Evaluating multimodal interaction patterns in various application scenarios

Frank Althoff, Gregor McGlaun, Manfred Lang, and Gerhard Rigoll

Institute for Human-Machine Communication
Technical University of Munich (TUM)
Arcisstr. 21, 80290 Munich, Germany
{althoff, mcglaun, lang, rigoll}@ei.tum.de

Abstract. In this work, we present the results of a comparative user study evaluating multimodal user interactions with regard to two different operation scenarios: a desktop Virtual-Reality application (DVA) and an automotive infotainment application (AIA). Besides classical tactile input devices, like touch-screen and key-console, the systems can be controlled by natural speech as well as by hand and head gestures. Concerning both domains, we have found out that experts tend to use tactile devices, but normal users and beginners prefer combinations of more advanced input possibilities. Complementary actions most often occurred in DVA, whereas in AIA, the use of redundant input clearly dominates the set of multimodal interactions. Concerning time relations, the individual interaction length of speech and gesture-based input was below 1.5 seconds on the average and staggered intermodal overlapping occurred most often. Additionally, we could find out that the test users try to stay within a chosen interaction form. With regard to the overall subjective user experiences, the interfaces were rated very positively.

1 Introduction

Parallel to the rapid development of computer systems the design of comfortable, robust, and intuitive user interfaces (UIs) has undergone significant changes, too, leading to various generations of advanced human-machine interfaces. The class of *multimodal interfaces* currently resembles the latest step in this development. Providing multidimensional input possibilities and employing innovative audio-visual feedback strategies, these types of interfaces facilitate flexible and intuitive access to the complex functionality of today's technical systems[1][2]. Moreover, multimodal systems offer an increased level of error-robustness, since they integrate redundant information shared between the individual input modalities[3]. Concerning the design of interactive multimodal systems, the fundamental problem consists in effectively combining the individual input modalities to interpret the primary intention of the user (*multimodal integration*). Thereby, the system has to cope with the evaluation of varying interaction patterns with regard to both the semantic content and the time relations of the individual information entities. According to the basic taxonomy developed in [4], the interfaces in this work operate on a late semantic level of synergistic information fusion.

From: GW 2003, Ed.: A. Camurri, LNCS 2915 Springer Verlag

Fig. 1. Multimedia test environment for the DVA (left) and the AIA (right) scenario

1.1 Application domains

In this contribution, we present selected results from a universal usability study evaluating prototypical multimodal interfaces with regard to two different scenarios: controlling a VRML browser in a native desktop Virtual-Reality application (DVA) and operating various audio and communication devices in an automotive infotainment application (AIA). The characteristics of the individual domains are extensively discussed in section 3. A first impression of the specific test environments can be taken from figure 1. Thereby, the user can freely choose among different input channels: conventional tactile interaction, natural speech utterances as well as dynamic head and hand gestures. In the DVA scenario, the user can absolutely concentrate on the tasks in the virtual scene, whereas in the automotive environment, the user additionally has to perform in a driving task. In this setup, operating the multimodal interface is the secondary task only.

1.2 Related work

The group of Oviatt reports on results of an empirical study of people interacting with a bimodal pen- and speech based interface for planning real-estate issues[3]. Unimodal commands given by speech clearly dominated the distribution of interaction patterns (63.5%), followed by command combinations (19.0%), and isolated pen-gestures (17.5%). Contradicting their initial expectations, multimodal draw-and-speak commands were used most often (86.0%), subdivided in simultaneous (42.0%), sequential (32.0%) and compound interactions (12.0%). Concerning intermodal time relations, written input was significantly more likely to precede speech. Detailed analysis revealed that the multimodal interactions occurred within a maximum time zone of four seconds.

Cohen evaluated multimodal time relations in a computer mediated system for battle planning[5]. Astonishingly, the quota of multimodal commands made up more than two thirds (69%) of all interactions which clearly contradicts the prior study. In direct analogy to Oviatt's results, gestural input mostly preceded speech (82%). Sequential interactions occurred less often (11%).

2 General experimental setup

This section describes the general parts of the test setup that are identical for both domains (DVA and AIA), including the overall goals, the available input modalities, the test methodology, and the framework for the test procedures.

2.1 Study background

The primary goal of our usability studies is to evaluate which modalities, and modality combinations, respectively, are preferred with regard to the particular application scenarios and the individual operation tasks at hand. We want to determine to which extent the overall user intention is distributed among complementary, redundant, and competing information streams of the available input modalities. By studying the individual modality transitions, prototypical multimodal interaction patterns of different user classes will be worked out. Additionally, we are interested in analyzing the time relations, i.e. the length of the individual interactions and the specific temporal overlapping. On the basis of these results we are planning to derive some fundamental requirements for a user-centered development of a generic multimodal system architecture.

2.2 Input modalities

To communicate with the target applications, our interfaces provide a wide range of interaction paradigms that can be classified into tactile-based input, speech-based input as well as gesture-based input. Thereby, the individual input devices are designed in a way to support the full set of functionalities if technically possible, i.e. a priori they are not restricted to device-specific interaction forms.

Concerning tactile interactions, our interfaces support input by a conventional touch-screen and a domain specific key-console, which is a standard keyboard in DVA and a special key-console in AIA. By introducing more advanced interaction styles, such as the use of speech and gestures as well as combinations of various input modalities, our interfaces provide a more natural and flexible form of human-machine communication. Speech allows direct interaction without losing eye-focus on the scenario by glancing at the specific input devices. Gestures often support speech in interpersonal communication. In various domains, e.g. when dealing with spatial problems, information can be communicated easier and even more precisely using gestures instead of describing certain circumstances by speech[6]. Moreover, gesture-based input represents an interesting alternative in noisy environments or for people with certain disabilities.

To benefit from the advantages of the individual modalities, in our experimental setup the user can arbitrarily choose among five different input channels: conventional tactile input by **t**ouch-screen (T) or **k**ey-console (K) as well as semantically higher level input by **s**peech (S) and dynamic **h**and (H) and h**e**ad gestures (E). Theoretically, the combinations form a bimodal interaction space of ten possible modality combinations (TK, TH, TE, TS, KH, KE, KS, HE, HS, ES), not regarding any order of usage. Concerning speech input, both natural spontaneous speech utterances and command speech is supported.

2.3 Test methodology

The functionalities of the test interfaces are partly realized according to the *Wizard-of-Oz* test paradigm[7]. In contrast to tactile interactions (touch-screen and key-console), that are directly transcribed by the system, the recognition of the semantic higher-level modalities (speech, hand and head gestures) is simulated by a human person supervising the test-subjects via audio- and video-signals. With regard to speech, an open-microphone metaphor[8] guarantees an optimal degree of naturalness and flexibility since the utterances are segmented automatically. The so-called *wizard* interprets the user's intention and generates the appropriate system commands, which are routed back to the interface to trigger the intended functionality. Thereby, the wizard is instructed to be extremely cooperative. In case of ambiguous user interactions, the input is to be interpreted at best in the current system context. For exchanging information between the individual modules of the system we have developed a special communication architecture based on an extended context-free grammar formalism[9]. As the grammar completely describes the interaction vocabulary of the underlying application on an abstract symbolic level, it facilitates the representation of both domain- and device independent multimodal information contents.

2.4 Test plan

First, the test subjects have to fill out an initial questionnaire. Hence standard data and specific previous knowledge of the users with regard to the application domain is ascertained. For pre-classifying the individual subjects, we apply a weighted evaluation scheme that divides the users into three distinct clusters: **b**eginners (B), **n**ormal users (N), and **e**xperts (E). Afterwards, the test subjects are learning the functionality of the interface in an interactive training period together with the wizard, mainly by employing tactile interaction. At the same time, the use of the other modalities and potential modality combinations are motivated. Although the test subjects are allowed to use their own vocabulary, a basic set of meaningful dynamic hand and head gestures is introduced [10].

A full test consists of two phases. In the first part, consisting of four separate blocks, the participants have to use prescribed modality combinations (TK, TS, HS, ES), each to solve identical operation tasks. The second phase exposes a much more complex operation task, but the test subjects are now allowed to combine all of the available input devices. After each part, the usability of the current interaction paradigm has to be rated according to a six-point semantic differential scale with regard to several adjectives. Additionally, the test users can state universal remarks and discuss upcoming problems with the wizard.

3 Domain-specific setup

Concerning both application scenarios, special preparations and adjustments have been made, which are described subsequently. Thereby, in each case, we explain the software prototype itself, the specific test environment, and finally discuss the individual tasks which the users are exposed to during the test runs.

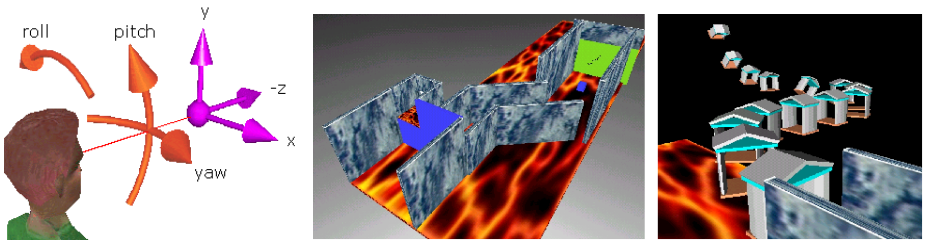

Fig. 2. Avatar centered coordinate system (left) and two screen-shots of the VRML scenario for evaluating the navigation tasks, showing the tunnel of the first phase (middle) and a fraction of the curved series of arch-ways of the second test part (right)

3.1 Desktop VR browser

Test system The first application is based on a multimodal front-end to the standard functionality of a common VRML browser[11]. In the test setup, we solely concentrate on the aspects of navigation. As we apply a first-person-view, the user can directly experience what a virtual avatar would see, thereby covering the full spectrum of continuous movements. Changing the field of view is equal to modifying the location and orientation of a central scene camera.

In VRML, both objects and any kind of interactions are defined with reference to a fixed world coordinate system (*wcs*). Navigating in this scene is equal to transforming an avatar centered coordinate system (*acs*) within this system. Mathematically, this can be described by a homogeneous transformation operation $p_{new} = T_{4x4} \cdot p_{old}$, covering both translational and rotational components.

With respect to the local *acs* shown on the left side of figure 2, six degrees of freedom can be identified. In the following, the names for the particular command clusters are given in parentheses. Concerning translational movements, the user can change the position along the z-axis, moving forward and backward (TFB), along the x-axis, moving left and right (TLR), and along the y-axis, moving up and down (TUD). Concerning rotational movements, turning the field of view left and right around the y-axis is called *yaw* (RLR), rotating up and down around the x-axis is called *pitch* (RUD) and a twisting move around the optical axis of the virtual scene camera (z-axis) is called *roll* (RRO).

Test environment The user study is carried out at the usability laboratory of our institute[8]. This laboratory consists of two rooms. The test subjects are located in a dedicated test room that is equipped with multiple, freely adjustable cameras and microphones. Separated from this area by a semi permeable mirror, the control room serves for recording and analyzing the individual user interactions. To carry out reproducible test runs with identical boundary conditions and to decrease the cognitive workload of the wizard, we have developed a special software suit[12] simplifying the management of various system parameters, semi-automatically announcing the operation tasks at specified points of time, and logging all kind of transactions on a millisecond basis.

User tasks In the first phase, the test users have to navigate through a kind of tunnel (see figure 2), mainly by employing translational and rotational movements (TFB, TLR, RLR). At the end of tunnel, they find a text string written on the wall. The test persons have to change their view in a way that the text becomes clearly readable in a horizontal position. This movement involves a rotation around the optical axis of the virtual camera (RRO). Finally, by the far end of the tunnel, there is a box located on the floor. The task is to look into the box which involves a combination of translational movements in the image plane and the horizontal plane as well as up/down rotations (TUD, TFB, RUD). In the second phase, the test subjects have to navigate through a curved series of archways. Thereby, they have to apply the full spectrum of movements which they have learned in the first four modality specific training blocks.

3.2 Automotive audio- and communication services

Test system The second application scenario deals with a multimodal front-end for controlling various audio devices as well as standard telecommunication units [13]. The player module provides well-known CD-player functionalities (*play, pause, stop, skip, etc.*). In radio mode, the user can switch between 25 different predefined radio stations. The telephone functions are restricted to basic call handling of predefined address-book entries. Moreover, the volume of the audio signal can be adjusted in a separate mode. As shown on the right-hand side of figure 1, the interface can be operated by the same class of input devices like the DVA, only differing in a specially designed key-console. As the central design element, the interface provides a list containing individual items that can vertically be scrolled through by means of two buttons on the right. Above the list, four buttons provide direct access to the individual modes of the application (MP3, radio, telephone, and control). In addition, the area beneath the list contains context specific buttons varying from three to five buttons in the specific modes as well as a feedback line. The design of the key-console is organized in direct analogy to the layout of the buttons on the touch-screen.

Test environment The user study is carried out at a usability laboratory that has specially been adapted to evaluate multimodal user interfaces in automotive environments[8]. To simulate realistic conditions in non-field studies, the lab provides a simple driving simulator, consisting of a specially prepared BMW limousine. Using steering wheel, gas and break pedals, the subjects have to control a 3D-driving task, which is projected on a white wall in front of the car[14]. Thus, they experience the driving scenario from a natural in-car perspective and can better anticipate the roadway. The individual parameters of the simulation can fully be controlled, e.g. the degree of the curves, day- or night sight conditions, speed regulations, obstacles, or passing cars. Besides touch-screen and key console, the car is equipped with a number of microphones and cameras to supervise the test subjects. The audio- and video signals from inside the car are transferred to a separated control room that serves for recording and analyzing user interactions with the test interface and the driving performance.

Table 1. Distribution of the command structure (left) and the proportion of unimodal and multimodal system interactions (right) in DVA for all user groups **b**eginners (B), **n**ormal users (N), and **e**xperts (E) as well as the average mean value (M) for all users

DVA	B	N	E	M
full command	58.2	73.6	72.6	69.5
partial command	41.8	26.4	27.4	30.5

DVA	B	N	E	M
unimodal	73.8	82.1	81.5	79.8
multimodal	26.2	17.9	18.5	20.2

User tasks In direct analogy to the DVA scenario, the user test in the automotive environment consists of two phases. Thereby, the test subjects are exposed to a wide variety of tasks that can be subdivided into five command clusters: player commands (PLY) for starting, stopping and pausing the currently selected track as well as skipping in the play-list, radio commands (RAD), telephone commands (TEL), mode-spanning list commands (LIS) for scrolling in the list display and selecting individual entries and, finally, universal control commands (CTL) for adjusting sound parameters or selecting various operation modes.

Concerning the modality specific training phase of the first four blocks, the test subjects have to accomplish 16 different operation tasks that are uniformly distributed among the various command clusters. To keep the test subjects from devoting most of their attention to control the test interface, they have to perform in a driving task simultaneously. In the second part, the subjects have to fulfill 23 operation tasks on the background of a slightly more difficult driving task and, additionally, they are distracted by changing boundary conditions like an increased frequency of curves, speed limits, and obstacles on the road.

4 Results desktop VR application (DVA)

A total of 40 persons participated in the usability tests (14 female and 26 male), with 11 beginners, 20 normal computer users, and nine experts. The average age of the users was 28.8 years. Besides many engineering students, people of different education and profession took part in the tests. Since the second test block symbolizes the core, non-restricted multimodal test conditions, we especially concentrate on evaluating the data of this part in the following subsections.

4.1 Command structure

The distribution of command types is shown on the left side of table 6. The average number of all system interactions in the second part is 237.4, varying from 85 to 679 (σ=115.33). Strongly related to the navigation tasks that have been superimposed on the subjects, as expected, the set of translational forward/backward movements (TFB) have been used most often, making up more than the half of all commands. With 16.0% on the average, the class of left/right rotations (RLR) provides the second-best frequented type of system interactions.

Concentrating on the semantic content, an isolated system command is composed of a *functional part*, i.e. indicating the special kind of navigation movement

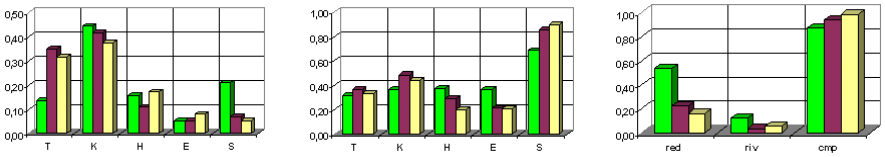

Fig. 3. Distribution of unimodal (left) and multimodal (middle) user interactions with regard to the individual modalities in DVA. The right diagram shows the distribution of multimodal information parts (red=redundant, riv=rivaling, cmp=complementary) in DVA. Concerning each diagram, the left column (green) stands for beginners, the middle column (purple) for normal users and the right column (yellow) for experts.

(translation or rotation) and a and *parameter value* qualifying the direction of the movement (forward, backward, left, right, up and down). A user interaction can either consist of both components (*full command*) or only one single slot (*partial command*). The left side of table 1 contains group specific details.

4.2 Use of modalities

The distribution of unimodal system commands is visualized in the left diagram of figure 3. Obviously, all user groups favor keyboard for unimodal interactions (B=44.5%, N=41.7%, E=37.7%). For next best choice, experts and normal users clearly prefer touch screen (N=35.1%, E=31.7%), whereas beginners tend to employ speech (20.9%), closely followed by hand gestures (15.7%).

As the navigation tasks were quite simple, unimodal interaction clearly overruled multimodal interaction. The respective quota for the individual user groups (B,N,E) and the average mean value (M) is listed on the right side of table 1. Yet, detailed analysis clearly prooves that with growing complexity, the use of multimodal interaction increases. Although only applied in about one fifth of all interactions, combined multimodal commands symbolize the core interaction style, as they were particularly used to change navigation contexts. For all groups, the use of speech clearly dominates the distribution of multimodal interactions (see middle of figure 3). When combining multiple input modalities, especially beginners favored gestures and speech. The other two groups strongly combined speech for mode setting with tactile devices to indicate directions.

4.3 Multimodal interaction

The distribution of redundant (red), rival (riv), and complementary (cmp) interactions is shown in the right diagram of figure 3. For all groups of users, complementary actions occurrs most often (B=86.9%, N=93.9%, E=98.3%). In only 12.5% of all complementary actions, more than two modalities are applied. Real multimodal commands (showing no intramodal dependencies) appears in 67.3% of combined interactions. Especially beginners seem to apply redundant interactions (53,2%) more than twice as much in comparison to normal users

Table 2. Distribution of relative modality transitions (RMT) in DVA (left) and AIA (right) from source modality mod_s to target modality mod_t, specified in % relative to all transitions of a given source modality, self transitions are marked in bold letters

DVA	T_t	K_t	H_t	E_t	S_t
T_s	**63.79**	12.61	2.59	2.50	12.20
K_s	22.48	**60.52**	5.22	4.65	11.48
H_s	7.84	7.94	**52.41**	7.65	32.57
E_s	9.23	21.00	21.56	**35.10**	33.05
S_s	16.25	31.64	23.07	13.75	**46.49**

AIA	T_t	K_t	H_t	E_t	S_t
T_s	**38.01**	18.77	7.53	2.32	35.44
K_s	16.73	**47.47**	9.20	2.74	23.97
H_s	17.69	24.45	**4.17**	2.07	48.14
E_s	16.67	4.17	14.58	**0.00**	67.36
S_s	20.49	8.59	7.10	1.52	**60.48**

(22.4%). These results also emphasize the observation that beginners show complementary behavior coupled with redundancy to a high degree.

The effectiveness of combined user interactions can be measured as the number of performed transactions per time (TpT). Experts and normal users work most efficient when combining touch screen with keyboard (1.04 / 0.95 TpT) or speech (0.54/0.47 TpT). As an outstanding result, we obtained that in part three of the first block (SH), beginners get even higher scores than the other two groups (0.30 compared to 0.27 and 0.23 TpT, respectively). Purely tactile interaction (TK) is still most time efficient, but concerning the other combination scenarios, all groups of test subjects work significantly more effective ($p < 0.01$) in the second block with free modality combination. In this regard, normal users nearly perform 32% better than experts (0.80 vs. 0.61 TpT).

4.4 Modality transitions

Concerning the transitions between the individual modalities, we have examined the relative modality transitions (RMT), defined as the percentage of transitions with respect to a fixed source modality $mod_s \in \{T_s, K_s, H_s, E_s, S_s\}$ to the potential target modalities $mod_t \in \{T_t, K_t, H_t, E_t, S_t\}$. The results for the DVA are shown on the left side of table 2 as an average mean value for all user groups. As in the case of redundant multimodal interactions several modalities contribute to a single system command, selective rows sum up to a value higher than 100%.

The results clearly confirm the primary expectations, that users tend to stick to tactile interactions much stronger than to speech or gestures. This interaction behavior can be observed for all user groups. Especially after an interaction with a head gesture, the modality is changed in most cases. Analyzing the average modality specific retention period ($AMRP$), which is defined by the number of self transitions divided by the total number of transitions to other modalities, the observations above are worked out even more clearly ($AMRP_T = 4.69$, $AMRP_K = 10.13$, $AMRP_H = 2.02$, $AMRP_E = 0.71$, and $AMRP_S = 5.42$).

4.5 Time relations

The left side of table 3 summarizes the medium interaction times for the semantically higher level modalities speech, hand and head gestures in DVA. For

Table 3. Average interaction times in seconds for semantic higher-level input by speech utterances, as well as by hand and head gestures in DVA (left) and AIA (right)

DVA	B	N	E	M
speech	0.99	0.90	1.06	0.96
hand gesture	1.02	0.99	1.32	1.08
head gesture	1.15	1.15	1.31	1.17

AIA	B	N	E	M
speech	0.87	0.87	1.10	0.94
hand gesture	0.90	0.90	1.20	1.08
head gesture	1.98	1.98	1.73	1.49

Table 4. Distribution of time relations in DVA (left) and AIA (right), subdivided in staggered, nested, and sequential interactions according to the scheme given in [3]

DVA	B	N	E	M
staggered	37.08	69.16	46.43	55.16
sequential	7.03	8.43	7.14	7.79
nested	55.89	22.41	46.43	37.05

AIA	B	N	E	M
staggered	50.00	66.67	50.00	43.75
sequential	50.00	0.00	12.50	39.06
nested	0.00	33.33	37.50	17.19

all groups of users, in the case of speech the shortest interaction times could be observed. Astonishingly, expert users exposed significant longer interaction times compared to beginners and normal users.

With reference to the scheme for analyzing intermodal time relations described by Oviatt[3], multimodal interactions are classified into *sequential*, *staggered*, and *nested* interactions. Assume, we have two interactions I_1 and I_2 lasting from time-stamp t_{1s} to t_{1e} and t_{2s} to t_{2e}, respectively (with $t_{1s} < t_{1e}$, $t_{2s} < t_{2e}$ and $t_{1s} < t_{2s}$). Then we speak of a sequential action, if $t_{2s} > t_{1e}$, a staggered action, if $t_{2s} < t_{1e}$ and $t_{1e} < t_{2e}$, and a nested action, if $t_{2e} < t_{1e}$.

The different user groups exhibit massively varying interaction patterns. While beginner and experts mainly use nested interactions with one modality completely enclosing the other, normal users tend to prefer staggered interactions. For each group, purely sequential interactions have been applied rarely. Finally, analyzing the interaction time by separately regarding only unimodal and multimodal interactions, respectively, we did not find any significant differences in the lengths of the individual interactions.

4.6 Subjective user experiences

As an overall result, analyzing the questionnaires reveals good accordance with the measured values discussed above. Experts and normal users rate touch-screen in combination with speech best, whereas beginners state to prefer speech in combination with hand gestures. Interestingly, head gestures have been rated very bad, which contradicts the measured values. In fact, combined with speech, head movements make up at least 20.0% of all combined interactions. Concerning the remarks of the closing questionnaire users have asked for advanced navigation features, i.e. they want the system to continuously react on head and hand movements. Moreover subjects have demanded to phrase browser commands applying context knowledge of the current navigation situation, e.g. "go to the wall" or "turn the view until the arch-way is in a horizontal position".

Table 5. Distribution of unimodal (left) and multimodal (right) interactions in AIA

AIA	T	K	H	E	S
B	24.98	31.09	3.49	0.96	39.49
N	38.69	41.14	0.00	0.13	20.04
E	36.21	34.92	4.81	0.43	23.63
M	31.87	34.43	2.77	0.93	30.00

AIA	T	K	H	E	S
B	0.00	0.00	50.00	0.00	50.00
N	0.00	0.00	33.33	16.67	50.00
E	0.00	0.00	25.00	25.00	50.00
M	0.00	0.00	33.33	16.67	50.00

5 Results automotive infotainment application (AIA)

A total of 28 persons participated in the second usability test (five female, 23 male). The participants are grouped in eight beginners, 13 normal computer users and seven experts. The average age of the test users was nearly 29 years. An intersection of five participants have taken part in both usability studies. Just like the DVA szenario, the evaluation of the test data mainly concentrates on the analysis of the second block.

5.1 Command structure

The distribution of the command types is shown on the right side in table 6. Again, the results are strongly related to the specific tasks in the test. Thereby, the individual users mostly differed in the number of list commands (LIS) and control commands (CTL), because some test subjects changed volume parameters several time during the test run. The average number of all system interactions (ASI) in the second block is 48.50 with a standard deviation $\sigma = 9.98$. Concerning the semantic content of an isolated system command, we exclusively find full commands in the AIA.

5.2 Use of modalities

Although the test subjects have been allowed to use the full spectrum of input devices, a detailed analysis of the audio-visual material offers an unambiguous tendency towards a concentration on two devices. For most of the users, this has been the combination touch-screen and speech which is supported by the data on the left in table 5 showing the distribution of unimodal interactions and on the right in table 5 containing the distribution of multimodal interactions.

Regarding all test subjects, in total 51 multimodal interactions have been observed. On the background of 1358 transcribed system interactions, this corresponds to an overall multimodal quota of only 4.2%. This result contradicts to our expectations, but can be explained by the fundamental differences in the experimental setups. The tactile devices have not been used in combination with speech and gestures at all, which is an important difference compared to the results in DVA. Concerning the distribution of multimodal information, redundant components make up 85.71% of all multimodal interactions. Since rivaling interactions have not been observed, the complementary parts cover

the remaining 14.29%. As the total number of multimodal interactions is very small and only two test users showed more than two multimodal interactions, no significant differences between the individual user groups can be identified. The underlying test material does not provide a sufficiant basis for statistically valuable arguments.

5.3 Modality transitions

The average RMTs for the automotive scenario are shown on the right in table 2. In contrast to the DVA, self-transitions do not symbolize the primary form of modality changes. Concerning hand gestures, the users change to speech instead of continuing purely gesture-based interaction. If head gestures are applied, none of the test subjects uses them for a second time in a row.

5.4 Time relations

The right side of table 3 summarizes the medium interaction times for the semantically higher level modalities. According to the interaction times, speech offers the fastest access, closely followed by hand gestures. With regard to the head gestures, we find significantly longer interaction times, compared to DVA. The observation that experts show longer interaction times can be confirmed in AIA. In direct analogy to the results in the DVA scenario, the analysis of the interaction times in the unimodal and multimodal case does not reveal any significant differences. The different user groups again showed strongly varying interaction patterns documented on the right in table 4. While beginner and experts mainly use nested interactions with one modality completely enclosing the other, normal users tend to prefer staggered interactions. For each group, purely sequential interactions have been applied rarely.

5.5 Subjective user experiences

Compared to the prescribed modality combinations in the first block, the system offering the full functionality obtaines significantly better ratings with regard to various usability criteria. Only the combination of speech and touch-screen was rated better concerning the quality of effective usage.

Natural speech has clearly been preferred for almost all system functionalities. While the key-console has obtaines best ratings for scrolling in the list and adjusting the volume, the touch-screen is chosen most often to skip between individual tracks in a play list. Head gestures represent an interesting special case. With eight of the test subjects favoring them as the primary input modality for yes/no decisions, like in accepting or denying an incoming phone call, the other users totally dislike this input form. When applying speech, the test subjects make extensive use of natural speech utterances, that are mainly applied for complex system functions involving context knowledge of the application domain (e.g. directly saying "play radio station *ABC*" instead of scrolling in the list and selecting the appropriate entry). Moreover, the test users expect the system to understand combined commands, like "go to the last entry".

6 Comparative discussion

Although the two application domains cannot be compared directly due to massively varying boundary conditions, certain characteristics of multimodal interaction patterns can be worked out. Concerning the general distribution of the individual input modalities, speech clearly dominates the use of the semantic higher-level modalities in the automotive environment. The results of the direct comparison are shown in the left diagram of figure 4. In the automotive setup, head and hand gestures have been used very rarely. This partly results from the fact that speech has been recognized on a basis of 100% due to the simulated Wizard-of-Oz recognition module and a highly tolerant wizard.

Comparing multimodal and unimodal interactions, the DVA exposes a much higher quota of combined interactions in general (more than 20% in DVA and only 2.1% in AIA). The ratio is visualized in the middle diagram of figure 4. While experienced users tend to focus on standard tactile devices, beginners prefer speech and gestures and combine modalities more often. Especially beginners show complementary interactions accompanied by simultaneous redundant information. Regarding the distribution of multimodal information (shown on the right in figure 4) we have determined an increased level of redundant interactions in DVA compared to more complementary interactions in AIA.

A detailed analysis of the audio-visual data material clearly reveales that the lengths of the interactions by the semantical higher-level modalities speech, head and hand gestures stay within a maximum time-period of 1.5 seconds. Concerning the intermodal time relations, staggered interactions are used in most cases. With reference to the modality transitions we have found out that once adapted to a specific modality, the test users try to stay within the chosen interaction form. This especially holds for the tactile devices and less for head-gestures. Concerning selected functionalities like yes/no decisions head-gestures represent the preferred form of input, at least for about one fourth of all users in the test trials. Analysing the closing questionaires with reference to several usability criteria, the full multimodal interaction paradigm has been rated significantly better compared to the individual bimodal combinations.

7 Conclusion and future work

In this work, we have presented the results of a comprehensive evaluation on multimodal system interaction regarding five input modalities and two different application scenarios: a desktop Virtual-Reality browser and an automotive infotainment application. The individual multimodal interfaces have been compared with regard to command clusters, the use of modalities, multimodal combinations, modality transitions, time relations and subjective user experiences.

Currently, we are working on running detailed statistical tests on the data material, especially evaluating the results with regard to specific tasks and potential user errors that may always occur during the interaction. For the nearest future, we plan to integrate the results of this study as empirical data in the design of a generic integration architecture for multimodal interactive systems.

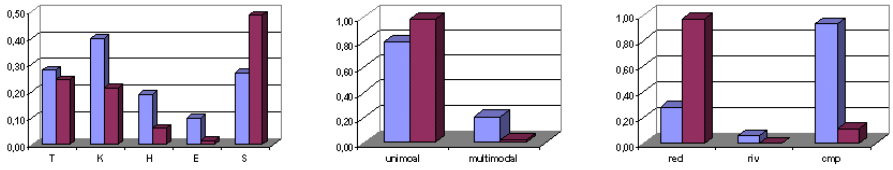

Fig. 4. Comparing DVA and AIA: the distribution of the individual modalities (left), the proportion of unimodal and multimodal commands (middle) and the distribution of multimodal information parts (right); concerning each diagram, the left column (blue) represents the DVA scenario and the right column (purple) the AIA domain

Table 6. Distribution of the command types for DVA (left) and AIA (right), listed in % for the individual user groups (B,N,E) and the average mean value (M) for all users

DVA	TFB	TLR	TUD	RRO	RLR	RUD
B	50.7	8.9	8.4	5.5	16.8	9.5
N	51.4	10.1	4.6	5.2	16.5	12.1
E	52.4	10.3	6.4	8.7	13.9	8.4
M	51.4	9.8	6.1	6.1	16.0	10.6

AIA	PLY	RAD	TEL	LIS	CTL
B	18.6	4.0	16.1	31.1	30.1
N	19.8	4.3	17.2	36.6	22.1
E	20.9	4.2	17.0	35.0	24.3
M	20.0	4.3	17.0	34.8	24.3

References

1. Oviatt, S.L.: Multimodal interface research: A science without borders. Proc. of the 6th Int. Conf. on Spoken Language Processing (ICSLP) (2000)
2. Cheyer, A., Julia, L.: Designing, developing and evaluating multimodal applications. In: WS on Pen/Voice Interfaces (CHI 99). (1999)
3. Oviatt, S.L., et al.: Integration and synchronization of input modes during multimodal human-computer interaction. Proc. of the 6th ICSLP (2000)
4. Nigay, L., Coutaz, J.: A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion. In: Proc. of INTERCHI '93, ACM Press (1993) 172–178
5. Cohen, P., et al.: Multimodal interaction during multiparty dialogues: Initial results. Proc. of 4th IEEE Int. Conf. on Multimodal Interfaces (2002)
6. Sowa, T.: Coverbal iconic gestures for object descriptions in virtual environments: An empirical study. Post-Proc. of Int. Conf on Gestures: Meaning and Use (2000)
7. Nielsen, J.: Usability Engineering. Morgan Kaufmann Publishers Inc. (1993)
8. Neuss, R.: Usability Engineering als Ansatz zum Multimodalen Mensch-Maschine-Dialog. PhD thesis, Technical University of Munich (2001)
9. McGlaun, G., et al.: A new approach for the integration of multimodal input based on late semantic fusion. In Proc. of USEWARE 2002 (2002)
10. Zobl, M., et al.: A usability-study on hand-gesture controlled operation of in-car devices. Proc. of 9th Int. Conf. on HCI (2001)
11. Althoff, F., et al.: A generic approach for interfacing VRML browsers to various input devices. Proc. of ACM Web3D Symposium (2001) 67–74
12. Schuller, B., Lang, M., et al.: Towards automation of usability studies. Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics (SMC02) (2002)
13. Althoff, F., Geiss, K., et al.: Experimental evaluation of user errors at the skill-based level in an automotive environment. Proc. of CHI 02 (2002)
14. McGlaun, G., Althoff, F., Lang, M.: A new technique for adjusting distraction moments in multitasking non-field usability tests. Proc. of CHI 02 (2002)