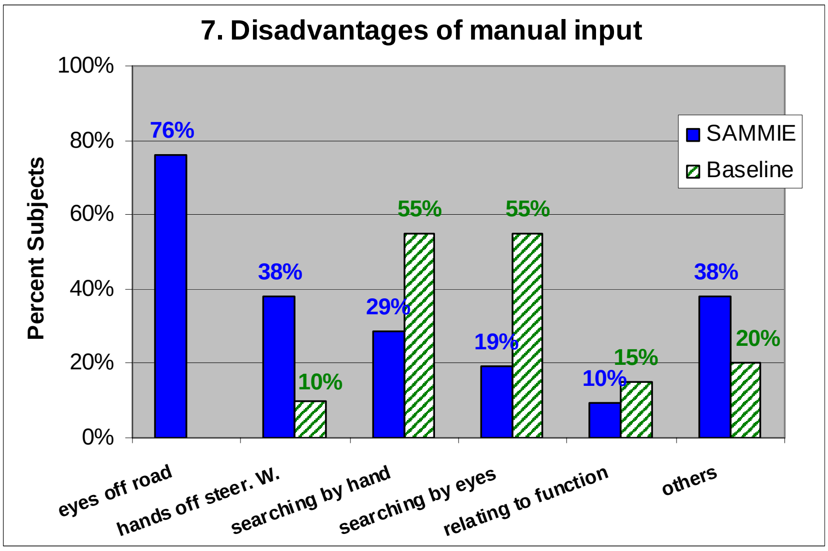
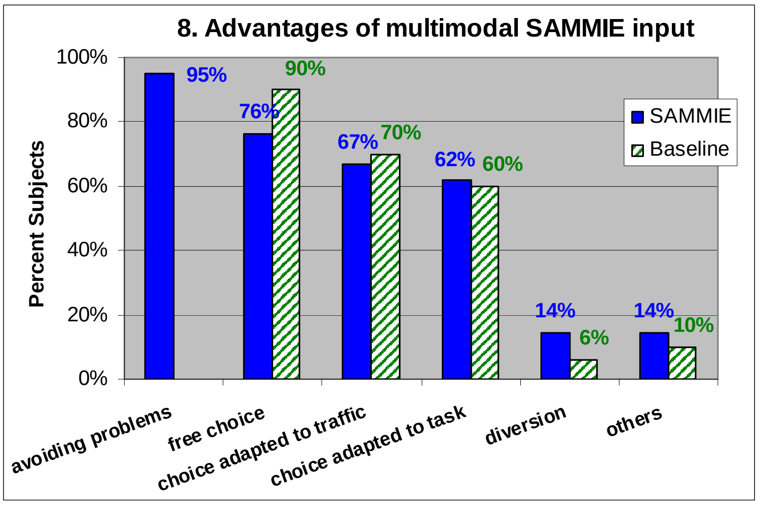


Figure 64: Answers to the question "6./7. Which of the following aspects represented advantages / disadvantages of manual input for you in relation to speech input?"

The following Figure 65 represents the answer frequencies concerning advantages and disadvantages of the multimodal input. The most frequently pronounced advantage concerned avoiding the problems of the other modality, which was pronounced by nearly all Subjects (This question was not asked in the baseline study.) The next frequently pronounced option was ´free choice of the operation mode´. I.e. a main motivation for SAMMIE – the free option of input modality – was felt positively by a considerable part of the sample. More than the half of the Subjects pronounced the aspects of adaptation to traffic and tasks.

Compared to the advantages, there were much less disadvantages pronounced. The main disadvantage was the uncertainty as to which task was feasible by which input device, which was more severe in the baseline study. This result is somehow astonishing, since all but one task (creating a new playlist) was feasible by both modalities. The need for a choice between input modalities was no longer pronounced by anyone.



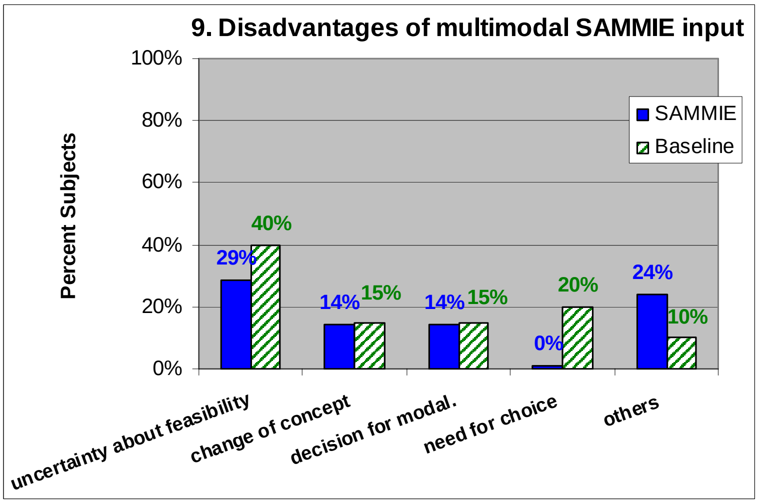


Figure 65: Answers to the question "8./9. Which of the following aspects represented advantages / disadvantages of multimodal input for you?"

Being asked, which functions the Subjects would like to use in the car by the multimodal interaction, including the natural speech system SAMMIE there was partly a different order as compared to the baseline study (s. Figure 66). Now most of the Subjects pronounced the more advanced functions like desk diary, navigation and internet, while infotainment functions were represented more frequently in the baseline study.

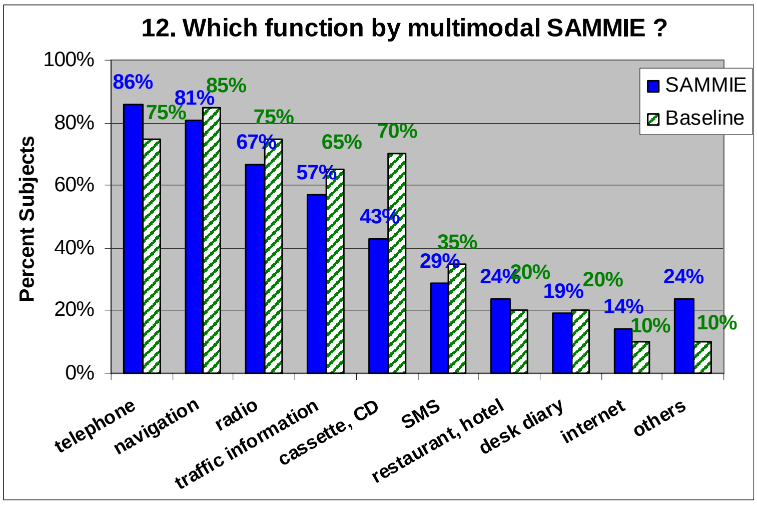


Figure 66: Answers to the question "12. Which functions in the car would you like to use with the multimodal input?"

In the last question 15 the Subjects were requested to give improvement suggestions. Following statements were done:

• Better speech recognition	4x
• Better adaptation to speech level	2x
• Other iDrive position (possibly at the steering wheel)	2x
• Random selection	2x
• Other display design / better display quality	2x
• Higher flexibility concerning formulation	1x
• Stopping speech output when manual input starts	1x
• Combination between SAMMIE and Command system	1x
• Higher sensitivity to the dialogue context (e.g. no erasing instead of playing back)	1x
• Additional functions (charts, statistics etc.)	1x
• Submenu	1x

4.3 Questionnaire: Adaptive / Non-adaptive SAMMIE

After the runs with the Full (Adaptive) SAMMIE and C&C system the Non-Adaptive SAMMIE system was presented at the end of the session in form of 6 examples (s. Attachment 5). An example consisted of a double presentation first with Full SAMMIE, second with NA SAMMIE. Each example was dedicated to one specific feature, which was differing between Full and NA SAMMIE. After each example the equivalent question to this feature was asked (s. Attachment 5).

The following figures show the answers to these questions. While all other features were judged positively, the usefulness of the personal differentiated addressing by "Sie / Du" was scored rather negatively (Figure 67). But there was a group of 30% who pronounced the second highest category, i.e. the Subjects were divided in respect to this feature.

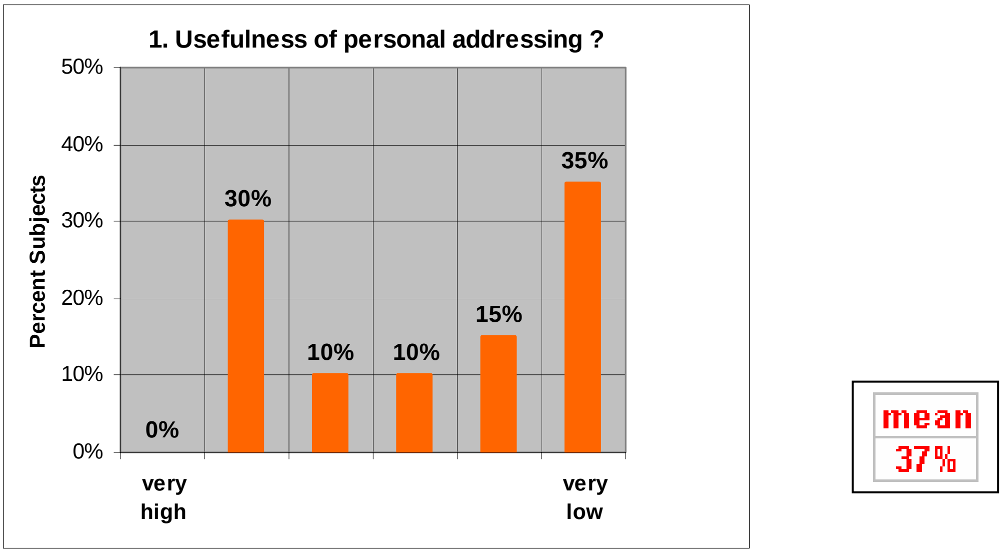


Figure 67: Answers to the question "1. How useful is the differentiated addressing by "Sie / Du"?"

The feature of a differentiated function for visual / acoustical presentation of artists / albums / songs was basically positively judged (Figure 68). But there was a certain range over the positive categories, which may be interpreted as a possibly disturbing effect of long spoken lists.

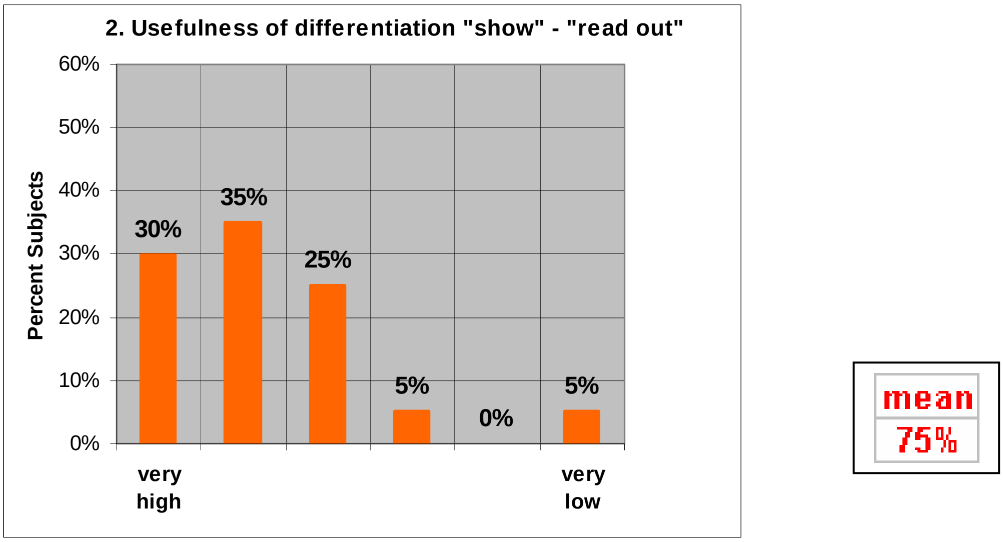


Figure 68: Answers to the question "2. How useful is the differentiation between "show" and "read out"?"

The features ´Presentation of albums with artists´ (Figure 69) and ´Implicit confirmation´ (e.g. the feedback of the entered artist as a headline, Figure 70) are clearly positively judged.

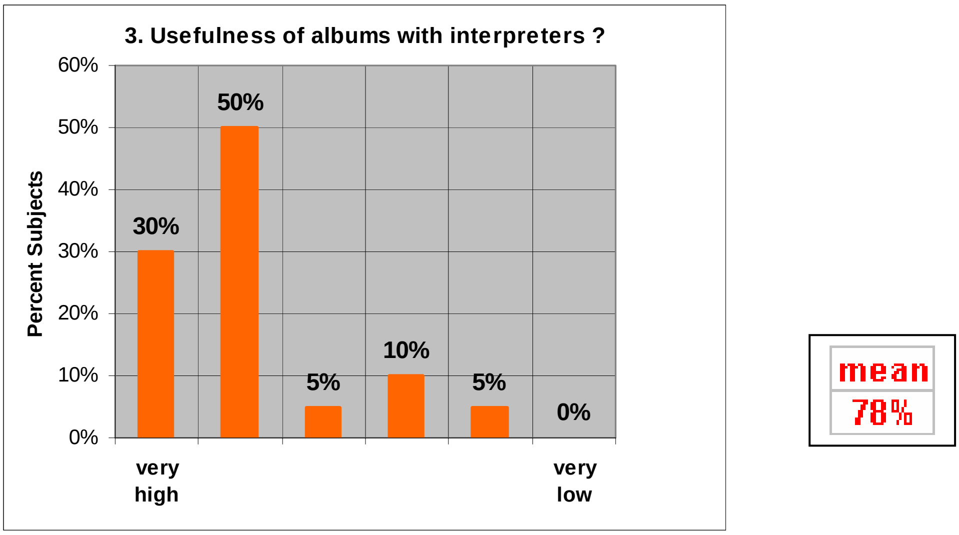


Figure 69: Answers to the question "3. How useful is album presentation with artists?"

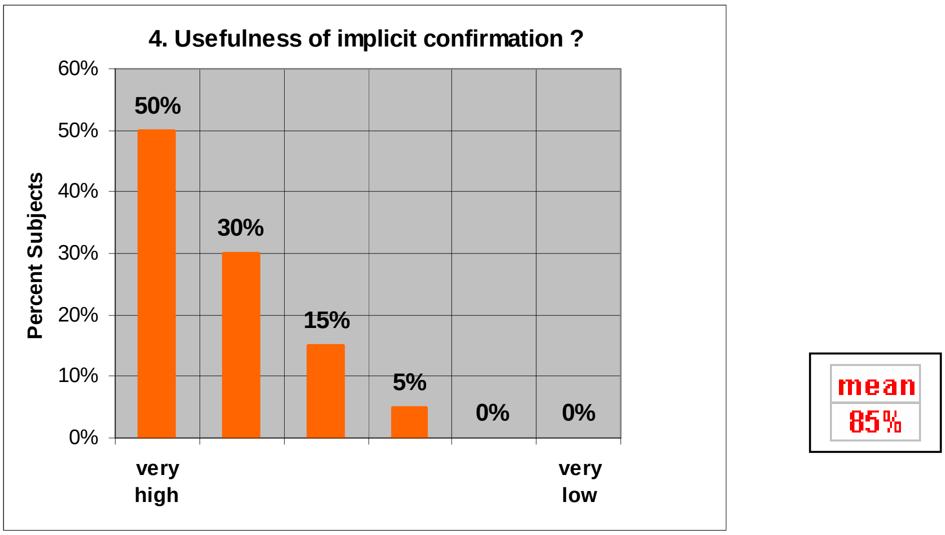


Figure 70: Answers to the question "4. How useful is the implicit confirmation?"

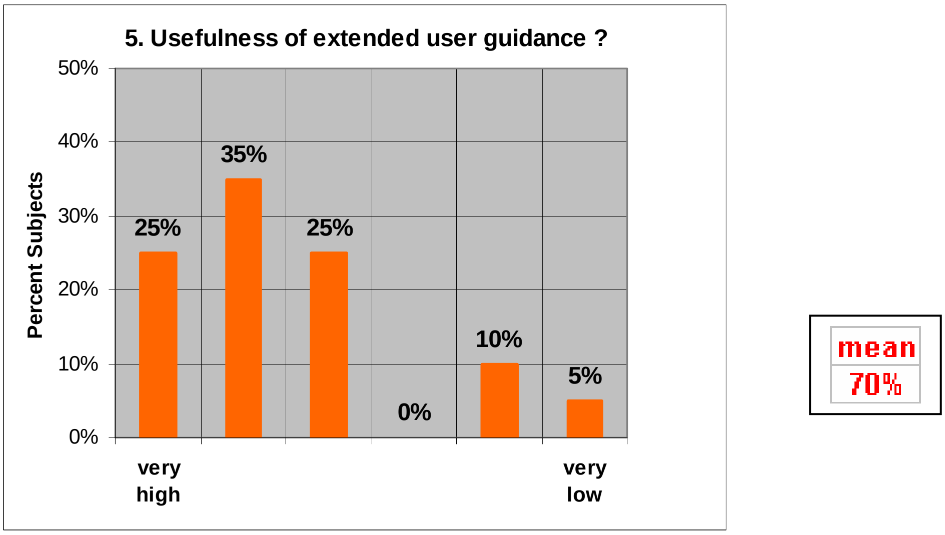


Figure 71: Answers to the question "5. How useful is the comprehensive user guidance?"

Concerning the extended user guidance was regarded as basically positive (s. preceding Figure 71). But most of the answers were distributed over the three positive, partly non-maximal rating categories. The step-by-step guidance was not totally accepted, presumably because of the somewhat lengthy dialogue.

The feature of an adaptation to the user's vocabulary was judged very diversely (Figure 72). In spite of the tendency to a positive acceptance, there was a group of 40%, who had a more or less negative attitude to the usefulness of this feature. As the informal statements showed, this was seen as a marginal feature.

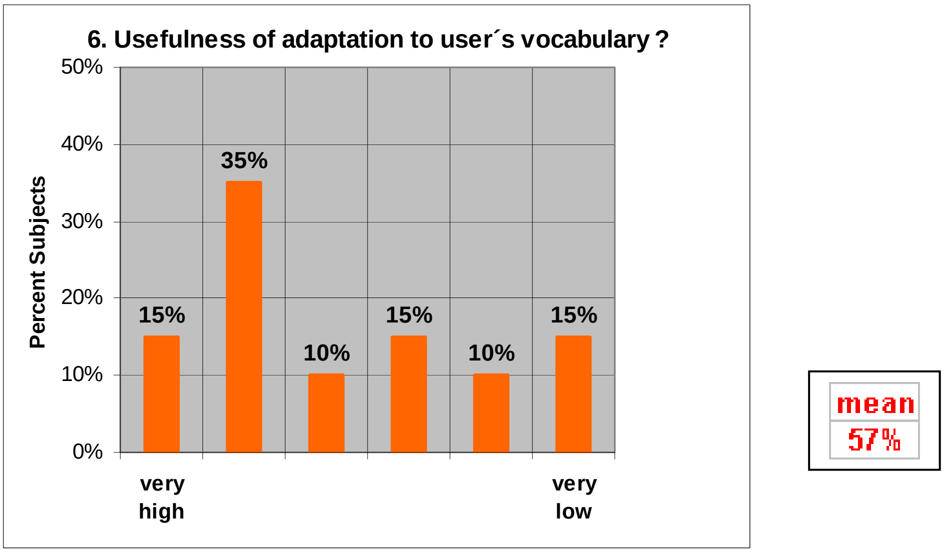


Figure 72: Answers to the question "6. How useful is the adaptation to the user´s vocabulary?"

After the questions to the single features a general question summarized all features with an emphasis on "advantage for you" (Figure 73). The order is reflecting to some extent the individual judgements. But the extended user guidance is now ranking higher, the implicit confirmation lower. If no immediate presentation is preceding, an extended user guidance seems basically to be positive.

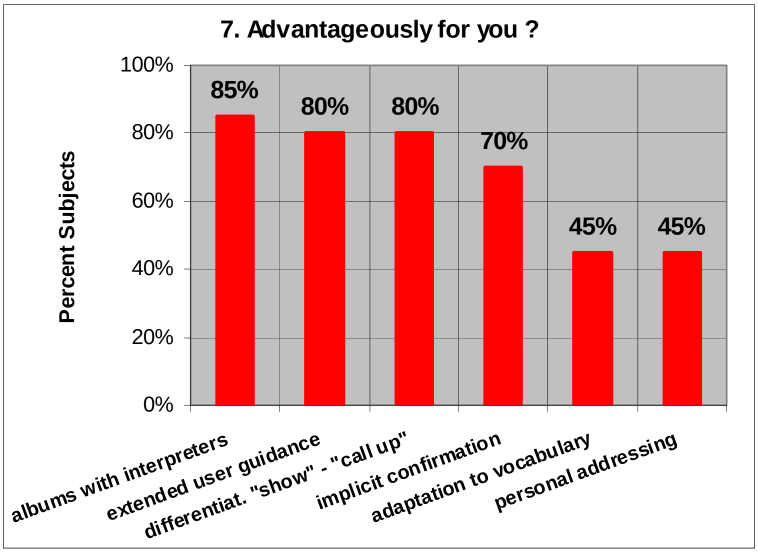


Figure 73: Answers to the question "7. Which features are useful for you?"

In the last question the preference for one of the systems is asked for (Figure 74). [50] There was no single vote neither for NA SAMMIE nor for the baseline system. The C&C system is after the direct experience better accepted than the Full SAMMIE system! The NA SAMMIE was preferred by nobody since all features were presented and judged negatively in the additional part of the session.

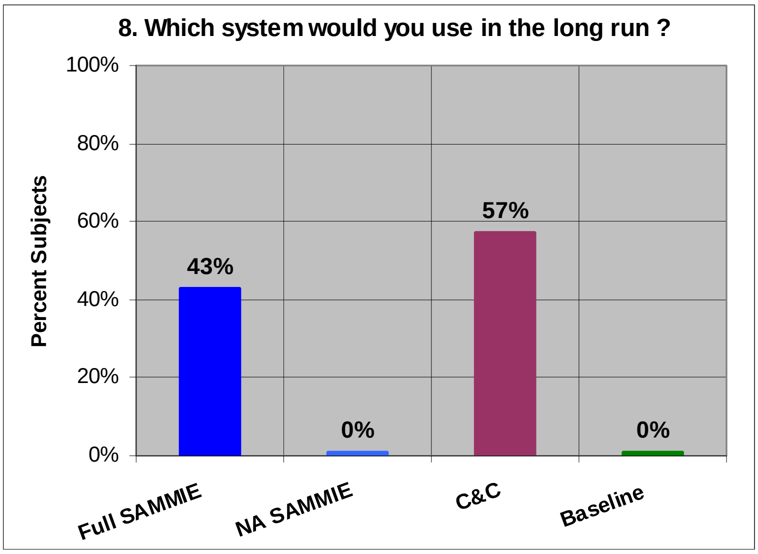


Figure 74: Answers to the question “8. Which system would you use on the long run?”

[50] This question – here with the NA SAMMIE included - was repeated in the final questionnaire.

4.4 Statistical tests

Originally, there were several hypotheses as to several performance and dialogue criteria of the systems. The hypothesis concerning the acceptance of the systems was:

"The SAMMIE system achieves a higher user acceptance as the NA system"

Acceptance is relatively well operationalised by the question 1. of the intermediate questionnaire ("General impression?") and question 8 of the NA SAMMIE ("Which system preferred?"). Concerning the general impression of the systems, asked in the intermediate questionnaire, a Wilcoxon Matched Pair test revealed, that there is no significant difference between systems (Wilcoxon Matched Pairs: n=18, T=19,5, p=0,41) [51]. I.e. immediately after the runs the Subjects had a similar positive impression of both systems (s. Figure 39).

Concerning the preference for one of the systems, asked in the NA SAMMIE questionnaire, a χ^2-Test was performed. The alternatives in the question were

a) Full SAMMIE
b) NA SAMMIE
c) C&C
d) Baseline

The alternative d) was only for those Subjects, who already participated in the baseline study. The result of the χ^2-Test depends on which systems are considered and which frequency was expected. If all three present systems or all four systems (including the baseline system) are considered, then the result was highly significant towards the preference of Full SAMMIE / C&C (e.g. four systems: χ^2=22, f=3, p<0,001). If just the two systems SAMMIE and C&C are considered, there was no statistical difference (χ^2=0,43, f=1, p=0,51).

Altogether, there is a tendency to a spontaneous better impression of SAMMIE (s. Figure 39), but for preferring the C&C system on the long run (s. Figure 74). But both results are missing significance.

The hypothesis concerning the distraction of the systems was:

"Full SAMMIE distracts from driving less than C&C"

Since the objective data of driving errors are very similar between the systems (s. chapter 3.6), the subjective evaluation concerning distraction is cited (intermediate questionnaire, question 3.) There was no significant difference between the systems (Wilcoxon Matched Pairs: n=18, T=19,5, p=0,41).

The statistical tests for the objective data are included into the corresponding chapters.

[51] Those Subjects were excluded, where the SAMMIE system was active instead of C&C plus Subject 1, where the questionnaires were structured differently, so that a n=18 resulted.

4.5 AttrakDiff

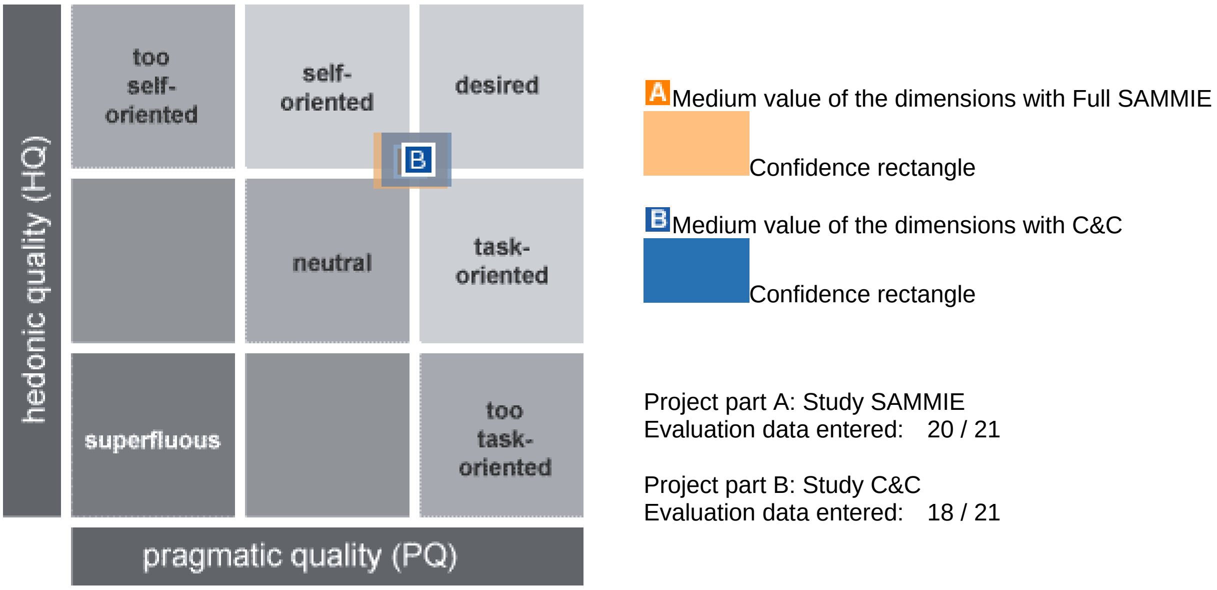
AttrakDiff™ [3] facilitates the evaluation of a chosen product by customers, user etc. The evaluation data makes it possible to assess how the attractiveness of the product is experienced, in terms of usability and appearance and whether optimisation is necessary.

AttrakDiff-1 was applied as an instrument of measurement in the form of semantic differentials. It consists of 23 seven-step items whose poles are opposite adjectives (e.g. "confusing - clear", "unusual - ordinary", "good - bad"). Each set of adjective items is ordered into a scale of intensity. Each of the middle values of an item group creates a scale value for pragmatic quality (PQ) , hedonic Quality (HQ) and attractiveness (ATT). The two constituent aspects of hedonic quality, namely stimulation and identity are separated.

The hedonic and pragmatic qualities are perceived consistently and independently of each other. Both contribute equally to the rating of attractiveness.

The data of the present study was used to simulate the participation of 20 Subjects (SAMMIE: all Subjects but No. 1) and 18 Subjects (C&C: all Subjects but No. 1, 14, 21). The following results were reported by AttrakDiff:

Overview of AttrakDiff Results



Project part A: Study SAMMIE
Evaluation data entered: 20 / 21

Project part B: Study C&C
Evaluation data entered: 18 / 21

Figure 75 shows the results for the dimensions pragmatic quality (PQ) , hedonic Quality (HQ) – identity (I) and stimulation (S) – and attractiveness (ATT).
For all dimensions the C&C systems performs slightly better than the SAMMIE system. This difference is however statistically *not* significant.

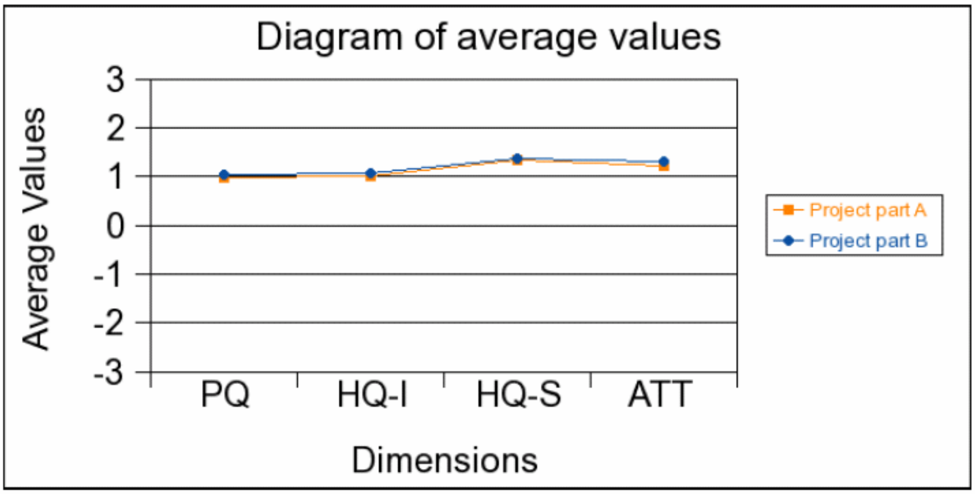


Figure 75: Mean values of the four AttrakDiff dimensions for the products "Full SAMMIE" (project part A) and "C&C" (project part B)

Word Pairs (Adjectives)

Figure 76 shows the mean values of the word pairs. Of particular interest are the extreme values. These show which characteristics are particularly critical or particularly well-resolved.

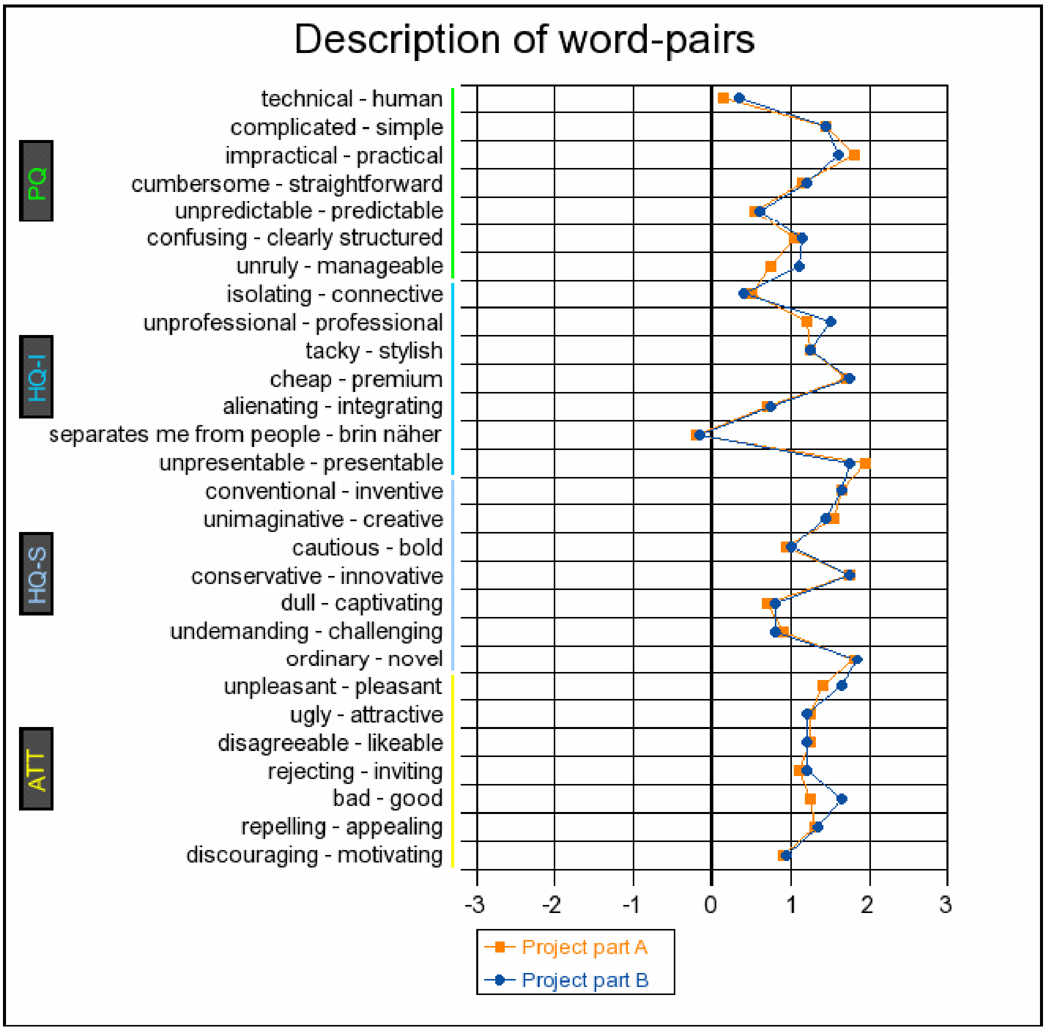


Figure 76: Mean values of the AttrakDiff word pairs for products "Full SAMMIE" (project part A) and "C&C" (project part B)

5 Summary

5.1 Objectives

The objectives of the evaluation study were to find out

- the usage of the multimodal systems
- the efficiency of the dialogue
- the acceptance of the systems
- the efficiency of the speech system
- influence onto driving quality

5.2 Methods

The experimental set-up for the user test comprised the experimental car BMW 335 including the iDrive button, a MP3 system, the full and non-adaptive SAMMIE system as well as the Command&Control (C&C) system and a video system including two cameras for the Subject and the traffic scene. An experimenter and a supervisor controlled the experiment and recorded the data.

The resulting experimental course was 34,5 km long for the SAMMIE run and 19 km long for the C&C run and was driven within 35 – 40 min and 20 – 25 min, respectively. The streets had two lanes with few or medium traffic or four lanes with medium or dense traffic. There were speed limits between 70 and 130 km/h.

A sample of 21 Subjects was recruited. Essential requirements for the participation were some or much experience with MP3 hardware or software and participation in the TALK baseline evaluation study, if possible. They were safe driver without strong dialect. The age was limited to the young and middle age group.

The basic principles for the tasks were to use a considerable number of tasks from the baseline study and covering the performance of the SAMMIE system. A sample of 10 tasks was chosen with browsing, playing back and information functions as well as playlist functions.

The study was conceived as critical experiment. The main variable was the multimodal interaction system. The Full SAMMIE system was the main system. The C&C system was used as a reference system as well as the baseline system. The Non-Adaptive (NA) SAMMIE system was presented at the end of the session to get a comparison to the Full SAMMIE system.

The SAMMIE and C&C system were balanced across Subjects, to get a fair comparison in respect to traffic situation, order and learning effects. A further balance between low and much MP3 experience and between day times was included.

After the preparation with the setting-up of all devices the Subject was successively introduced into car functions, the MP3 and the interaction systems, including several video clips.

Within the two test runs the experimenter gave the tasks at the specific marks on the course. The Subject signalized the finishing of a task. If a task was not completed within the given segment, it was broken off at the corresponding mark and the Subject was asked for his mental load.

5.3 Objective Results

The SAMMIE evaluation of the Final In-Car Showcase revealed a number of results about the use and usefulness of different variants of the SAMMIE dialog system for the MP3 domain during real driving. Detailed dialogue and driving performance data as well as subjective evaluation data about speech input and output, manual iDrive input and about the display were collected. The main concern was the usability and usefulness of multimodal interaction.

Basically, the multimodal combination of speech and manual input was extensively used. The users changed in about 30 – 60% of tasks from speech to manual operation and in about 15 – 30% from manual to speech operation. The main reason for the first result are system errors or dialogue deadlocks, where the user does not succeed to solve a task. The main reason for the second result are functions, where the user does not find the correct item in a list or does not recall the right manual action.

At the beginning of a task, there was a very clear preference for speech input with all systems. For the first action SAMMIE speech input was used five times more frequently than iDrive (respectively three times for C&C speech input). The reasons were the felt potentials of speech input like low distraction, easy operation, comfort etc. Another reason could be the novelty of speech input.

With ongoing interactions during a task processing there was a clear reduction in speech preference. The rejections and false reactions of the systems during speech interaction led to changes to iDrive mode, where the Subjects were sure to get the tasks done. Sometimes, a long cumbersome speech interaction was followed by a short successful iDrive interaction.

For the tasks in the SAMMIE mode there was still a considerable preference for speech input even during the ongoing task performance. Speech input was exclusively used in almost 60% of the tasks. For the tasks in C&C mode, however, there was a balance between the preferred modalities during the ongoing interactions.

Within free interaction periods, however, iDrive was used relatively often, more frequently than in the mandatory tasks. This can be a hint, that the experimental situation affected the modality choice.

MP3 experienced Subjects tend to use speech more than the less experienced Subjects and vice versa for iDrive. This younger group took more advantage of the natural speech interaction mode. By this, they achieved a higher task completion rate (TCR) with SAMMIE than with C&C. The older group with less MP3 experience relied more on the well known manual operation with a direct connection between input device and display.

With SAMMIE there was a similar behaviour in relation to modality choice as compared to the TALK baseline study, even if there was a tendency speech input to be used and preferred somewhat more frequently.

The TCR results of the present study were on a level of about 80%. This has to be interpreted as a general high level, considering the partly tight time conditions. The tasks with SAMMIE were completed somewhat (but not significantly) more frequently than the tasks with C&C.

The SAMMIE TCR was 6% above the baseline TCR. Considering the possibility of 5 attempts within the baseline study as compared to usually less attempts that were possible within the course segments of the present study, this is a clear advantage of the SAMMIE system over the baseline system. Actually, many tasks were completed rather quickly, often with the minimal number of turns. Without the helps of the experimenter, however, a lower TCR would have been yielded. The helps of the experimenter concerned the repetition of the parameters (≈10% of the tasks) and more substantial helps (explanation of a task, loudness, etc.; ≈10% of the tasks).

The reason for not completed tasks often was a combination of understanding, dialogue and system problems, particularly by Subjects with less MP3 experience. Experienced Subjects achieved a higher TCR with SAMMIE than with C&C. Those Subjects relied more on speech input and accomplished tasks more frequently with fewer turns and somewhat faster.

With the SAMMIE system 4,9 turns and with the C&C system 5,4 turns were necessary on the average to complete a task, the difference being significant. Considering the complexity of most of the tasks, this seems to be an acceptable level.

With the SAMMIE system, however, there were not more than 0,5 turns less than with the C&C system. This is also due to the fact that subjects frequently did not use the direct and shortest dialogue path. In addition the number of necessary iDrive actions are independent from the respective system.

There was a tremendous difference of number of turns between the tasks. Much more turns were necessary to perform tasks with more parameters or/and where the system performance was lower than else. In all tasks more turns occurred than the minimum number necessary to fulfill the task, which was very pronounced with the SAMMIE system.

The average task duration with SAMMIE and C&C took about 40 – 50 s. The minimal task durations were about 10 s – 12 s. The comparable tasks in the baseline study, however, took clearly longer.

For both, number of turns and task duration, there were no very prominent differences of the results with regard to MP3 experience. But MP3 experienced Subjects were somewhat faster with SAMMIE according to their fewer turns. A general impression was, that the task duration was not a critical factor in cases when task processing progressed.

There were as many false reactions with the SAMMIE as with the baseline system but more than with the C&C system. On the average nearly each second task was affected by a false reaction of the system, which irritated the user usually more than a rejection.

There was about one rejection / task with the SAMMIE system, which was fewer as compared to the C&C and baseline system. The rejections correlated with the number of turns, i.e. more rejections corresponded to more turns.

The driving quality was measured by recording the driving errors online during the runs and by scoring the overall driving quality and normalizing it to one minute. There was no pronounced difference of the mean number of driving errors between systems. With some Subjects there were not more than occasional driving errors, while others crossed the lane boundaries continuously during task processing.

Lane departures and low speeds were the most frequent driving errors. More than one lane departure error per minute and about 0,7 speed too low errors seem to be relatively high and can be attributed to the visual distraction when observing the display. The experimental car was relatively often overtaken, even on the two-lanes roads. No definite statement, however, can be made about the effect of multimodal operation on driving safety in general. (For that a reference trial without any interaction tasks would be necessary, including additional measurements, e.g. of the eye movements.)

The driving quality scores were calculated by averaging those of the experimenter and supervisor. This subjectively judged driving quality of the Subjects was nearly equal for both systems, which confirms the objective driving quality results.

As could be observed, some Subjects drove very cautiously and relatively slowly during the complete session, more or less independent from system and tasks / no tasks. They wanted to perform well and did not “play” with the MP3 system and the car. Often they relied somewhat more on manual input by iDrive.

Some other Subjects (mostly the younger ones) drove in a superior style, played with the MP3 system and the car and operated often with speech input. Those individual differences affected the driving quality more than the respective interactive system.

The mental load was on a generally low level of about two (scale 1 – 5). There was no difference of mental load between systems. Higher scores resulted from operating the MP3 system within a demanding traffic situation and dialogue or speech recognition problems. The processing of tasks with a good progress and without serious driving or operation problems were generally not considered to be demanding.

With the SAMMIE system the thinking about the formulation or reformulation after rejections was felt to be straining by many Subjects. With the C&C system the Subject was more bound to the menu and had to do more turns. These factors seem to be more or less equivalent as to the subjectively felt mental load.

5.4 Subjective Results

The Subjects filled out intermediate questionnaires after both runs and a final questionnaire at home. For reasons of comparison most of the questions were identical to the final baseline study questionnaire and included mainly 6-point rating scales.

With the present systems by far most of the Subjects tended to a positive rating (Summarized and normalized scores of general impression: SAMMIE: 75%, C&C: 70%). With the baseline system there had been a lower rating (61%). I.e., there was a clear improvement concerning the subjective overall impression from the baseline to the SAMMIE systems, the more so as the present systems were judged to be easier to use (≈ 75%) than the baseline system (65%).

SAMMIE (65%) was assessed to be less distracting than C&C (59%) and much less distracting as compared to the baseline system (46%). But a certain distracting effect was felt by most of the Subjects.

A markedly higher comfort was felt with SAMMIE system (82%) as compared to the C&C system (69%). This distinct vote for SAMMIE as to comfort should be attributed to the Subjects´ experiences of one-input-tasks with SAMMIE speech input.

The decision for a modality and the change between modalities was easy for most of the Subjects (about 80 – 85%). This is an important result for the concept of multimodality, since a change between modalities at pleasure is easily possible.

The information output was not fully accepted as to liking (≈ 70%), support (≈ 50 - 55%), information distribution (≈ 65-70%) and assistance (≈ 65%).

The speech output was assessed to be more or less good (≈ 65 – 70%), sufficiently extensive, with relatively good formulation (≈ 75%) and very good quality (≈ 90%). The judgment of speech output was better than in the baseline study as to quality, extent, formulation, but not as to content and assistance. These aspects were highly appreciated in the baseline study.

The display was relatively well judged. This holds true for the assistance (≈ 80%), contents (≈ 70 – 80%), design (≈ 75 – 80%) and extent. Here, the SAMMIE display was mostly better judged than the baseline display (difference ≈ 7 - 15%), apart from the extent.

Concerning the dialogue there was a tendency to a positive judgment, too. SAMMIE was generally better judged than C&C. The best scores got the statement concerning the understanding of what the system said (≈ 90%). Relatively bad judgments referred to the statements, that it was easy to get the information which the user wanted and that the system worked as expected (≈ 55 – 65%). Actually, the Subjects were relatively often disoriented about the present system state.

The Subjects who participated already in the baseline study often stated spontaneously an increased performance of the present systems as compared to the baseline system. This concerned particularly the recognition performance and speed of the systems. Recommendations for further improvements concerned the extent of speech output and display, the selection of items in the lists and the position of the iDrive button.

Concerning the preference of a system in the long run, there was a slight preference of the C&C system in the final questionnaire (SAMMIE: 48%, C&C: 52%). In the intermediate questionnaire the difference was even more pronounced (SAMMIE: 45%, C&C: 60%). This is an unexpected result, because the C&C system was meant as a reference system for the SAMMIE system. It can presumably be attributed to the better system performance of the C&C system concerning speech recognition. Possibly, the better orientation along the menu with C&C is another reason for it. A change to the iDrive operation was easier with C&C, since the Subjects were always up-to-date with the display.

Concerning the advantages of the natural speech input with SAMMIE as compared to the manual iDrive inputs the safety aspects dominated as in the baseline system (´no averting glances´). The possibility to formulate freely and the new technology were pronounced much more frequently than in the baseline study.

Concerning the disadvantages of the natural speech input ´Looking for formulations´ was pronounced nearly as often as in the baseline study. That can be interpreted either by a missing acceptance of the still restricted formulation freedom or by the instruction to formulate in whole sentences.

Concerning the advantages of manual input the option ´correct system reaction´ was pronounced most frequently, even more frequently than in the baseline study. The main disadvantages were the safety aspects like ´eyes off road´, ´hands off steering wheel´ and ´searching by hand´.

The subjectively felt most important advantage of the multimodal input was avoiding the problems of the other modality. The ´free choice of the operation mode´ was another important argument. I.e. one main motivation for SAMMIE – the free option of input modality – was felt positively by a considerable part of the Subjects.

Compared to the advantages, there were much less disadvantages of the multimodal input pronounced. The main disadvantage was the uncertainty as to which task was feasible by which input device, which was more severe in the baseline study. The need for a choice between input modalities was no longer pronounced by anyone.

Being asked, which functions the Subjects would like to use in the car by the multimodal interaction, including the natural speech system SAMMIE there was partly a different order as compared to the baseline study. Now most of the Subjects pronounced the more advanced functions like desk diary, navigation and internet, while in the baseline study infotainment functions were represented more frequently.

Besides the runs with the Full (Adaptive) SAMMIE and C&C system six example videos were presented at the end of the session contrasting the Adaptive and the Non-Adaptive variants of the SAMMIE system. After each example a corresponding question related to features of the adaptive presentation strategy was asked to the subject.

While all other features of the Full SAMMIE were judged positively, the usefulness of the personal differentiated addressing was scored rather negatively (37%). The feature of a differentiation between a visual and an acoustical presentation of items (75%), presentation of albums with artists (78%) and the implicit confirmation (85%) were judged positively.

The extended user guidance was basically regarded as positive (70%). The step-by-step guidance was not totally accepted, presumably because of the somewhat lengthy dialogue. The feature of an adaptation to the user’s vocabulary was judged very diversely (57%).

6 OUTLOOK

The field test with different variants of the final In-Car Showcase SAMMIE revealed an extensive use of the multimodal interaction. Though some Subjects tended to use the systems exclusively by speech or manually in some tasks, all Subjects changed between modalities particularly when problems arose. The multimodal systems allowed a faster and more efficient interaction with the MP3 system as compared to the baseline system and it was clearly more accepted. It offered some kind of freedom, so that even playing with the system either verbally or manually could be observed.

The multimodality, however, often served as a chance to avoid the respective other modality. Changes from speech to manual input often occurred when system errors occurred.

There was a considerable progress of the SAMMIE system relating to the baseline system. This concerns most of the objective and subjective results. The most obvious improvements apply to the speed, the TCR and the display.

It was very striking, that the C&C system was somewhat more preferred than the Full SAMMIE system. This has to do with the system performance, the tight connection of input to output and the possibility to enter single commands, i.e. to avoid looking for a formulation in a sentence. A future system featuring natural language interaction should also allow for a C&C like interaction.

Natural speech input seems to be coupled to a quite different inner model of the user in respect to formulating all wanted functions and parameters within one or few sentences. A pure acoustical dialogue would be possible. In cases of lists or system problems, however, falling back to the display is necessary and represents a rupture within the model. With a command based system speech input goes along with the display presentations and allows an easy change to the manual input.

Even if not verified within this experiment, some kind of distraction from driving can be assumed. A possible distraction can affect the lane keeping and speed. While speech input per se is not very prone to distraction, the coupled visual activities towards the display does. Nevertheless, a mere speech system without display would not be accepted.

The experimental conditions affected the results particularly when giving predefined tasks. Most of the Subjects felt some time pressure and acted differently than else. There may have been even the artefact to comply with assumed expectations of the experimenter, e.g. ´Speech input is a relatively new interaction system. Prefer it.´

The free interactions showed, that the Subjects behaved partly differently when choosing their own music and interacting with the system in their own – possibly more known - way. In free interactions the iDrive was used as often as speech input. This is a hint that the familiar manual input is still a well accepted input modality, at least without a considerable familiarity with speech input. Long term studies could show some changes in the interaction behaviour and the choice of modality.

Hypotheses: Most of the hypotheses missed significance. But the tendencies confirmed a part of them:

Hypothesis:	**Tendency**	**Significance**
1. Users prefer speech input more with the SAMMIE system than with the C&C system	yes	yes
2. Users with much MP3 experience tend to manual operation	contrary	no

3. Users with much MP3 experience achieve a higher operation efficiency, particularly with a lower number of turns	yes	no
4. Users get a higher Task Completion Rate with SAMMIE than with C&C	yes	no
5. Users are faster with the SAMMIE system than with the C&C system	yes	no
6. The number of turns per task is higher with C&C than with SAMMIE.	yes	yes
7. SAMMIE needs less iDrive actions.	no	no
8. The number of system errors with SAMMIE is only marginally higher than with C&C.	False reactions clearly higher, rejections lower	no
9. SAMMIE does less distract from driving than C&C.	no	no
10. The SAMMIE system leads to a higher user acceptance than the C&C system.	contrary	no
11. Users can assess well what the system has understood.	yes	--

On the basis of the objective and subjective results as well as on the basis of observations and informal discussions following recommendations can be given:

Generally:

- Pursue the concept of multimodality, i.e. fully parallel input modes with a free choice of modality at any time.
- Keep most of the features, e.g. free access to any menu levels by speech, back function etc.
- Optimise all acoustic signals in respect to a clear differentiability between microphone opening and closing.

Speech input:

- Further improve speech recognition and language understanding performance
- Either: Improve the grammar by e.g. extending the coverage. Do not claim a natural language system, if a lot of common German expressions are not covered.
- Or: reduce the vocabulary/grammar to a very limited one and provide a user manual.
- Make the automatic opening of the microphone configurable. In favour of a consequent user-driven concept for each single speech input an activation of the PTT-button should be provided.
- Allow a verification dialogue for low confidence understandings.

iDrive:

- Reposition the iDrive device in the centre console more to the front.
- Mark the possible actuations on the iDrive device.

Speech output:

- Keep the concept of barge-in (by PTT-button). Possibly extend barge-in concept for modality changes.
- Reduce amount and length of speech output.
- Do not read lists when not explicitly requested by the user.
- Provide a button to switch off speech output completely, so that the user is free to have speech output or not.
- Do not announce very obvious system activities, e.g. “Die ersten sieben werden auf dem Bildschirm dargestellt”. A short tone is often enough for signalising a display output.
- Do not refer to the incomplete help system.

Optical display:

- Keep the display basically as it is.
- Leave out any unnecessary information, particularly the picture of the albums and increase instead the size of the actual artist, album and song or playlist.
- Increase the graphics resolution.
- Position the display centrally, i.e. at or above the dashboard.
- Signalise the pause status of the MP3 player optically.

7 References

[1] Tilman Becker, Nate Blaylock, Ciprian Gerstenberger, Andreas Korthauer, Nadine Perera, Peter Poller, Jan Schehl, Frank Steffens, Rosmary Stegmann, Jochen Steigner: “In-Car Showcase Based on TALK Libraries”, Deliverable D5.3, TALK project, 2006.

[2] Andreas Korthauer, Holger Banski, Frank Steffens, Hartmut Mutschler, Peter Poller: ”Evaluation of the Baseline System”, Deliverable D6.3, TALK project, 2006.

[3] AttrakDiff website: http://www.attrakdiff.de

[4] M.A. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Owen Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff, D. Stallard, “DARPA Communicator: Cross-System Results for The 2001 Evaluation”, ICSLP-2002:Inter. Conf. on Spoken Language Processing, vol. 1, pp 269-272, Denver, CO USA, Sept. 2002.

8 Attachments

8.1 Tasks

1. Aufgabe:

SAMMIE 1	Kommando
multimodal lösen.	mit Spracheingabe anfangen.

Bitte finden Sie heraus, welche Alben im System vorhanden sind. -
Sie wollen also wissen, welche Alben es gibt.

2. Aufgabe:

SAMMIE 1	Kommando
multimodal lösen.	mit Spracheingabe anfangen.

Lassen Sie sich bitte das Lied ´Der Weg´ von Herbert Grönemeyer auf dem Album Mensch abspielen. - Sie möchten also das Lied ´Der Weg´ von Herbert Grönemeyer auf dem Album Mensch hören.

3. Aufgabe:

SAMMIE 1	Kommando
multimodal lösen.	mit Spracheingabe anfangen.

Finden Sie nun bitte heraus, welche Lieder in der Playliste „Pur Klassiker“ vorhanden sind. - Sie wollen also wissen, welche Titel die Wiedergabeliste „Pur Klassiker“ enthält.

4. Aufgabe:

SAMMIE 1	Kommando
multimodal lösen.	multimodal lösen.

Bitte gehen Sie durch die Alben, suchen Sie das Album ´Live´ von Pur bis es angezeigt wird. und lassen es abspielen. - Also das Album ´Live´ von Pur, indem Sie die Liste durchgehen, und anhören

5. Aufgabe:

SAMMIE 1	Kommando
multimodal lösen.	multimodal lösen.

Suchen Sie ein Swing-Stück von Michael Buble und lassen es abspielen. - Sie wollen also ein Stück der Musikrichtung Swing von Michael Buble und es anhören.

Selbstständiger Dialog:

SAMMIE 1	Kommando

Sie können nun das System selbstständig nach eigenem Wunsch bedienen. -
Bitte probieren Sie Funktionen beliebig aus.

6. Aufgabe:

SAMMIE 1	Kommando
multimodal lösen.	Entfällt !

Fügen Sie bitte das Lied ´99 Luftballons´ im Album Leuchtturm von ´Nena´ zur Playliste ´Autofahrt´ hinzu. - Also das Lied ´99 Luftballons´ im Album Leuchtturm von ´Nena´ in die Wiedergabeliste aufnehmen.

Selbstständiger Dialog:

SAMMIE 1	Kommando

Sie können nun das System selbstständig nach eigenem Wunsch bedienen. - Bitte probieren Sie Funktionen beliebig aus.

7. Aufgabe:

SAMMIE 1	Kommando
multimodal lösen.	Entfällt !.

Finden Sie heraus, ob es das Lied ´Yesterday` auf dem Album ´Number One Hits ´ von den Beatles gibt. Sagen Sie es mir und falls ja, spielen Sie es ab. – Ist das Stück ´ Yesterday ` von den Beatles auf dem Album ´ Number One Hits ´, eventuell anhören?

8. Aufgabe:

SAMMIE 1	Kommando
Mit Spracheingabe lösen.	Entfällt !

Bitte erstellen Sie eine neue Playliste. – Sie wollen also eine neue Wiedergabeliste anlegen.

9. Aufgabe:

SAMMIE 1	Kommando
multimodal lösen.	multimodal lösen.

Von welchem Künstler ist ´Romeo und Julia´ auf der Playliste ´Cool Hits´. -
Wer ist der Interpret von ´Romeo und Julia´ auf der Wiedergabeliste ´Cool Hits´.

10. Aufgabe:

SAMMIE 1	Kommando
multimodal lösen.	multimodal lösen.

Bitte wählen Sie aus der Musikrichtung ´ Rock´ ein Lied Nach Ihrem Geschmack und spielen es ab. – Finden Sie also ein beliebiges Stück der Musikrichtung ´Rock´ und hören Sie es sich an.

8.2 Introduction to the experiment

Vor-Ort-Erklärung des Versuchs:

Sehr geehrter(e) Teilnehmer(in),

vielen Dank, dass Sie gekommen sind. Wir gehen davon aus, dass Sie die Vorab-Erklärung gelesen haben. (Falls nicht, tun Sie dies bitte jetzt.). Hier sind nun weitere Einzelheiten zum heutigen Versuch:

Sie testen heute das SAMMIE-Dialogsystem für MP3-Player im Fahrzeug in verschiedenen Varianten. Dies sind die Varianten, die Sie bei den zwei Versuchsfahrten benutzen, also:

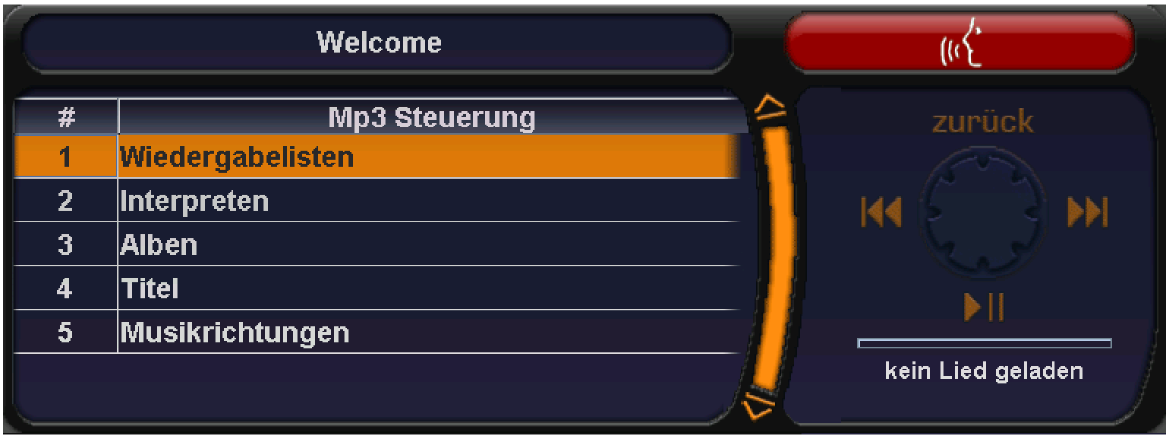
A. SAMMIE-System: Spracheingabe mit natürlicher Sprache oder manuell mit iDrive-Knopf

B. Kommando-System: Spracheingabe mit einzelnen Wörtern oder manuell mit iDrive-Knopf

Außerdem werden wir Ihnen abschließend noch eine weitere Variante vorführen und Sie um eine Beurteilung bitten:

C. SAMMIE-System Variante: Im wesentlichen ähnlich zu A, allerdings mit einigen Besonderheiten

Es kann z.B. folgendes Display dargestellt werden:



Die Aufgaben werden wahlweise mit natürlicher Spracheingabe oder manuell mit einem Bedienelement (iDrive) gelöst. Dies nennen wir „multimodal“.

Am Lenkrad befinden sich mehrere Tasten. Für den Versuch sind lediglich die beiden inneren markierten Tasten auf der rechten Seite von Interesse:

Obere innere Taste auf der rechten Seite: Öffnen des Mikrofons für die Spracheingabe. Dabei wechselt die Mikrofonanzeige auf dem Display von Rot nach Grün und es ertönt ein bestimmtes Gong-Signal. Das Mikrofon schließt nach jeder Spracheingabe automatisch mit einem anderen Gong-Signal.

Untere innere Taste auf der rechten Seite: Schließen des Mikrofons bzw. Beenden der Sprachausgabe. Diese Taste können Sie bei Bedarf benutzen.

Der große iDrive-Knopf befindet sich rechts von Ihrem Sitz auf der Mittelkonsole.

Drehen:	Markierung eines Elements auf der angezeigten bzw. vorgelesenen Liste
Kurzes Drücken:	Auswahl des Elements, z.B. Spielen eines Liedes
Langes Drücken:	Aufnahme des ausgewählten Liedes in die Playliste
Nach rechts/links verschieben:	Auswahl des nächsten/vorigen Liedes
Verschieben nach unten:	Stopp/Pause des Liedes oder Albums
Verschieben nach oben:	Zurückgehen zur vorigen Menüebene, d.h. zur vorigen Darstellung

Außerdem befindet sich daneben eine Taste „Hauptmenü“, mit der Sie in die oberste Ebene des Menüs kommen. (Diese Ebene wird auch der Ausgangspunkt vor jeder Aufgabe sein.) Wie Sie es von anderen MP3-Systemen kennen, gibt es hier unter anderem den Menüpunkt „Musikrichtungen“, z.B. „Pop, Rock, Deutsch Rock, Jazz etc.“

Sie werden bei der folgenden Versuchsfahrt das Dialogsystem verwenden, das Ihnen der Versuchsleiter nun sagt:

SAMMIE-System	**Kommando-System**
Sie bedienen das SAMMIE-System mit Spracheingabe wahlweise in natürlicher Sprache oder manuell über den iDrive-Knopf.	Sie bedienen das Kommando-System mit Spracheingabe wahlweise in Kommando-Sprache oder manuell über den iDrive-Knopf
Wenn Sie für eine Aufgabe oder einen Teil einer Aufgabe die Spracheingabe wählen, dann sprechen Sie im Prinzip so, als wenn Sie mit einer Person sprechen würden, also in natürlicher Sprache. Sie sollten möglichst in einfachen, ganzen Sätzen sprechen. Sie führen den Dialog etwa so wie bei der zwischenmenschlichen Kommunikation, also im Wechselgespräch mit dem System.	Wenn Sie für eine Aufgabe oder einen Teil einer Aufgabe die Spracheingabe wählen, dann sprechen Sie im Prinzip einzeln die Wörter, die Sie auf dem Display sehen oder die unten erklärten Steuerwörter „Weiter“ etc.“ – also in Kommandoform. Sie können auch Interpreten, Alben, Titel und Playlisten nennen, die sich nicht sichtbar weiter unten oder oben in der Liste befinden.
Bei Verständnisschwierigkeiten hilft evtl. eine Neuformulierung der Spracheingabe. Außerdem können Sie jederzeit – also auch während einer Aufgabe - auf die manuelle Eingabe übergehen und andersherum!	Bei Verständnisschwierigkeiten hilft evtl. eine erneute Spracheingabe. Außerdem können Sie jederzeit – also auch während einer Aufgabe - auf die manuelle Eingabe übergehen und andersherum!

Für beide Systeme gelten die folgenden Steuerbefehle:

Mit „Weiter“ oder einem ähnlichen Befehl blättert das System auf den nächsten / vorigen Teil einer Liste, ähnlich dem mehrfachen Drehen des iDrive-Knopfes.

Mit „Zurück“ oder einem ähnlichen Befehl geht das System in eine der vorigen Darstellungen zurück. Dies entspricht oft dem Hochschieben des iDrive-Knopfes.

Mit „Hauptmenü“ oder einem ähnlichen Befehl geht das System in das Hauptmenü, analog zur Betätigung der Taste ´Hauptmenü´.

Sie können den Dialog abbrechen und mit dem Hauptmenü erneut beginnen. Wenn Sie die weitere Bearbeitung einer Aufgabe für ganz aussichtslos halten, können Sie ebenfalls abbrechen. Wenn wir während der Bearbeitung der Aufgaben an bestimmten Marken auf der Strecke angekommen sind, dann fordert Sie der Versuchsleiter auf, die Aufgabe abzubrechen.

Falls eine Aufgabe nicht zu Ende geführt werden kann, ist das kein Misserfolg Ihrerseits, sondern für uns ein Erkenntnisgewinn. Fahren Sie einfach mit den weiteren Anweisungen fort.

Für alle Systeme gilt: Sie können nur sprechen, wenn Sie die Mikrofon-Taste vorher gedrückt haben und die Mikrofonanzeige auf Grün geschaltet wurde. Sie können der Sprachausgabe jederzeit mit der Mikrofon-Taste „ins Wort fallen“ und danach selbst einsprechen.

Noch einmal: Bis auf wenige Ausnahmen besteht grundsätzlich freie Wahl zwischen Spracheingabe und manueller Eingabe (iDrive), auch während der Bearbeitung einer Aufgabe = „multimodale Eingabe“.

Die Ausgabe des Systems erfolgt optisch auf Display und akustisch als Sprachausgabe.

Von besonderem Interesse ist für uns Ihre Benutzung und Bewertung des Dialogsystem, incl. der Systemausgaben. Falls Sie bei der Fahrsimulation im November 2005 beteiligt waren, ist außerdem Ihr Vergleich mit früherem TALK-System der Fahrsimulation interessant.

MP3-Aufgaben:

- Lieder anhören
- Einholen von Informationen
- Arbeit mit Playlisten (=Wiedergabelisten).

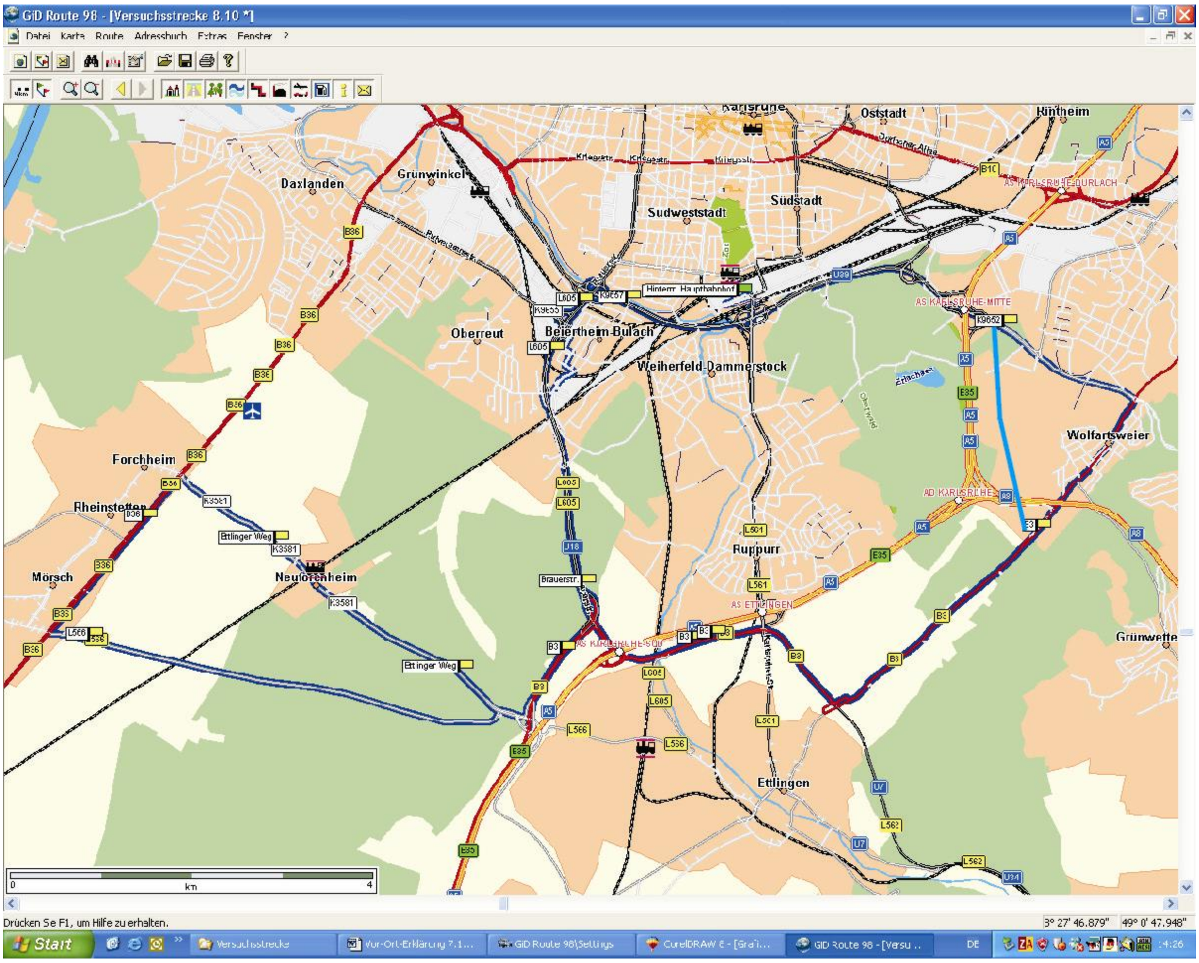
Jede Aufgabe wird zu bestimmten Zeitpunkten zweimal hintereinander angesagt. **Beginnen Sie mit der Bearbeitung der Aufgabe bitte erst nach der zweiten Ansage.**

Sagen Sie bitte laut oder geben Sie ein Handzeichen, wenn Sie mit der Bearbeitung der Aufgabe fertig sind. Danach können Sie auch wieder mit den anderen Fahrzeuginsassen sprechen.

Wir fragen Sie nach den Aufgaben nach Ihrer Beanspruchung, die Sie bitte auf einer Skala von 1 bis 5 ohne Zwischenstufen angeben: 1=keine Beanspruchung, 5=große Beanspruchung. Dabei ist die gesamte Beanspruchung gemeint, also das Fahren und Bedienen.

Nach jeder Fahrt geben wir Ihnen einen kleinen Fragebogen zum sofortigen Ausfüllen.

Fahrstrecke: Südtangente Richtung Wolfartsweier → B3 Richtung Ettlingen → B3 Richtung Rastatt → Straße Richtung Mörsch → B36 → Straße von Forchheim → B3 Richtung Karlsruhe → BAB Zubringer → Südtangente Richtung Hauptbahnhof.



Das sichere Fahren hat auch bei der Bearbeitung der Aufgaben stets unbedingten Vorrang. Achten Sie dabei bitte auf die Straßenverkehrsordnung.

Zum Schluss erhalten Sie einen großen Fragebogen mit Rückumschlag und wir bitten Sie, ihn zuhause heute oder allerspätestens morgen auszufüllen.

Viel Spaß.

8.3 Intermediate questionnaire

For SAMMIE and C&C run identical, apart from the system name „SAMMIE-Bediensystem“ and „Kommando-Bediensystem“

Zwischenbefragung nach dem SAMMIE-System

Datum: ________________ Name: ____

1. Wie ist Ihr allgemeiner Eindruck vom gesamten SAMMIE-Bediensystem ?

   sehr gut □ □ □ □ □ □ sehr schlecht

2. Die Bedienung des gesamten SAMMIE-Bediensystems, also mit Sprach- und iDrive-Bedienung war für Sie ?

   sehr einfach □ □ □ □ □ □ sehr schwierig

3. Wie stark fühlten Sie sich während der Bedienung des SAMMIE-Bediensystems vom Fahren abgelenkt? ***Unterscheiden Sie dabei nicht zwischen den Eingabearten, sondern betrachten es als Gesamtsystem.***

   überhaupt nicht abgelenkt □ □ □ □ □ □ sehr abgelenkt

4. Wie sicher fühlten Sie sich bei der Bedienung des SAMMIE-Bediensystems ? ***Unterscheiden Sie dabei nicht zwischen den Eingabearten, sondern betrachten es als Gesamtsystem.***

   sehr sicher □ □ □ □ □ □ sehr unsicher

5. Welchen Komfort empfanden Sie bei der Bedienung des gesamten SAMMIE-Bediensystems?

   sehr groß □ □ □ □ □ □ sehr gering

6. Welchen Spaß hatten Sie bei der Bedienung des gesamten SAMMIE-Bediensystems ?

   sehr groß □ □ □ □ □ □ sehr gering

7. Wie leicht oder schwer fiel Ihnen die jeweilige Entscheidung für eine Eingabeart?

   sehr leicht □ □ □ □ □ □ sehr schwer

8. Wie leicht oder schwer fiel Ihnen der Wechsel zwischen Spracheingabe und Bedienteil?

   sehr leicht □ □ □ □ □ □ sehr schwer

9. Wie beurteilen Sie das Verhalten des Systems, das Mikrofon während eines Dialogs selbstständig zu öffnen?

   unterstützt sehr gut □ □ □ □ □ □ sehr verwirrend

10. War es für Sie verständlich, wann Sie sprechen konnten?

    immer □ □ □ □ □ □ sehr selten

Nächste Seite

Informationsausgaben:

Hier zunächst Fragen zur Informationsausgabe allgemein, d.h. unabhängig davon, ob sie optisch oder akustisch erfolgten.

11. Wie hat Ihnen die Systemausgabe (optisch und akustisch) gefallen?

sehr gut □ □ □ □ □ □ sehr schlecht

12. Wurden Sie bei Problemen im Dialog vom System unterstützt?

sehr gut □ □ □ □ □ □ sehr schlecht

13. Wie gut fanden Sie die Verteilung der Information zwischen Sprachausgabe und optischer Anzeige?

sehr gut □ □ □ □ □ □ sehr schlecht

Hier Fragen zu den Sprachausgaben, d.h. zu den Sprachansagen des Systems an Sie.

14. Wie hat Ihnen die Sprachausgabe gefallen?

sehr gut □ □ □ □ □ □ sehr schlecht

15. Wie hilfreich waren für Sie die Sprachausgaben?

sehr hilfreich □ □ □ □ □ □ überhaupt nicht hilfreich

16. Wie gut fanden Sie den Inhalt der Sprachausgaben?

sehr gut □ □ □ □ □ □ sehr schlecht

17. Wie beurteilen Sie den Umfang der Sprachausgaben?

nicht ausreichend □ □ □ □ □ □ zu umfangreich

18. Wie beurteilen Sie die Formulierung der Sprachausgaben?

sehr gut □ □ □ □ □ □ sehr schlecht

19. Wie beurteilen Sie die akustische Qualität der Sprachausgaben?

sehr gut □ □ □ □ □ □ sehr schlecht

20. Haben Sie noch Bemerkungen zu den Sprachausgaben*? (Falls Sie bei der Fahrsimulation dabei waren, können Sie auch mit den damaligen Sprachausgaben vergleichen)*

____

____

Hier sind nun Fragen zur Displaydarstellung, d.h. zu den Darstellungen auf dem Bildschirm.

21. Wie hat Ihnen die optische Anzeige gefallen?

sehr gut □ □ □ □ □ □ sehr schlecht

22. Wie hilfreich waren für Sie die optischen Anzeigen?

sehr hilfreich □ □ □ □ □ □ überhaupt nicht hilfreich

23. Wie gut fanden Sie den Inhalt der optischen Anzeigen?

sehr gut □ □ □ □ □ □ sehr schlecht

24. Wie gut fanden Sie die Gestaltung der optischen Anzeigen?

sehr gut □ □ □ □ □ □ sehr schlecht

25. Wie beurteilen Sie den Umfang der optischen Anzeigen?

nicht ausreichend □ □ □ □ □ □ zu umfangreich

26. Haben Sie noch Bemerkungen zu den Displaydarstellungen*? (Falls Sie bei der Fahrsimulation dabei waren, können Sie auch mit den damaligen optischen Anzeigen vergleichen)*

____

____

Bitte nehmen Sie zu den folgenden Aussagen über das System Stellung. Es handelt sich also um Aussagen, denen Sie mehr oder weniger zustimmen oder nicht zustimmen sollen.

Ab hier handelt es sich um Skalen mit 5 Optionen!

27. Es war einfach für mich zu verstehen, was das System sagte.

stimme vollkommen zu □ □ □ □ □ stimme gar nicht zu

28. Es war einfach, die Informationen zu bekommen, die ich wollte.

stimme vollkommen zu □ □ □ □ □ stimme gar nicht zu

29. Ich wusste zu jeder Zeit im Dialog, was ich sagen oder machen kann.

stimme vollkommen zu □ □ □ □ □ stimme gar nicht zu

30. Das System funktionierte in der Weise, wie ich es von ihm erwartet habe.

stimme vollkommen zu □ □ □ □ □ stimme gar nicht zu

31. Ich denke, ich würde das System zukünftig gerne nutzen.

stimme vollkommen zu □ □ □ □ □ stimme gar nicht zu

Nächste Seite

32. Haben Sie noch Bemerkungen zum SAMMIE-Bediensystems ? (Falls Sie bei der Fahrsimulation TALK dabei waren, können Sie auch das damalige TALK-System einbeziehen.)

____

____

____

Ab hier handelt es sich um Skalen mit 7 Optionen!

33. Nachfolgend finden Sie Wortpaare, mit deren Hilfe Sie die Beurteilung des soeben verwendeten Systems vornehmen können. Sie stellen jeweils extreme Gegensätze dar, wischen denen eine Abstufung möglich ist. Bitte bewerten Sie das System möglichst spontan mit Hilfe der unten angegebenen Adjektiv-Paare indem sie das zutreffende Feld mit einem Kreuz markieren. Wenn Sie der Meinung sind, ein Adjektiv-Paar nicht zuordnen zu können, kreuzen Sie bitte den Mittelpunkt der Skala an (0).

Das System war...

technisch								menschlich
Kompliziert								einfach
Unpraktisch								praktisch
Umständlich								direkt
Unberechenbar								voraussagbar
Verwirrend								übersichtlich
Widerspenstig								handhabbar
Isolierend								verbindend
Laienhaft								fachmännisch
Stillos								stilvoll
Minderwertig								wertvoll
Ausgrenzend								einbeziehend
trennt mich von Leuten								Bringt näher
nicht vorzeigbar								vorzeigbar
Konventionell								originell
Phantasielos								kreativ
Vorsichtig								mutig
Konservativ								innovativ
Lahm								fesselnd
Harmlos								herausfordernd
Herkömmlich								neuartig
Unangenehm								angenehm
hässlich								schön
unsympathisch								sympathisch
zurückweisend								einladend
schlecht								gut
abstoßend								anziehend
entmutigend								motivierend
	-3	-2	-1	0	+1	+2	+3	

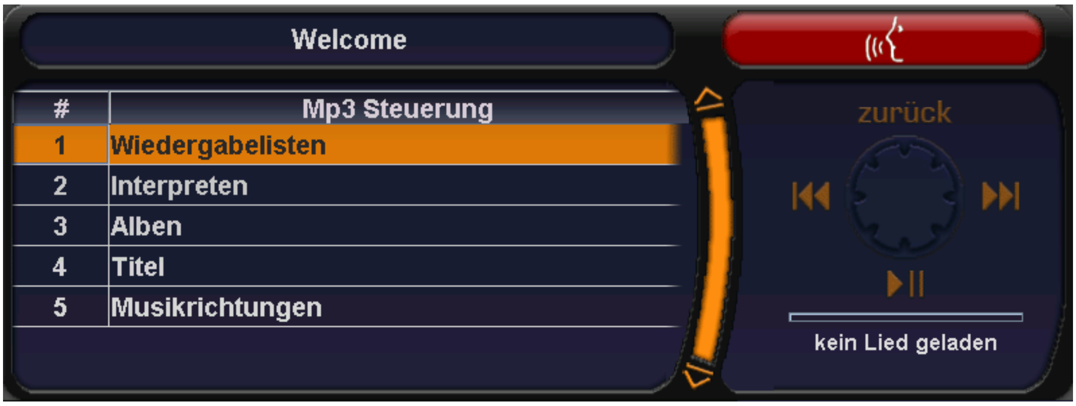
8.4 Final questionnaire

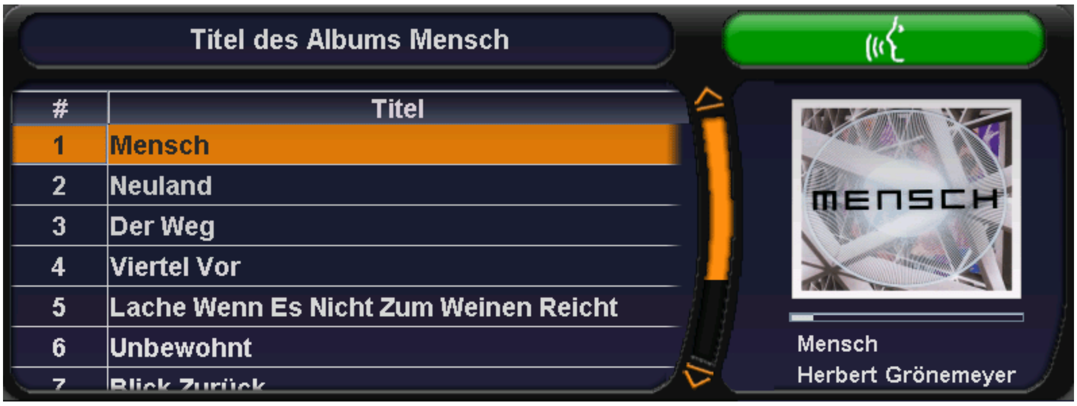
Postexperimenteller Fragebogen nach dem SAMMIE- Versuch

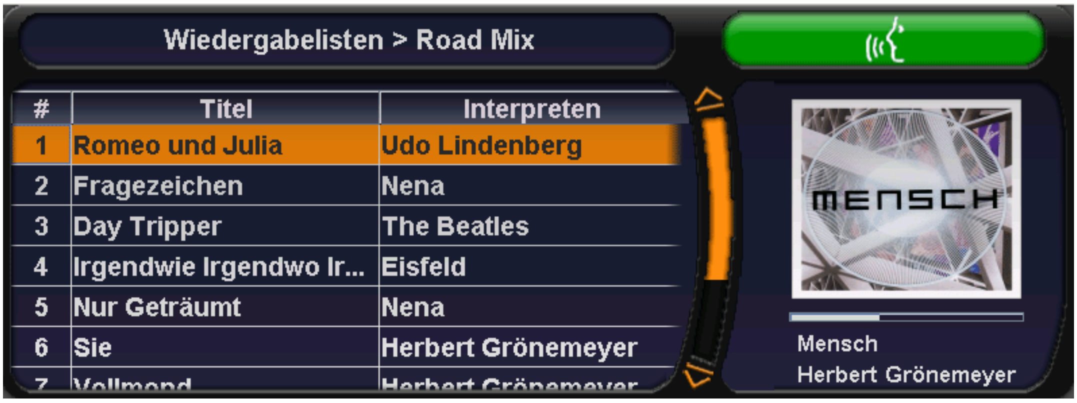
Name: ____________________________ Datum: ____________

Bitte beantworten Sie die folgenden Fragen und beurteilen Sie das multimodale SAMMIE- und Kommando-System, das Sie heute im Fahrversuch kennen gelernt haben. Wir benötigen Antworten, die genau Ihre Erfahrungen und Beurteilungen wiedergeben.

Zur Erinnerung: Sie haben beim Fahrversuch drei Systeme kennen gelernt: A) Multimodales SAMMIE-System während der Fahrt B) Kommando-System während der Fahrt C) Multimodale SAMMIE-Variante als Vorführung. Die Systeme A) und B) haben Sie möglicherweise in einer anderen Reihenfolge getestet. Hier sind einige der Bilder von A) als Erinnerungshilfe:



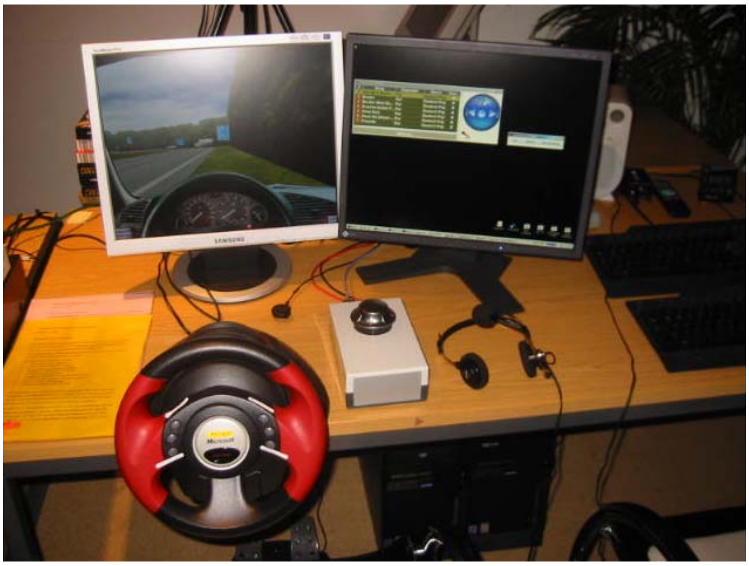




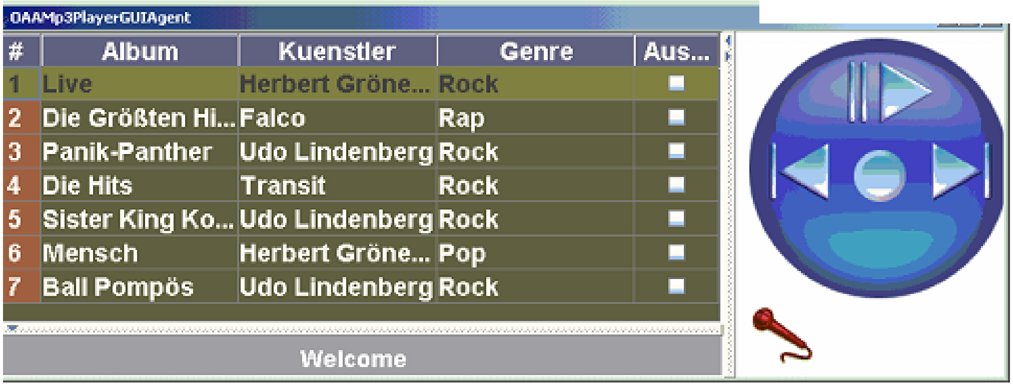
Wiedergabelisten > Road Mix

#	Titel	Interpreten
1	Romeo und Julia	Udo Lindenberg
2	Fragezeichen	Nena
3	Day Tripper	The Beatles
4	Irgendwie Irgendwo Ir...	Eisfeld
5	Nur Geträumt	Nena
6	Sie	Herbert Grönemeyer
7	Vollmond	Herbert Grönemeyer

Mensch
Herbert Grönemeyer



Falls Sie den Fahrsimulationsversuch TALK im November 2005 im BEF mitgemacht haben, bitten wir Sie an verschiedenen Stellen des Fragebogens um einen Vergleich der jetzt getesteten SAMMIE-Systeme mit dem damaligen TALK-System. Als Erinnerungshilfe sehen Sie im folgenden den Versuchsaufbau der Fahrsimulation sowie das TALK-Display.



OAAMp3PlayerGUIAgent

#	Album	Kuenstler	Genre	Aus...
1	Live	Herbert Gröne...	Rock	■
2	Die Größten Hi...	Falco	Rap	■
3	Panik-Panther	Udo Lindenberg	Rock	■
4	Die Hits	Transit	Rock	■
5	Sister King Ko...	Udo Lindenberg	Rock	■
6	Mensch	Herbert Gröne...	Pop	■
7	Ball Pompös	Udo Lindenberg	Rock	■

Welcome

Die Antwortoptionen der meisten Fragen sind mit 6 Kästchen gekennzeichnet, die z.B. von „sehr gut" bis „sehr schlecht" reichen. Bitte entscheiden Sie sich bei diesen Fragen für genau ein Kästchen, nicht mehr und nicht dazwischen ankreuzen.

Bei anderen Fragen, deren Antworten mit Kreisen gekennzeichnet sind, können Sie mehrere Antworten ankreuzen.

Bei den offenen Fragen, die mit Linien versehen sind, sind keine Antworten vorgegeben. Hier können Sie frei formulieren, aber bitte so kurz und bündig, dass der Platz ausreicht.

Vergleich der Eingabeverfahren:

1. Welche Eingabeart würden Sie mit mehr Übung als heute wohl auf Dauer verwenden? Bitte nur ein Kreuz! *Falls Sie bei der Fahrsimulation TALK dabei waren, haben Sie die Auswahl zwischen allen drei Optionen, ansonsten zwischen den oberen beiden.*

   SAMMIE-System (Versuchsfahrt) □

   Kommando-System (Versuchsfahrt) □

   TALK-System (Fahrsimulation) □

2. Wie gut waren die Systeme im Vergleich zu den jeweils anderen Systemvarianten zu bedienen? Bitte pro Zeile ein Kreuz. *Falls Sie bei der Fahrsimulation TALK dabei waren, bitte auch in der 3. Zeile ein Kreuz machen.*

   SAMMIE-System:
   viel einfacher □ □ □ □ □ □ viel schwerer

   Kommando-System:
   viel einfacher □ □ □ □ □ □ viel schwerer

   TALK-System:
   viel einfacher □ □ □ □ □ □ *viel schwerer*

Einzelne Aspekte der Eingabeverfahren:

Spracheingabe:

Bitte denken Sie bei der Beantwortung der folgenden drei Fragen an die Versuchsfahrten mit dem natürlich-sprachlichen SAMMIE-System sowie an die Aufgaben, wo Sie (vor allem) die Spracheingabe benutzt haben.

3. Stellten ein oder mehrere Aspekte der folgenden Liste für Sie persönlich Vorteile der Sprachbedienung mit dem natürlich-sprachlichen SAMMIE-System im Vergleich zur manuellen Eingabe dar?

   ○ Keine Blickabwendung vom Verkehr

   ○ Höhere Konzentration mit den Gedanken auf den Verkehr

   ○ Relativ freie Formulierung der Fragen

   ○ Moderne Technik

   ○ Sonstiges____

4. Stellten ein oder mehrere Aspekte der folgenden Liste für Sie persönlich Nachteile der Sprachbedienung mit dem natürlich-sprachlichen SAMMIE-System im Vergleich zur manuellen Eingabe dar?

   ○ Fehlerkennung von Spracheingaben

   ○ Notwendige Suche nach einer passenden Formulierung

- ○ Länger dauernde Eingaben
- ○ Gegenseitige Störung von Spracheingaben und menschlicher Kommunikation / Geräuschen
- ○ Sonstiges____

5. Wie beurteilen Sie die Möglichkeit, mit dem natürlich-sprachlichen SAMMIE-System relativ frei zu formulieren im Vergleich zur Verwendung von Kommandoworten?

viel besser □ □ □ □ □ □ viel schlechter

Manuelle Eingabe:

Bitte denken Sie bei der Beantwortung der folgenden beiden Fragen an die Aufgaben, wo Sie in einer der beiden Fahrten das manuelle Bedienteil benutzt haben.

6. Stellten ein oder mehrere Aspekte der folgenden Liste für Sie persönlich Vorteile der manuellen Eingabe im Vergleich zur Sprachbedienung dar?
   - ○ Betätigung mit der Hand („Ich kann etwas greifen")
   - ○ Korrekte Reaktion des Systems („Es macht genau das, was ich will")
   - ○ Auswählen aus einer Liste (Drehen des Knopfes + Drücken)
   - ○ Sonstiges____

7. Stellten ein oder mehrere Aspekte der folgenden Liste für Sie persönlich Nachteile der manuellen Eingabe im Vergleich zur Sprachbedienung dar?
   - ○ Betätigung mit der Hand („Ich muss die Hand vom Lenkrad wegnehmen")
   - ○ Suchen des manuellen Bedienteils mit der Hand
   - ○ Suchen des Bedienteils mit den Augen
   - ○ Blickabwendung vom Verkehr
   - ○ Zuordnung der einzelnen Betätigungsarten (Drehen, Schieben, Drücken) zu den Funktionen (Cursor verschieben, Wiedergabefunktionen, auswählen etc.)
   - ○ Sonstiges____

Multimodale Bedienung:

Bitte denken Sie bei der Beantwortung der folgenden Fragen an die Fahrt mit multimodaler Eingabe in der natürlich-sprachlichen SAMMIE-Version.

8. Stellten ein oder mehrere Aspekte der folgenden Liste für Sie persönlich Vorteile der multimodalen Bedienung in der natürlich-sprachlichen SAMMIE-Version dar, d.h. Vorteile der freien Auswahl zwischen sprachlicher und manueller Eingabe?
   - ○ Freie Wahl des Eingabemediums nach eigenem Geschmack
   - ○ Anpassung des Eingabemediums an die Aufgabe

○ Anpassung des Eingabemediums an die Fahrsituation

○ Vermeidung von Problemen des einen Mediums durch Wahl des anderen

○ Abwechslung

○ Sonstiges____

9. Stellten ein oder mehrere Aspekte der folgenden Liste für Sie persönlich Nachteile der multimodalen Eingabe in der natürlich-sprachlichen SAMMIE-Version dar, d.h. Nachteile der freien Auswahl zwischen sprachlicher und manueller Eingabe dar?

○ Konzeptionelles Umdenken zwischen den Eingabemedien erforderlich („Bei der Spracheingabe muss ich formulieren, bei der manuellen Eingabe muss ich auf eine bestimmte Art greifen“)

○ Unsicherheit, ob Aufgabe mit dem gewünschten Eingabemedium tatsächlich durchführbar ist

○ Entscheidung für ein Eingabemedium, da beide Eingabearten möglich sind

○ Ungewohnte Wahl zwischen zwei Eingabemedien

○ Sonstiges____

10. Welche Gründe hatten Sie dafür, bei der natürlich-sprachlichen SAMMIE-Version die Spracheingabe zu nutzen in Fällen, bei denen Sie auch manuell mit dem Bedienteil hätten eingeben können?

____

____

11. Welche Gründe hatten Sie bei der natürlich-sprachlichen SAMMIE-Version dafür, das manuelle Bedienteil iDrive zu nutzen, da Sie ja auch per Spracheingabe hätten eingeben können?

____

____

12. Welche weiteren Funktionen würden Sie gerne mit dem multimodalen natürlich-sprachlichen SAMMIE-System im Fahrzeug nutzen?

○ Navigation/dynamische Zielführung
○ SMS
○ Radio
○ Kassette, CD Spieler
○ Internetzugang
○ Sonstige ____

○ Restaurant-, Hotelreservierung,
○ Terminkalender
○ Telefon
○ Verkehrsinformation

13. Haben Sie noch Bemerkungen zur multimodalen Bedienung mit dem natürlich-sprachlichen SAMMIE-System während des Fahrens, d.h. zur Bedienung mit beliebigem Wechsel von sprachlicher und manueller Eingabe? (*Falls Sie bei der Fahrsimulation dabei waren, können Sie auch mit der damaligen kombinierten Eingabe vergleichen)*

____

____

14. Bitte überdenken Sie jetzt noch einmal den gesamten Versuch. Wenn es noch Aspekte aller Bediensysteme gibt, die Ihnen aufgefallen sind, zu denen Sie aber noch nicht befragt wurden, dann erläutern und beurteilen Sie sie bitte hier. Also ergänzende Bemerkungen zum

SAMMIE-System im Stand, zum SAMMIE-System bei der Fahrt und zum Kommandowort-System (bei der Fahrt):

15. Welche Verbesserungsvorschläge haben Sie für die Weiterentwicklung des multimodalen natürlich-sprachlichen SAMMIE-Systems?

Falls Sie zufällig noch andere Personen kennen, die an diesem Versuch teilnehmen, ist es wichtig, dass Sie keine Informationen und persönlichen Beurteilungen austauschen, bis Sie alle den Versuch unabhängig voneinander durchgeführt haben.

Vielen Dank für Ihre Teilnahme!

8.5 NA SAMMIE questionnaire

Zwischenbefragung für SAMMIE-Variante im Stand

Name: __________________________ Datum: ___________

Bitte beantworten Sie die folgenden Fragen jeweils nach den entsprechenden Videoclips:

Informationsausgaben:

Videoclips: „Zeigen Sie mir alle Alben“
„Zeige mir alle Alben“

1. Wie hoch bewerten Sie den Nutzen der differenzierten persönlichen Ansprache mit Sie / Du?

sehr hoch □ □ □ □ □ □ sehr gering

Videoclips: „Zeigen Sie mir alle Alben“
„Nennen Sie mir alle Künstler“

2. Wie hoch bewerten Sie den Nutzen der Unterscheidung zwischen „Zeige“ und „Nenne“?

sehr hoch □ □ □ □ □ □ sehr gering

Videoclip: „Welche Lieder sind auf der Playliste Cool Hits?“

3. Wie hoch bewerten Sie den Nutzen der Darstellung der Alben mit den Interpreten?

sehr hoch □ □ □ □ □ □ sehr gering

Videoclip: „Zeige mir die Alben von Herbert Grönemeyer“

4. Wie hoch bewerten Sie den Nutzen der impliziten Bestätigung?

sehr hoch □ □ □ □ □ □ sehr gering

Videoclip: „Ich will ein Rock-Lied“

5. Wie hoch bewerten Sie den Nutzen der ausführlicheren Benutzerführung?

sehr hoch □ □ □ □ □ □ sehr gering

Videoclip: „Nennen Sie mir alle Künstler“

6. Wie hoch bewerten Sie die Anpassung des Systems an das Vokabular des Benutzers („Künstler / Interpreten“)?

sehr hoch □ □ □ □ □ □ sehr gering

Bitte beantworten Sie die folgenden Fragen abschließend, nachdem Sie alle Videoclips gesehen haben:

Vergleich der Eingabeverfahren:

7. Stellen ein oder mehrere Aspekte der folgenden Liste für Sie Vorteile des SAMMIE-Systems (Versuchsfahrt) im Vergleich zur SAMMIE-Variante (Stand) dar?

○ Persönliche Ansprache Sie / Du

○ Unterscheidung zwischen „Zeige“ (→ optische Darstellung) und „Nenne“ (→ akustische Darstellung)

○ Darstellung der Alben mit den Interpreten

○ Implizite Bestätigung („Alben von Herbert Grönemeyer“)

○ Ausführlichere Benutzerführung

○ Anpassung des Systems an das Vokabular des Benutzers („Künstler / Interpreten“)

○ Sonstiges____

8. Welche Eingabeart würden Sie wohl auf Dauer verwenden?
Bitte nur ein Kreuz!

□ SAMMIE-System (Versuchsfahrt)

□ SAMMIE-Variante (Stand)

□ Kommandowort-System (Versuchsfahrt)

□ *TALK-System (Fahrsimulation)*

9. Welche sonstigen Bemerkungen haben Sie noch zur SAMMIE-Variante im Vergleich zum SAMMIE-System?

____

____

____