Effects of Shortening Speech Prompts of In-Car Voice User Interfaces on Users Mental Models

Julia Niemann[1], *Kati Schulz*[1] , *Ina Wechsung* [1]

[1] Deutsche Telekom Laboratories, Quality and Usability Lab, Berlin Institute of Technology

julia.niemann@telekom.de, ina.wechsung@telekom.de, schulzkati@gmx.net

Abstract

Shortening speech prompts reduce attention allocation display; but might on the other hand deteriorate the users' mental model. Thus the paper investigates effects of reducing time effort of speech via a transfer task, retrieval tasks and navigation-orientation tasks for three different strategies: (1) using earcons for menu orientation, (2) using commando based speech for interaction options, and (3) using uptempo speech for content based information. Results show that earcons are well qualified to not impair navigation-orientation performance. Commando based speech leads to even better retrieval performance than sentence based representation of interaction. Solely uptempo speech decreased retrieval performance.

1. Introduction

Driving is mainly a visually and motor demanding task. Visual distraction causes most of the accidents [17]. Thus and in line with multiple resource model [16], interacting with in-car systems via speech compared to visual-haptic interfaces provides a lower interference with driving. However, speech as an interaction modality can also be cognitive distracting and leads to lower driving performance compared to just driving [11]. Developers of Voice User Interfaces (VUI) should thus not only provide a hands- and eyes free interaction but also enable a mindful conduction of the driving task. More evidence for necessary improvements regarding non-distracting speech output is shown by [10]. They observed that drivers tend to take their eyes more often from the road to look on the display of a speech dialogue system (SDS) compared to a SDS without a display. Apparently despite of disregarding the driving task it is "easier" for the driver to receive information via the graphical user interface (GUI) instead of just listening to the speech prompts. In a driving simulator study [12] found that the percent dwell time (PDT) on display could be reduced by reducing the time effort of speech prompts. These were shortened by using earcons, commando based speech and uptempo speech. The present study deals with the problems and benefits of shortening speech prompts. The benefit is obviously the reduced time effort. But if shortened prompts allow the driver to build an equally good mental model compared to extended speech prompts has so far not been investigated.

2. Shortening Speech Prompts

In the following the strategies to shorten the speech prompts will be reported [12]. Different information and communication applications (e.g. email, news) were implemented on a smartphone. These applications should be usable while driving. Thus automatic speech recognition is activated by pressing the push-to-talk button on the steering wheel. The system gives feedback (direct prompt) after users' input which is limited to information regarding menu orientation (land marking) and content information (e.g. email header in the inbox). To enable users receiving all information presented on the screen also acoustically the push-to-talk button can be pressed longer (longterm push-to-activate = LPA). This way, the driver can request the information at any time which is especially relevant for parallel tasks as the infotainment task can be interrupted by the driving task. In addition, by reading all information (land marking, content information & interaction options) to the user, eye gazes on the display should be avoided. However, as shown in [12] the PDT is only reduced by providing all information and simultaneously shortening the speech prompts.

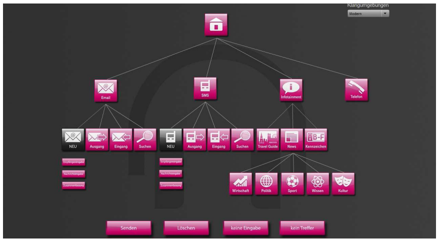
2.1. Shortening of interaction options

Interaction options for VUIs are speech commands with which users initiate the next interaction step. The user needs to know which commands he/she can use since natural language systems still need to be improved. Additionally, for in-car infotainment systems interacting via commando-based speech is less distracting than natural speech [7]. Furthermore a rich grammar of the speech recognition system increases the "no match" rate. An often employed strategy to reduce speech prompts telling the user possible commands is the so called "speak-what-you see" principle. Here, speech commands are displayed on the GUI. To present the information acoustically often meta-information is used; e.g. "if you want to answer the mail, say answer". The meta-information helps the user to differentiate between content information, land marking and system feedback. However, this is very time consuming. As a solution we suggest employing different voices replacing the meta-information. Only speech commands will be read (like a list) to the user by a different voice than all other information. Thus, we replaced sentence-based representation of interaction options by commando-based speech and enriched this information by using different voices.

2.2. Shortening of land marking

To avoid disorientation systems should give users feedback about their current position in the menu. Menu orientations help the users to build a conceptual representation of the generic and narrow menu steps. If the prior knowledge fits the actual conceptual representation a faster learning performance can be expected [14]. To shorten the land marking information we used earcons instead of speech. Earcons are short musical motives representing menu items. They differ in the combination of rhythm, pitch, timbre, register and dynamics [4]. Earcons can build up menu hierarchies [4]. Every menu item on our start screen (email, sms, telephone and infotainment) has its own sound motive (timbre). The 4 sound motives exhibit a high differentiation. To represent the single steps of the main interaction path (e.g., writing an email) within the specific menu items, we varied the menu item sounds by changing the pitch. The pitch rises with every step nearer to the end of the operation. The higher pitched sound is added to the sound of the previous interaction step. According to [4] "icons can present information in a small amount of space compared to text; nonspeech sounds can present information in a small amount of time as compared to speech". Another strategy to label menu steps with nonverbal sound is the use of auditory icons. Auditory icons are metaphorical representations of a word or a concept [6]. Benefits of auditory icons are the strong identifiably and intuitive link with the word that is presented. Consequently the learning rate is lower compared to earcons [3]. However they are not as suitable for building up menu hierarchies as it is usually hard to find natural sounds for all menu steps that are semantically highly linked with the represented step. In Figure 1 the menu hierarchy of the implemented applications is shown. For email and sms we used earcons, for all infotainment applications and the telephone application auditory icons were introduced. This was to evaluate which sounds will be more qualified.

Figure 1: *Schematic menu hierarchy*



2.3. Shortening of content

Next to land marking information and interaction options also content has to be presented acoustically. For example in the emails application the email-header has to be read to the user as well as names or addresses he entered. By pressing the LPA to re-read this information has already been heard. To shorten it we suggested speeding up the output. The speech synthesis allows a manual setting of the playback velocity. We set it at 30% (value determined with a pre-test including 5 users).

3. Effects on Users' Mental Models

In average a decrease of 9 sec. of the speech prompt time was achieved employing all three strategies simultaneously which as mentioned before led to lower PDT on display. But since "speech is the most semantically rich acoustic medium" [5], shortening and therefore exclusion of speech could result in information loss. The present experiment investigates to which extent the users' knowledge and the learnability of the system is decreased through shortening. The users' knowledge of a system is represented in his/her mental model. Mental models contain humans' structural analogies of the world [9]. In the context of HCI the mental model can be seen as the conceptual representation of a device. Due to the human cognitive resource restriction mental models are incomplete [13] and simpler than the actual situation or problem [9]. The developer of a device should keep in mind what mental model he/she wants the user to build of the system. According to [13] the users' mental model will be build upon the system image which communicates the designers' model to the user. The system image contains the user interface among training and manual. For our VUI we wanted to achieve a mental model resulting in quick learning about what to say and facilitating understanding of the menu structure. Our system as mentioned before is a commando-based, hierarchical orientated menu systems including broad as well as deep structures. Therefore menu-orientation and learning of the speech commands are essential aspects. However, we did not expect the users to know in detail the speech prompts and to know the exact menu labels since mental models are not expected to be overly detailed. Hence, we assume that the shortening (cf. Sec.2) will not result in a worse mental model. The shortening of interaction option via commando based speech is even expected to be more efficient since all relevant information is still presented while irrelevant information, the meta-information, is excluded. Please note, that we did not expect the users to learn the exact wording of the prompts. For the use of earcons an improvement is not necessarily expected. Earcons' semantically link compared to speech is determinately reduced. However, again we assume that the relevant information is still present by the hierarchical character of earcons and that they are more qualified for land marking than auditory icons. To prove these assumptions we had to measure the abstract mental model of the users.

4. Evaluation

48 subjects (19f, 29m, age ø=27y., SD=6) were tested. First a training phase was completed. This was followed by a transfer task, assessing the effects of all shortening strategies. Transfer tasks, requiring the transformation of abstract problem solving knowledge to new tasks, are useful to assess the quality of mental models [8]. Then retrieval and navigation-orientation tasks had to be performed to test specific effects of the shortening strategies on the mental models.

4.1. Transfer Task

For the transfer task we implemented three system versions on an Android-based smartphone (HTC Desire) each with different shortening strategies. With the system initiated direct prompt a speech based land marking was presented (e.g. "inbox" or "main menu"). By using the LPA an extensive prompt was given (including land marking, content information and interaction options). The system versions varied in the way the outputs were shortened (cf. Table 1). For system 1 no shortening strategies were used.

Table 2. *Independent Variable*

Shortening strategies	Independent Variable			
	System 1	System 2	System 3	System 4
Land marking (non verbal sounds)	Long	Long	Short	Long
Content (uptempo speech)	Long	Long	Long	Short
Interaction options (commando based speech)	Long	Short	Long	Long
N	12	12	12	12

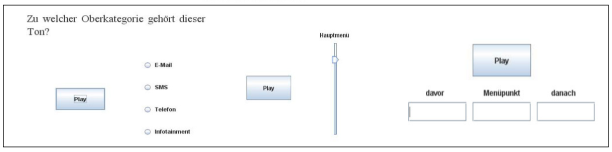
Participants were equally allocated to the 4 groups. The display was covered during the whole experimental phase simulating the dynamic driving situation requiring the visual attention allocated on the road. The subjects started in the main menu and first passed through an email task. Participants had to find the third mail in the inbox, read out the mail and answer by recording an audio file. For every interaction step they were asked to use the LPA to hear the extensive speech prompt with the respective shortening strategies. Subsequently they conducted the transfer task. Afterwards the participants again started in the main menu and were asked to search a sms in the inbox and let it read out by the system. Using the LPA was not allowed. Usability parameters as efficiency and effectiveness were measured. This way the number of speech commands errors and the mean reaction time after each interaction step were collected.

4.2. Retrieval and Navigation-Orientation Tasks

Testing commando based speech. For testing effects of commando based speech on learnability the subjects trained with system 1 and 2 were asked to indicate the speech commands (interaction options) of the email task in a multiple choice test. Recognition (e.g. in a multiple choice test) is a different process than free recall [1]. Thus we avoided to ask for specific wording and rather tested if the subject retrieved the speech prompts and attributed them to the right menu step (e.g. "Was the option "next" available in the email inbox? Yes/No"). Correct answers and reaction time were measured.

Testing uptempo speech. After testing the commando based speech all test subjects were allocated to two groups. It was ensured that the same amount of participants of each previous group was assigned to the new group. The trials for testing uptempo speech and nonverbal sounds were randomized to avoid learning effects. Five email headers were read after each other to the participants. For 24 of the subjects the headers were read out with a 30% faster velocity. The other 24 subjects received the same email headers but with normal speed. Again a multiple choice test was presented.

Figure 2. *Navigation-Orientation Tasks*



Zu welcher Oberkategorie gehört dieser Ton?

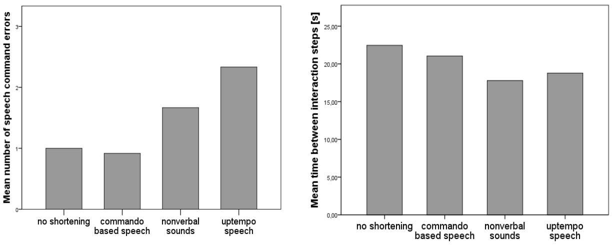
Testing nonverbals sounds. At first a training phase was conducted. One group heard the hierarchical order of the menu structure with the speech-based labels of every menu step. The other group heard the sound label in the hierachical order via earcons and auditory icons. As mentioned before the email and sms applications were presented by earcons while infotainment and telephone application were presented by auditory icons. The earcons included hierarchical information compared to the auditory icons. Also for the speech condition the menu labels for email and sms were hierarchical based since the superior category was always stated (e.g "email inbox", "sms inbox") while for the infotainment application only the label itself was named ("sport").

After one time listening to the menu order (by speech or nonspeech) we presented a sample of the labels to both groups and asked to name the superordinated category (cf. Fig. 2). Then we asked to define the menu depth by an abstract visual scale (cf. Fig. 2). According to [14] users' spatial representation of the menu can be measured by imaging the menu hierarchy on a vertical one dimensional scale. Subjects were asked to indicate the menu position on this scale. At last the menu point after and before and the menu step itself should be recalled.

4.3. Results

Transfer Task. We expected no impairments in performance data for the transfer task for all three shortening strategies. If anything, we assumed the commando based speech having positive effects. An oneway ANOVA was conducted for the time between two interaction steps. The time between two interaction steps includes the duration of the direct prompt and the time till the push-to-talk button (to activate the speech recognition) was pressed. Since the duration of the direct prompt was equal for all 4 groups we did not analyze it. The difference arises from the time participants need to react on the prompt. A significant effect was found for time between two interaction steps (F(3,41)=2.73, p=0.057) as well as for number of speech command errors (F(3,43)=3.25, p=0.031). However, the Bonferroni test showed no significant differences for reaction time which is likley due to the conservative justification of the test. A tendency (p<.10) for duration was shown for nonverbal shortening (sounds) compared to no shortening (indicating an improvement for the systems with sounds instead of speech). A significant difference (p<.05) of speech command errors of uptempo speech and no shortening was observed. The uptempo speech group made significantly more speech command errors than the non shortening group (cf. Figure 3).

Figure 3. *Number of speech command errors and reaction time*



Shortening by using commando based speech. No shortening (sentence based speech) and shortening (commando based speech) were compared. An independent t-test was calculated for correct answers of the multiple choice test and duration to make an answer. No significant effect for the number of correct answers was found. But the group with the commando based speech was significantly faster (M = 13.40) compared to sentence based speech (M = 19.18, *t*(12.20) = 2.24, *p* = .044).

Shortening by using uptempo speech. Shortening by using uptempo speech had no significant effect on the number of correct answers (group no shortening: M = 2.63, SD = 1.66, group uptempo speech: M = 2.43, SD = 1.41) as well as on the duration to set the answer (group no shortening: M = 9.44, SD = 3.00, group uptempo speech: M = 9.44, SD = 3.06).

Shortening by using nonverbal sounds. To test the effect of using nonverbal sounds instead of speech the rate of correct answers for the analogue scale and naming the menu steps before and after (generic and narrow menu steps) were compared. The difference between sounds and speech as well as the difference between hierarchical information or non hierarchical was investigated. Also correct answers of naming the superior category for the different sounds (auditory icons = non hierarchical vs. earcons = hierarchical) were analyzed. Therefore we conducted a two-way mixed ANOVA with hierarchical information as the repeated measure factor. But first we tested the presupposition that auditory icons are easier to remember than auditory icons since the semantically link is stronger.

Correct naming. A paired t-test was conducted. It was shown that the difference between the two groups was significant. The menu labels presented via auditory icons were significantly better remembered (t(21) = 3.87, p = .001).

However, we expected that the positives effects for auditory icons compared to earcons will be adjusted if different aspect of the mental model will be tested (e.g. spatial representation of the menu or order of the steps). Accordingly we also expected that the benefits of semantically highly linked speech will be adjusted.

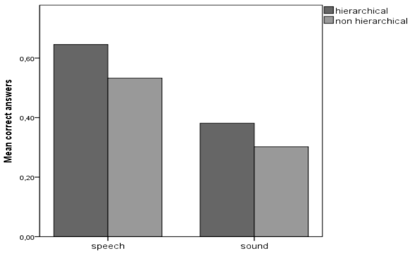
Naming superior category. A paired t-test for comparing auditory icons and earcons by naming the superior category was carried out. Neither a significant main effect and nor a significant interaction effect was found.

Analogue scale. No main effect hierarchy was found (F(1,459 = 1.48, p= .230). Also the main effect for shortening (speech vs. sound) and the interaction effect (F(1, 45) = 1.53, p = .469) were not significant. Although a tendency in the expected way was shown.

Naming menu steps before and after. A significant main effect for shortening (*F*(1,45)= 12.06, *p*= .001) as well as a significant effect for hierarchy (*F*(1, 45) = .87, *p* = 019) was found. The interaction effect was not significant. Means are shown in figure 5.

A t-test between non hierarchical speech and the hierarchical sounds (earcons) on the number of correct answers showed that the difference between sounds and speech was not significant anymore.

Figure 5. *Correct answers for naming menu steps before and after*



5. Discussion

A reduction of the time effort (shortening) of speech prompts inhibits allocating visual attention away from the road to the system [12]. Shorting was achieved by using sounds for land marking information, uptempo speech for content information and commando based speech for menu options. Since shortening is linked to information loss the present experiment tested the effect of the shortening on users' mental models. Since mental models are rather conceptual than detailed, we measured the effects on the mental model via a transfer task, multiple choice tests (instead of free recall) and the mental spatial representation of the menu structure. Additionally participants were asked to name generic and narrow menu steps as well as the superior category of menu steps. For the transfer task only shortening by uptempo speech had negative effects on the performance data. Sounds instead of speech had even a positive effect. Shortening of interaction options via commando based speech led not to an increased recognition performance in terms of correct answers but to a faster recognition time. This positive effect on users' mental model is inline with findings that irrelevant information interfere with recognition [15] and the word-length effect [2]. Earcons led to a worse recognition of the exact label they are presenting than auditory icons. Asking for the speech-based menu labels is not meaningful as these are 100% semantically linked. However, as developers we did not want the user to know every menu step by its correct label we would rather have a good overall representation of the menu hierarchy and structure. As the earcons we used included hierarchical information we expected that the positive effects of a high semantic link (non hierarchical speech or auditory icons) will be adjusted. This was shown in tendency for navigation-orientation tasks. Table 2 gives an overview of the amount of time which is reduced by using the single shortening strategies as well as the effects on the mental model. Also recommendations are given for shortening speech for in-car VUIs.

Table 2. *Recommendation based on time effort and effects on the mental model*

Effort	Shortening of…		
	Interaction options by commando based speech	Land marking by earcons	Content by uptempo speech
Time reduction	++	+	+
Effects on mental model	+	0	-
Recommendation	++	+	0

6. References

[1] R.C. Atkinson & J.F. Juola, Search and decision processes in recognition memory. In R.C. Atkinson, et. al. (Eds.), *Contemporary developments in mathematical psychology: Learning, memory and thinking*, Freeman, San Francisco, 1974.

[2] A.D. Baddeley, N. Thomson, & M. Buchanan, Word length and the structure of short-term memory. *Journal of Verbal Learning & Verbal Behavior, 14*, 575–589.1975

[3] T.L. Bonebright, & M.A. Nees, Memory for Auditory Icons and Earcons with Localization Cues. In Proc. ICAD 2007, 2007.

[4] S.A. Brewster, Non-speech auditory output. In: J.A. Jacko, & A. Sears (Eds.), *Human-Computer Interaction Handbook*. Lawrence Erlbaum, Mahwah, 220-239, 2002

[5] S. Garzonis, S. Jones, T. Jay, et al., Auditory icon and earcon mobile service notifications: intuitiveness, learnability, memorability & preference, In Proc. CHI 2009, 1513-1522, 2009

[6] W.W. Gaver, Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction*, 2(2), 167-177,1986

[7] R. Graham, L. Aldridge, C. Carter, & T.C. Lansdown, The design of in-car speech recognition interfaces for usability and user acceptance. In: *Proc. EPCE* 1998. 1998.

[8] F.G. Halasz, & T.P. Moran, Mental models and problems solving in using a calculator, In *Proc. CHI'83* , 212-216. 1983

[9] P.N. Johnson-Laird, *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, Harvard University Press, 1983

[10] A.L. Kun, T. Paek, Z. Medenica, et. al., Glancing at personal navigation devices can affect driving: Experimental results and design implications. In *Proc. AutomotiveUI* 2009, 2009

[11] J.D. Lee, B. Caven, S. Haake, T.L. Brown, Speech-based interaction with invehicle computers: the effect of speech-based e-mail on drivers' attention to the roadway. *Human Factors, 43*, 631–640, 2001

[12] J. Niemann, A. Naumann & F. Oberle, Entwicklung eines nutzerzentrierten Sprachdialogsystems im Fahrzeug. In Proc. *USEWARE 2010*, VDI, 107-119, 2010

[13] D.A. Norman, Some observations on mental models. In D. Gentner,& A.L. Stevens (Eds.) *Mental Models*. Lawrence Erlbaum, 7-14, 1983

[14] Totzke, I., Schmidt, G. & Krüger, H.-P. Mentale Modelle von Menüsystemen - Bedeutung kognitiver Repräsentationen für den Kompetenzerwerb. In M. Grandt (Ed.), *Entscheidungsunterstützung für die Fahrzeug- und Prozessführung*, 33-158, 2003.

[15] R.Vilimek, *Gestaltungsaspekte multimodaler Interaktion im Fahrzeug. Ein Beitrag aus ingenieurpsychologischer Perspektive*. PhD Thesis, Universität Regensburg, 2007

[16] C.D. Wickens, Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*, 159-177, 2002

[17] W. Wierwille, & L. Tijerina, Eine Analyse von Unfallberichten als ein Mittel zur Bestimmung von Problemen, die durch die Verteilung der visuellen Aufmerksamkeit und der visuellen Belastung innerhalb des Fahrzeugs verursacht wird. *Zeitschrift für Verkehrssicherheit,* 164-168,1995.