# A GENERIC OPERATION CONCEPT FOR AN ERGONOMIC SPEECH MMI UNDER FIXED CONSTRAINTS IN THE AUTOMOTIVE ENVIRONMENT

Gregor McGlaun[*], Frank Althoff[*], Hans-Wilhelm Rühl[+], Michael Alger[°], Manfred Lang[*]

[*] Institute for Human-Machine Communication, Technical University of Munich, 80290 Munich, Germany
[+] SiemensVDO AG, Philipsstraße 1, 35576 Wetzlar, Germany
°Vodafone Pilotentwicklung GmbH, Chiemgaustraße 116, 81549 Munich, Germany

{mcglaun, althoff, lang}@ei.tum.de, hans-wilhelm.ruehl@de3.vdogrp.de, michael.alger@v-pe.de

## ABSTRACT

In this contribution we focus on the evaluation of a generic operation concept for a speech-based MMI integrated in a car radio and mp3-player. The system is additionally subject to several restrictive economical and geometrical boundary conditions. Nevertheless, the interface has to meet high technical requirements: its handling must be easy to learn, comfortable and, above all, intuitive. In several trial series, two competing generic dialogue concepts (Tree Structure Principle vs. Action-Object Principle) are evaluated. Moreover, designing the layout of the limited display user distraction effects are studied. As a general result, we obtained 88% of the test subjects to be very satisfied with the operation comfort of the final system.

## 1. INTRODUCTION

In this study, the main goal is to design and evaluate a concept for an ergonomic speech-based MMI. Since the application is to be used in mid-range cars or compacts, the design of this system is noticeably influenced by economical as well as geometrical constraints:

- The system is to be placed into the car radio chute. Thus, from the outset, the matrix display is geometrically restricted to two lines of text of at most 16 characters length plus an additional small set of LEDs.
- The system has a speaker independent full word recognizer that is based on an existing cost-efficient low-end hardware module (Hello IC). Therefore, the size of the active vocabulary per command level is restricted to 30-50 words. Yet, shifting to another command level, a new vocabulary set can be reloaded in real-time.

Besides the speech channel, for some functions a haptic input device in form of a small keypad is applied. For providing a multimedia feedback the system has a separate speech output channel. In the evaluation phase the system feedback consists of a simple speech synthesis, sc. text to speech (TTS).

Despite the constraints mentioned above, the operation concept should, as far as possible, be generic, easy to learn as well as interactively explorable, and, above all, intuitive.

## 2. METHODOLOGY

The dialogue behavior of a human being is quite complex and thus difficult to predict. A purely abstract and theoretical data ascertainment without any user feedback cannot meet the requirements of an ergonomic dialogue concept. Therefore, over the whole investigations, repeatedly usability studies are carried out. Most of these tests are based on the Wizard-of-Oz method [Nie99]. Thus, a control panel has to be created with which the wizard can control all functionalities. There is a bi-directional information flow based on a socket communication between the control panel and the application, which is represented in the form of a GUI inside a specially equipped test car (see figure 1). In consideration of the test subjects' impressions evaluated from the questionnaires and videotapes the system prototype is optimized by iterative redesigns.

### 2.1 Preliminary investigations

First of all, existing concepts and basic approaches of existing systems are compared with respect to an ergonomic and intuitive application concept. Based on the results of this basic study, a suggestive functional specification of the system to be developed is compiled. As the active vocabulary of the recognizer is restricted in a baseline study, the most frequent command words the test persons use intuitively for given functionalities are determined. The probands get different tasks concerning the operation of different functions of our application. Commands can be

From: Proc. HCI 2001, Lawrence Erlbaum Ass.

chosen arbitrarily regarding the premise that only one or more command key words are allowed (no natural speech). There are no further instructions, optical feedback nor help function to not influence the test persons.

## 2.2 Analysis and evaluation of two competing generic dialogue concepts

The aim of the second study is to bring out the advantages of each of the two operation concepts, which have been developed on basis of the preliminary investigations. The **T**ree-**S**tructure **P**rinciple (TSP) is adapted from search tree topologies (like in popular operating systems), and contains four main command levels. A detail of the TSP is sketched in figure 2. Changing to another level, a new set of vocabularies can be used. Commands of inactive branches or command levels cannot be recognized by the system.
The **A**ction-**O**bject-**P**rinciple (AOP) is based on an abstract grammatical formalism. All valid commands are, in this steady order, a combination of an action command, an object command, and an optional flag, e.g. "play title 5" (for a detail of the action-object matrix see figure 3). Once an action keyword is recognized, a set of object vocabularies is reloaded.

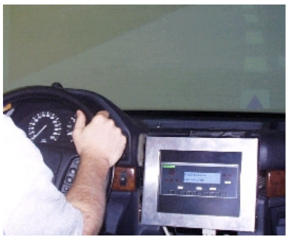


**Fig. 1: Test environment**

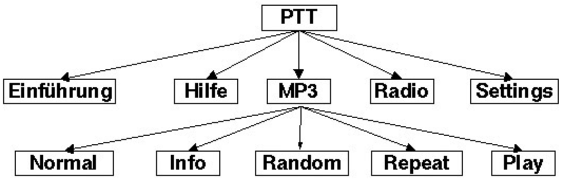
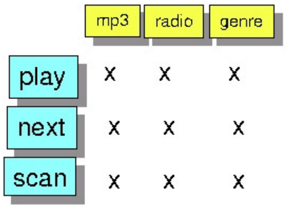


**Fig. 2: Detail of the TSP**



| | mp3 | radio | genre |
|---|---|---|---|
| play | x | x | x |
| next | x | x | x |
| scan | x | x | x |

**Fig. 3: Detail of the AOP**

In both concepts, important functions like “play” can be quickly accessed for an intuitive handling. A single track or song list is addressed by a number (digits must be entered). Each command concept is evaluated in a separate test series. In the trials, the probands are given a short general introduction (no command words nor a structure). Moreover, there is no optical readout feedback. The users have to cope with a set of different operation tasks. To assist them in operation, a three-stage adaptive help function is implemented. Depending on the user’s individual system experience all or a subset of valid command words are resumed. This function is triggered by the keyword “Hilfe”. The system can be individualized for 10 different persons. Entering the system, the user name has to be provided to the system. Personal settings (like the stage of the help function) are saved on exit.

## 2.3 Design and analysis of acoustical and visual system feedback

Concerning the system feedback, three aspects have to be taken into account: what kind of information at which point of time does the user need and how should it be presented? The system gets the information about an MP3-file by reading out the ID3v2-tags [Nil01], which is a standard providing information about the track (artist, title, genre, etc.). As the display field is very small, only the most important pieces of information (e.g. title and artist) can be displayed coevally. On user’s demand, further information in this context is provided. For displaying texts of more than 16 elements, five different readout types are evaluated: terms are displayed as they are, single terms are abbreviated (AT readout), sections of the readout are displayed one after another (**C**yclic **P**aging, CP readout), the display text is shifted from left to right and back (**R**ight-**L**eft-**O**scillation, RLO readout) or it is passed through the readout (**C**yclic **R**un-**T**hrough, CRT readout). Every time the user changes to another level of the tree structure, depending on his settings all or a subset of keywords in this level are automatically displayed. Moreover, there are two LEDs showing global settings (e.g. repeat or random mode). Besides, an optical feedback is implemented: the recognized commands are displayed for a freely adjustable period (default: 2 sec.), then the display switches back to the anterior readout. In the final application, the driver’s attention will be primarily targeted on the traffic. Thus for safety reasons, his visual workload must not be taken up too much by the display. For this purpose, a context-sensitive acoustic feedback information is implemented. If the action to be achieved with the command is per se acoustical (e.g. command: “play title”, action: track is played) no additional feedback is necessary. Else (e.g. given the command “repeat all”), an acoustical feedback (e.g. repetition by the system or a beep) is put out. In the test series, several combinations of optical and acoustical feedback configurations were evaluated. Moreover, to simulate realistic conditions, the probands have to perform in a simultaneously running driving simulation (the so-called attention task). To ensure the test persons take the attention task seriously, the number and the intensity of deviations are summed up in an error score.

## 3. RESULTS

The evaluation of the baseline study (21 probands) showed that sometimes, two terms of an important function denoting the same action were specified with almost the same cumulative percentage, e.g. “play” (47%) and “spiele” (38%). In those cases, for an intuitive handling two commands were mapped on the same action. As another result, analyzing the sequence of the commands, different navigation strategies could be found which helped to build an intuitive command structure. All except for one applied the commands in a tree hierarchy. It was outstanding that 79% used command combinations of exactly two commands. Analyzing the structure, in 85% of all cases a partition into an action and an object could be detected. In passive avoidance of errors, some functions found out critical to be operated by speech (e.g. the setting of volume) can now only be performed by hard-keys. For the most important functionalities (on / off, play / stop etc.), also hard-keys are available as a haptic fallback solution. As at first many of the test subjects were confused about not having a complete concept on how to use the system, it became important to provide some information at start-up. This information should be put across acoustically (55%), by an instruction manual (35%) or solely on the display (10%). At that time, there was no system feedback after a command. A group of 94% claimed a help function, and a kind of system feedback, respectively, where 55% favored a both acoustical and optical, 32% a purely acoustical, and 13% an exclusively optical echo.

For each command concept (TSP and AOP, respectively), 10 persons were tested, 4 participated in both trials for a comparison. The TSP was rated positive by 70% of them. They pointed out that all command words and the structure were self-explanatory (most of them known from entertainment electronics) and easy to memorize (mark 2.8 for reduced, 3.7 with full functionality, where 1 means “very good” and 6 means “very bad”). Yet, the mixture of German and English command words has been criticized by 30% of the test persons. 90 % found that all functions could be easily controlled by speech. 30% of the users declared they would need an instruction manual on any account (e.g. an overview over the tree diagram). Most of them (90%) rated the help function very effective, some (60%) would prefer additional help texts on the display. Yet, most of the test persons (80%) would like to cancel the help function having received assistance they needed. Some probands mentioned that the AOP command structure is, with respect to some commands (e.g. “play random”), unfamiliar and too complex for an intuitive access due to the rigid grammatical order of the commands. The most frequent substantiation was that the users only had known the icons, but had not made up their mind about the intrinsic spoken commands.

A total of 15 test subjects attended the trial concerning the optical and acoustical feedback of the system. Although some of the probands found the RLO / CRT readout more informative than the CP, 87 % of the test persons pointed out that, with regard to RLO / CRT, the permanent movement in the peripheral field of vision is very distractive and bothering while driving the car in the attention task. These subjective impressions were objectively emphasized by the evaluation of the video recordings showing the distribution of the users’ areas of interest (AOIs). Hence, the Relative Glimpse Retention Periods (RGRPs) per task could be computed by the formula

$$RGRP_{display} = GRP_{display} \; / \; TPT,$$

where $GRP_{display}$ is the period of time in which the probands’ AOI was focused on the display and TPT means the total processing time the test subjects needed to cope with the task. The evaluation of the users’ error rates in comparable trial parts revealed that there were made more mistakes with the RLO / CRT display (87%). This underlines that RLO / CRT - because of the distraction effect on the driver - should not be implemented for security reasons. A detailed report on this study we will describe elsewhere. The combination of CP and the context sensitive optical and acoustical feedback once a command had been understood was rated most efficient and helpful (80%). Many of the users (73%) would have liked to stop or skip the display readout (e.g. play lists) by button or PTT. In some special cases (e.g. remaining play time of a track), some icons or a graphical visualization (ray etc.) was preferred rather than text (33%), whereas 53% rated it explicitly unnecessary.

## 4. CONCLUSIONS

The final operation concept was based on the structure of the TSP command levels. We finally applied the AOP to a reduced set of plausible command combinations. 35 probands participated the closing test for the evaluation of the whole concept. The acceptance rate (marks “good” and “very good”) was 88%.

## REFERENCES

[Nie99] Jakob Nielsen, Usability Engineering, Morgan Kaufmann Publishers, Inc., San Francisco, California, 1999, pp. 93-114.

[Nil01] Martin Nilsson, ID3 homepage, http://www.id3.org/, April 2001