Towards Multimodal Error Management: Experimental Evaluation of User Strategies in Event of Faulty Application Behavior in Automotive Environments

Gregor MCGLAUN, Frank ALTHOFF, Manfred LANG, and Gerhard RIGOLL

Institute for Human-Machine Communication
Technical University of Munich
Arcisstr. 16, 80290 Munich, Germany
phone: +49 89 289-28541

{mcglaun, althoff, lang, rigoll}@ei.tum.de

ABSTRACT

In this work, we present the results of a study analyzing the reactions of subjects on simulated errors of a dedicated in-car interface for controlling infotainment and communication services. The test persons could operate the system, using different input modalities, such as natural or command speech as well as head and hand gestures, or classical tactile paradigms. In various situational contexts, we scrutinized the interaction patterns the test participants applied to overcome different operation tasks. Moreover, we evaluated individual user behavior concerning modality transitions and individual fallback strategies in case of system errors. Two different error types (Hidden System Errors and Apparent System Errors) were provoked. As a result, we found out that initially, i.e. with the system working properly, most users prefer tactile or speech interaction. In case of Hidden System Errors, mostly changes from speech to tactile interaction and vice versa occurred. Concerning Apparent System Errors, 87% of the subjects automatically interrupted or cancelled their input procedure. 73% of all test persons who continued interaction, when the reason for the faulty system behavior was gone, strictly kept the selected modality. Regarding the given input vocabulary, none of the subjects selected head or hand gesture input as the leading fallback modality.

Keywords: error, management, user, behavior, interaction, multimodal, automotive;

1. INTRODUCTION

Today's growing complexity of in-car infotainment and communication systems strongly implicates an enlargement of input modalities in cars. Multimodal interfaces (MI) offer a lot of advantages to the driver. Compared to monomodal systems, MIs allow for shorter learning phases and a highly intuitive and individual interaction [1]. Prior studies of Oviatt et al. showed that in purely speech-based systems, the recognition rate dropped by 20-50%, when input was provided during natural or spontaneous interaction, by different user groups (e.g., accented speakers, children, or speech impaired people), or in noisy mobile environments [2].

Error-prone situations are very likely to occur during interaction with various applications in a car environment. If caused by heavy traffic noise, the signal-to-noise ratio gets drastically worse, e.g., speech recognition will probably no longer work properly. Hence, multimodal interfaces have great potential for a significant enhancement of error robustness. Oviatt et al. mention that in dedicated scenarios, up to 86% of all task-critical errors can be avoided, if an alternative input modality is provided [3]. A special set of multimodal systems facilitates user interaction in a *synergistical* [4] way, i.e. the user can enter input temporally overlapping in different modalities. Besides the gain of efficiency, in case of *redundant* [4] input, recognition errors of a single modality could directly be avoided by *mutual disambiguation* [5]. For example, if a speech recognizer issues an *n*-best list with low confidence for the potential output candidates, additional visual information by lip-reading can result in correct recovery of the input. On the other hand, the user can at any time choose freely amongst the provided modalities, which allows for a highly natural and intuitive way of human-machine communication. In case the selected modality channel fails for some reason, it is necessary to have a comprehensive *error management* that assists the user in performing the desired interaction (e.g., offering so-called *fallback modalities* dependent on the context of the application and the system environment). One step in a targeted development of an effective error handler is to evaluate how the multimodal interaction behavior of the user changes in case of system errors.

2. THEORETICAL BACKGROUND

In the field of error theories, many researchers have contributed significant work.
Strictly following an absolute philosophical point of view, Festinger [6] has developed an approach of cognitive dissonance for describing user errors. In his model, human error is always an expression of certain habits that cannot automatically be used in specific situations and thus result in an error during the operation.
Rigby [7] differentiates between sporadic, accidental, and systematic errors. In his phenomenological approach, sporadic errors are singular events, and are often considered as outliers.

Accidental errors have a high mean variation with regard to the intended target status, but in contrast to systematic errors, they do not show any clear tendency towards a special direction.
However, these two approaches can hardly be used in a practical application since they suffer from a significant drawback. As the flow of interactions is assumed to be controlled by the *system* exclusively, the *user* is not involved sufficiently.
Reason [8] has given the theoretical basis for modeling potential error-prone user interactions. Related to the skill-rule-knowledge framework of Rasmussen [9], he differentiates between errors on three different performance levels (see table 1).

levels (*J. Rasmussen*)	**skill-based (SBL)**	**rule-based (RBL)**	**knowledge-based (KBL)**
description	routine actions	productions	analytical processes
error type (*J. Reason*)	slips, lapses	planning-failure (stored rules)	planning-failure (novel situation)
causation	deviation from a trained routine	misclassification of situations	unpredictable changes

Table 1: The skill-rule-knowledge framework of Rasmussen

User interactions at the *skill-based level* comprise operations that have already become routine actions by multiple execution. Characteristic errors are either execution failures (slips) or failures of memory (lapses). They imply a deviation (normally known in advance) from a well-trained routine. At the *rule-based level*, human performance is determined by stored rules (productions). Hence, error patterns are planning failures (mistakes), and typically related to the misclassification of situations. At *the knowledge-based level*, in novel situations, problems are solved by applying conscious analytical processes and stored knowledge. Significantly, errors arise from unpredictable environmental changes one is not prepared for.

Interaction Errors
Based on the formal description and abstract classification of human errors discussed above, we will derive an expedient definition of an interaction error that additionally covers system failures and faults. In the following, we briefly list some prototypical error-prone situations in human-machine communication.
In the first case, the user gives a command that is interpreted by the system in a certain context that does not match the primary intention of the user. In a second scenario, a given command is interpreted in the wrong way, and then executed. The system as well as the user can be the etiological cause of an error. If the mental model of the user (which is a combination of the task model and the system model) and the user model of the system differ to a certain degree, an issued command will be interpreted in the wrong context. The significance of the error potential becomes higher, the later the proceeding divergence of the two models is detected.
Covering these individual cases, we can give the following definition of an interaction error:
An error in human machine communication is a consequent result if the requirements and the intention of the acting part are not covered in a sufficient way by the reacting part.
Thereby, the acting part can be both the system and the user.

Evaluating errors that appear during human-machine interaction, it is very important to distinguish whether the user or the system actually caused an error. This work exclusively focuses on scenarios in which the user as a correctly acting part faces a certain malfunction of the system. In this regard, we can identify two different error classes.
Hidden System Errors (HSE) are spontaneous errors that occur independently from any contextual conditions (e.g., sudden break down of a module). In the current situation, it is not apparent or comprehensible for the user why the error happened. This class of errors is characterized by partial or total recognition failures in the used input modality.
Moreover, we evaluated so-called *Apparent System Errors* (ASE). Hence, the cause that leads to the error is clearly evident to the user (e.g., the user interrupts the interaction with the system due to an incoming call on her or his mobile phone).

3. EXPERIMENTAL SETUP

For a dedicated analysis of user strategies induced by the errors of the classes listed above, we designed a *non-field user study*. In the following subsections, we will describe the boundary conditions, the basic method, and the schedule of the test.

Test Platform
The study was conducted in the car laboratory of the institute, which is specially adapted to evaluate multimodal user interfaces in automotive environments. In a separate control room, a test supervisor monitors the run of the experiment. For simulating realistic test conditions, the laboratory is equipped with a simple driving simulator consisting of a specially prepared BMW limousine with a force-feedback steering wheel, gas and break pedals, as well as an automatic transmission. The test subjects have to use these devices to control a 3D driving task, which is projected on a white wall in front of the car. This allows for experiencing the driving scenario from a natural in-car perspective and a better anticipation of the driving course. The test designer can control a large set of individual parameters of the simulation by a dedicated run chart, e.g., day- or night sight conditions, speed regulations, or passing cars. To carry out reproducible test runs, we have developed a special software suit [10] enabling a precise time management of various system parameters, semi-automatically announcing operation tasks at specified points of time, and logging all kind of transactions. The concept has successfully been applied in many former experiments, e.g., [11].
For permanently recording audio and video signals, the car is equipped with a microphone array and two CCD cameras. Together with the data of the driving simulation, we were able to effectively analyze the individual interaction style of the subject.

Test System
The test vehicle is equipped with a prototypical multimedia interface for controlling an infotainment and communication application consisting of an MP3-player, a radio tuner, and an integrated telephone application. The MP3 player features commonly known standard functionalities (like play, skip, stop, etc.). In the radio mode, the test participant can tune to 25 different previously stored stations. The telephone functions are limited to basic call handling (call, accept, deny, etc.) of 30 predefined address-book entries. Moreover, the volume of the audio signal can be controlled in a separate mode. As depicted in figure 1, the interface itself is organized in four separated horizontal areas.

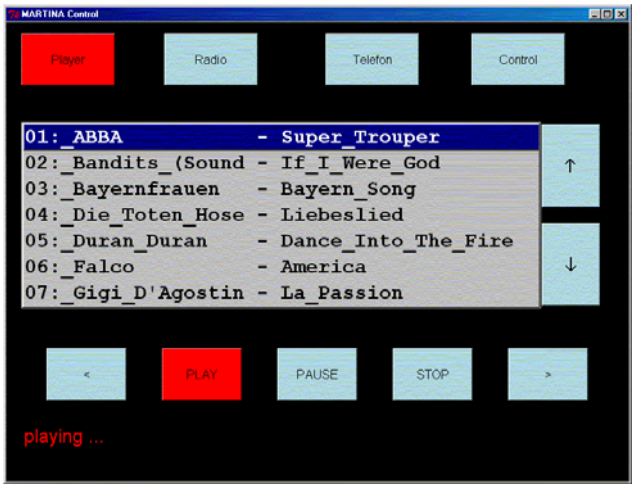

Figure 1: Screenshot of the test interface used in the study

Figure 2: Illustration of the Wizard-of-Oz principle

The top line is composed of four buttons representing the individual *main modes* of the application (MP3, radio, telephone, and control). Directly beneath this button line, as the central design element, the interface provides a list containing individual items that can be vertically scrolled through by the two buttons on the right. The area in the in the lower part contains context specific buttons varying from five buttons in MP3 mode, three in radio and control, and two in telephone mode. In dependence of the current application mode, the system provides the particular functionality by displaying the respective buttons. In addition, the interface contains a feedback line continuously informing the user of the status of the interface, e.g., indicating an incoming call, the currently tuned radio station, or the system volume.
Using a key word ("computer") for initialization, the system can be operated via *natural speech* (SPC). Furthermore, subjects can use *head-* (HEG) and *hand-gestures* (HAG). For interaction via HEG or HAG, there is no initialization paradigm, but subjects are told to make sure that their head or the hand is not outside the focus of the camera. For tactile interaction, there is a 10" *touch-screen* (TSC) located in the middle of the center console, as well as a *keypad* that was *integrated in the armrest* (AKC) of the test car. The AKC consists of a 2x4 button array, which is organized in direct analogy to the position of the buttons on the touch-screen. The buttons of the first row allow for controlling the main modes, the buttons of the second row change their functionality in dependence of the current system mode. The two turning knobs are used for adjusting the volume and for browsing in the list display. By pressing these knobs, subjects can mute the volume and select the current list item, respectively.
The test persons are given a set of six head and 15 hand gestures, as well as 30 speech commands that can be provided in natural speech expressions. Concerning the composition of the interaction vocabulary, six commands (e.g., "yes" and "no") can be entered in any modality channel.

Test Methodology
The study is performed as a partial Wizard-of-Oz (WOO) test [12]. In our evaluation, the test supervisor (also referred to as "wizard") simulates the recognizers for the semantic higher-level modalities (HEG, HAG, and SPC). The wizard interprets the user's intention and generates the appropriate system commands, which are sent back to the interface in the car to trigger the intended functionality (see figure 2).

Haptic interactions via TSC and AKC are directly transcribed by the system, but for simulating error scenarios, the wizard can also interfere with this process.
The test supervisor is instructed to be extremely cooperative. In case of ambiguous user inputs or actions that are similar or synonymous to the given vocabulary set, the test supervisor tries to interpret the interaction at best in the current system context. We have chosen the WOO principle, as it allows for a deterministic system behavior and an arbitrarily adjustable recognition rate.
As presented in former work [13], the driving simulation demands each test subject in a different way. For normalizing the cognitive load induced by the driving task, we have developed a dedicated baseline technique. This method rates the individual driving performance of the subject in a separate test run, and consequently allows for adjusting the degree of difficulty of the driving task in subsequent parts of the trial.

Test Procedure
At the beginning of the test procedure, there is a short training period in which the subjects get to learn the different ways of interaction with the system. Before the main part starts, we carry out the baseline analysis to make sure that each subject is exposed to the same cognitive load, as outlined above. The main phase of the trial is split up into three parts, as follows:

Part 1: Reference Phase: In this phase, which contains 16 different operation tasks, the user can arbitrarily select and combine the given modalities. Regarding the vocabulary set mentioned above, in ten of 16 tasks, the respective functionality of the interface can be accessed via any modality channel. The driving task comprises a simple course (straight road, no obstacles). The goal in this test part is to determine individual modality preferences and the quota of synergistic multimodal input.

Part 2: HSE Scenarios: This step of the test consists of 21 tasks. Again, the user can freely choose and combine all modalities. In five scenarios, the system does not react on any kind of user input (e.g., the user gets the task "Call Mr. Miller," but independent from the chosen input modality, the dial command does not work). As a significant feature of the HSE scenarios, the actual reason why the system does not react is not at all evident for the user. To get comparable conditions, the driving task in this part is identical with that of the reference phase.

Part 3: ASE Scenarios: In this trial part, which comprised 21 tasks, eight ASE scenarios are interspersed. These error situations are simulated by dazzling lights, noise (e.g., braking sounds or honks), or incoming telephone calls interrupting the current action of the user. Moreover, in eight tasks, the test subject is forced to take a certain initial modality. Using a more complex driving task (obstacles on the road, speed limits that have to be kept) than in part 1, we increase the workload of the test participants.

4. RESULTS

In the study, 15 subjects (47 % female, 53 % male, average age 25.5 a) participated.
Regarding the left columns for each modality in figure 3, it can be seen that in the reference phase, tactile interaction followed by speech were the leading modalities.

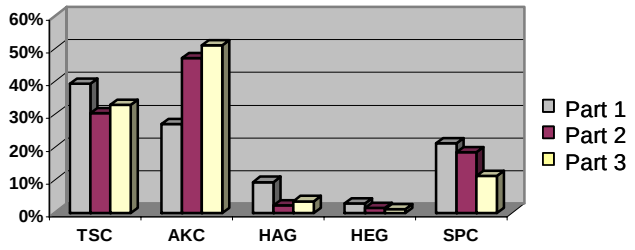

Figure 3: Modality distribution over all test parts

Very few used HAG (8%) or HEG (2%), respectively. 73% of the subjects stated that it was a new experience for them to operate a system via HAG or HEG and that it took time to get used to this kind of input paradigm. Despite massive system failures, the preference of the fallback modalities in part 2 and 3 of the test (middle and right column for each modality in figure 3) was nearly the same. SPC decreased, whereas AKC was even used more often than in the reference phase.
During the whole trial, we could only very sparsely observe synergistic multimodal input (8% of all interactions). Complementary input was mainly delivered sequentially or expressed in a single modality. When we asked test participants for the reason, they pointed out that while driving a car, they tried to keep interaction as simple as practicable. Twelve of 15 subjects would rather execute two actions successively to reach a task goal, even if it eventually took more time.
In the HSE error scenarios, the subjects repeated a command 2.1 times on average, until they changed the modality. This is less than they pointed out in a rating before the test (3.3 repetitions on average). Most retrials were done with tactile interaction via TSC (2.6 on average), AKC (2.4), and with speech (1.9). In contradiction to the subjective data, the average number of command repetitions, using HAG (2.3) or HEG (2.2) was higher than AKC (1.7). If the system did not react for the third time, independently from the initial modality, the subjects used speech commands charged with various emotions in combination with tactile interaction, i.e. hence, they performed redundant synergistic inputs.
We could observe that towards the end of the test, the trial persons showed a tendency to directly change the modality than to retry it in the current one. All subjects pointed out that they increasingly lost faith in the reliability of the modality and thus switched over to another one.

In the questionnaires, we also asked subjects to which fallback they would change if they could no longer use their preferred modality. Concerning the situational context, we assumed a relaxed driving situation on an interstate. As a result, we got the transition matrix containing the averaged ratings (see table 2).

	TSC	**SPC**	**HEG**	**HAG**	**AKC**	***median***
TSC		2.93	2.33	2.53	2.53	***2.58***
SPC	1.13		1.93	1.60	1.67	***1.58***
HEG	2.60	2.86		3.00	3.00	***2.86***
HAG	3.33	3.40	3.15		3.36	***3.31***
AKC	2.27	2.60	2.60	2.07		***2.38***

Table 2: Transition matrix of the input modalities;
first row: initial input modalities,
first column: modalities the user tends to fall back to

For the data ascertainment, we used a semantic differential scale without forced rating [12] with "1" standing for "definitely prefer" and "6" meaning "definitely disprefer."
Most test persons prefer SPC, followed by tactile interaction. All participants dispreferred HEG and HAG. In the eyes of the subjects, some functionalities (e.g., a "random" or a "repeat" command) could hardly or only very intricately be executed by gesturing (particularly HEG). 75% of the test persons switched from SPC to TSC. With AKC failing, only 33% of the subjects changed to TSC, whereas 47% switched over to SPC. In good agreement with the subjective ratings, HEG (0%) and HAG (6%) were hardly used as a fallback. Moreover, subjects tended to keep their modality as long as possible. In the ASE scenarios, 87% automatically interrupted the input, when an external event interfered with their action. 27% of the test participants forgot to finish the task they had begun. All of these subjects pointed out that in such a case, they expected the system to remind them of the unfinished task in a way that they could proceed exactly from the point where they had suspended. Those who continued interaction, when the derangement was over, strictly kept the selected modality.

5. DISCUSSION

As mentioned above, the study was designed as a laboratory wizard-of-oz trial. For interpreting the semantic higher-level modalities, a wizard reacted instead of real recognizers, which was the base for a deterministic system behavior and an arbitrarily adjustable recognition rate. The laboratory setup allowed for an integrated reproducibility of the scenarios, which is an important feature for the inter-individual comparison of the results. This methodology provided a fast and cost-efficient implementation of the problem compared to a field test. However, it has to be mentioned that in a real traffic scenario, the conclusions might partially differ from the findings of this contribution: the anxiety of economical or physical damages additionally influences the interaction behavior of the test persons. Even in a high-level driving simulator, this situational awareness can hardly be generated due to an irreducible lack of immersion.
An outstanding result is the strong dispreference of the subjects regarding HEG and HAG. For this, several reasons can be accounted. First, our test application had not perceivably been

optimized for HEG and HAG. There are existing systems implementing highly specialized gestural concepts. For example, the application presented in [14,15], always displays a set of valid gestures in the current system context. But for our purpose, the system should not advise, emphasize, or force any modality in order to influence the user as less as possible. Moreover, compared to a speech command, most of the gestures are not standardized. A total of 13 subjects pointed out that they forgot most of the gestures they were explained at the beginning of the test, and that they used there own ones later on. Finally, nearly all subjects stated that in some cases they simply could not figure out how to express the interaction command (e.g., the "random" functionality) with a gesture.

The study presented here is the first of a series of analyses regarding the development of an error-robust multimodal interface in the automotive environment. With 15 subjects evaluated, the test results cannot be generalized without further considerations. The main intention was to give a first impression and to show tendencies for a guideline on how the user accepts different modalities in case of system-intrinsic malfunctions.

6. CONCLUSIONS AND FURTHER WORK

The study clearly proved that the situational context, implied by the state of the user (like emotions), the current mode of the system (in which the error has happened), and the environmental parameters (traffic conditions, etc.) have to be considered in a purposive error management. To be effective and user-friendly, the system must make a sensible taxonomy whether the current modality should be changed or the action can be retried in the initial modality. The sparse use of HAG and HEG shows that these modalities are critical as fallback modalities, unless the driver is not used to this kind of interaction paradigm, e.g. by a special training or tutorial. Therefore, we currently work on the design of a long-time evaluation in this regard.

In ongoing work, the findings are iteratively integrated into an error-handling component of a multimodal in-car infotainment and communication system [16,17]. The system is based on a client-server architecture, where information of the monomodal recognizers is processed via a late semantic fusion approach. To verify the usability of the error management component, extensive user studies are currently conducted, using real recognizers for natural speech and gesture interaction [14,15,18,19]. To verify the usability of deduced error strategies, the system will also be tested in real traffic scenarios. The target group of our study consisted of users aged between 18 a and 35 a. Some evaluations (e.g., [13]) in this field of research show that the age has a certain impact on the driving performance and, as a consequence, on the interaction with in-car devices while driving. Thus, we plan a large analysis for several age cohorts in the near future, using the results of this study as a reference.

7. ACKNOWLEDGMENTS

The work presented in this paper has been supported by the FERMUS project [20], which is a cooperation between the BMW Group, DaimlerChrysler AG, SiemensVDO AG, and the Institute of Human-Machine-Communication at the Technical University of Munich. FERMUS stands for "Error Robust Multimodal Speech Dialogs," and addresses the development of error resolution strategies in multimodal in-car infotainment and communication systems.

8. REFERENCES

[1] S. Oviatt, "Taming Recognition Errors Within a Multimodal Interface," **Com. of the ACM**, 2000

[2] S. Oviatt, P.R. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro, "Designing the User Interface for Multimodal Speech and Gesture Applications," **Proc. of HCI 2000,** vol. 15, no. 4, pp. 263-322, 2000

[3] S. Oviatt and R. van Gent, "Error Resolution during Human-Computer Interaction," **Proc. of ICSLP '96**, vol. 1, pp. 204-207, 1996

[4] L. Nigay and J. Coutaz, "A Generic Platform for Addressing the Multimodal Challenge," **Proc. of CHI '95**, 1995

[5] S. Oviatt, "Mutual Disambiguation of Recognition Errors in a Multimodal Architecture," **Proc. of CHI '99**, ACM Press, pp. 576-583, 1999

[6] L. Festinger, "A Theory of Cognitive Dissonance," **Stanford University Press**, 1957

[7] L. Rigby, "The Nature of Human Error," **Annual Technical Conference Transact. of the ASQC**, 1970

[8] J. Reason, "Modeling the Basic Error Tendencies of Human Operators," **Reliability Engineering and System Safety**, 22, pp. 137-153, 1988

[9] J. Rasmussen, "Skills, Rules, Knowledge: Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models," **Transact. of SMC '83**, SMC-13, pp. 257-267, 1983

[10] B. Schuller, F. Althoff, G. McGlaun, M. Lang, and G. Rigoll, "Towards Automation of Usability Studies," **Proc. of SMC '02**, "Bridging the Digital Divide," Vol. 4, 2002

[11] F. Althoff, K. Geiss, G. McGlaun, B. Schuller, and M. Lang, "Experimental Evaluation of User Errors at the Skill-Based Level in Automotive Environments," **Proc. of CHI '02**, pp. 782-783, 2002

[12] J. Nielsen, "Usability Engineering," **Morgan Kaufmann Publishers, Inc.**, San Francisco, CA, 1999

[13] G. McGlaun, F. Althoff, B. Schuller, and M. Lang, "A New Technique for Adjusting Distraction Moments in Multitasking Non-Field Usability Tests," **Proc. of CHI '02**, pp. 666-667, 2002

[14] M. Zobl, R. Nieschulz, M. Geiger, M. Lang, and G. Rigoll, "Gesture Components for Natural Interaction with In-Car Devices," **Proc. of GW '03**, Vol. 2915, pp. 448-459, 2004

[15] M. Geiger, R. Nieschulz, M. Zobl, and M. Lang, "A Gesture-Based Control Concept for In-Car Devices," **Proc. of USEWARE '02,** VDI/VDE No. 1678, pp. 299-303, 2002

[16] G. McGlaun, F. Althoff, and M. Lang, "A New Approach for Integrating Multimodal Input via Late Semantic Fusion," **Proc. of USEWARE '02,** VDI/VDE No. 1678, pp. 181-185, 2002

[17] G. McGlaun, M. Lang, and G. Rigoll, "Development of a Generic Multimodal Framework for Handling Error Patterns during Human-Machine Interaction," **Proc. of SCI '04**, Vol. I, pp. 523-527, 2004

[18] B. Schuller, "Towards Intuitive Speech Interaction by the Integration of Emotional Aspects," **in Proc. of SMC '02**, "Bridging the Digital Divide," Vol. 6, WA2N1, 2002

[19] P. Morguet and M. Lang, "Comparison of Approaches to Continuous Hand Gesture Recognition for a Visual Dialog System," **Proc. of ICASSP '99**, Vol. 6, pp. 3549-3552, 1999

[20] Project FERMUS, "Error Robust Multimodal Speech Dialogs," **website: www.fermus.de**, 2002