Effects of Speech and Non-Speech Sounds on Short-Term Memory and Possible Implications for In-Vehicle Use

Research paper for the ICAD05 workshop "Combining Speech and Sound in the User Interface"

Roman Vilimek and Thomas Hempel

Siemens AG
Corporate Technology
Information & Communications
Otto-Hahn-Ring 6
81730 Munich, Germany

roman.vilimek.ext@siemens.com
thomas.hempel@siemens.com

ABSTRACT

Using auditory output for presenting non-critical but relevant events to the car driver, we compared the effect of four groups of sounds (two speech, two non-speech) on short-term memory and on response time and accuracy. The results indicate that longer speech messages can disrupt short-term memory performance whereas earcons, auditory icons, and single keywords do not cause this effect. Earcons, in turn, lead to comparatively long response times. Based on these experimental data, the suitability of such stimuli for in-vehicle representation is discussed. The type of experimental set-up may enable transfer of the results to comparable settings.[1]

1. MOTIVATION

With the emergence of more and more driver assistant systems, driver information systems, comfort functions and the integration of nomadic devices in modern vehicles, the question of how to transmit messages of these systems adequately to the driver obtains increasing priority. Today, in many cases information is given visually only. As most of the driver's visual attention is consumed by monitoring the traffic, using the auditory modality in addition is a widely accepted approach.

Among the characteristics of auditory system output both its omnidirectional nature and the capacity to capture attention even while a person is otherwise engaged are most prominent when compared to visual output. The latter aspect can be exploited particularly useful in the design of warning sounds. The same rationale holds if the purpose of sounds in system design is not warning, but notification of non-critical though not negligible events. If a sufficient signal to noise ratio is provided, it can safely be assumed that information presented auditorily will be noticed by the driver and thus contributes to reducing demands on visual attention. However, in contrast to warnings, non-critical informative sounds must reside in the background and inform the user in a non-distracting manner.

[1] The experiment reported in this paper has been conducted in cooperation with Prof. Dr. Alf Zimmer, Department of Experimental Psychology, Regensburg University (Germany) and is part of the first author's dissertation research.

2. UTILIZING AUDITORY OUTPUT

One of the fundamental decisions to be made in the design of a user interface for the auditory channel is the one between speech and non-speech sounds. The most established design options for speech are fully verbalized speech messages versus short prompts using only keywords. To present information non-verbally, two techniques have emerged in the late 1980s: *earcons* and *auditory icons*. Earcons [1] can be described as non-speech audio messages consisting of abstract, musical tones that are used in structured combination. As there is no intuitive link between an earcon and what it represents, earcons have to be learnt. Brewster and colleagues [2], [3] showed, that they are an effective means of communicating information in sound. In contrast to earcons, auditory icons are natural, everyday sounds that are recorded in the environment (e.g. slamming a door) and mapped to system events by analogy [4]. The advantage of auditory icons is that only a minimal learning effort is required to understand the connection between sound and the to-be-represented object. A clear disadvantage is of course, that abstract interactions and interface objects cannot be represented by a natural sound.

The decision to use speech or non-speech sounds in a user interface is subject to design considerations concerning optimization criteria (see [5]) most likely for single task settings. But in dual or multiple task scenarios like driving a car, another important aspect has to be kept in mind: Auditory system output should interfere as little as possible with cognitive processes supporting driving the car. Considering Michon's [6] hierarchical structure of the driving task, the strategic level – which is preparatory to the maneuvering and operational level – consists of general planning activities. One of the most prominent cognitive demands in this respect is the processing of information in short-term memory, as it serves to maintain temporarily relevant route information. If it comes to the design of sounds for non-critical information, great diligence has to be exercised in creating instructive, nevertheless non-distracting auditory cues that do not disturb short-term memory performance. This can get quite challenging as even unattended sounds at low sound pressure levels can disrupt recall from short-term memory [7].

3. SOUNDS AND SHORT-TERM MEMORY

In a comparison of speech and non-speech sounds for interface design, Brewster [5] points at the so called "unattended speech effect", also known as "irrelevant speech effect" (ISE) [8], [9], which is brought forward as an argument for using non-speech sounds. Speech in the background causes information to drop out of short-term memory, although subjects in these studies were told that the speech in the background is completely irrelevant to the task and should be ignored.

If only speech causes this effect and non-speech audio does not, this would be a clear argument for the use of non-speech sounds. A number of experimental studies show that this is not the case. Interestingly, it is not important to understand the irrelevant speech in the background: For instance, Jones, Miles, and Page [10] showed that monolingual English subjects were disturbed by the same amount if they were exposed to English, Welsh or even reversed speech. Studies by Salamé and Baddeley [11] demonstrated that the effect of unattended vocal music is equivalent to unattended speech. In the case of instrumental music, the effect was present, but less marked. It seems that non-speech sounds might as well have a negative influence on recall from short-term memory.

Further evidence for this is provided by Jones and Macken [12]: They showed that irrelevant sounds also produce the irrelevant speech effect. Banbury and colleagues [13] summarize the necessary conditions for non-speech sounds to produce an "irrelevant sound effect": The main causal factor for disruption is not speech itself, but acoustic change. This change may be manifested by changes in pitch, timbre, or tempo. A change in sound level as well as a simple repeating of sounds, tones, or utterances is not disruptive. The magnitude of the effect of sounds is equal to the effect of speech when non-speech sounds are equated in terms of their acoustic variation.

To summarize, there is convincing evidence that the irrelevant sound effect depends on physical characteristics of the sound and not on its semantic, lexical, or associative content. However, it is unclear in how far sounds as short as typical earcons or auditory icons produce this effect.

4. EXPERIMENT

Auditory system messages clearly are not meant to be irrelevant sounds, because users have to decide in which way to respond to them. Nevertheless, as mentioned above, as long as the sounds are not time-critical warning sounds the disruption of cognitive processing should be kept to a minimum. This holds especially for information messages with lower priority in multiple task settings. To assess to what extent verbal messages, earcons and auditory icons have detrimental effects on short-term memory performance, we conducted an experiment using a modified version of the standard irrelevant sound paradigm.

The ISE standard paradigm as described in [9], [12] consists of a simple short-term memory task in which subjects are visually presented with randomly permuted items like digits or letters one after another. After a retention interval of several seconds, the presented material has to be reproduced in the correct serial order. The background sounds can be presented during presentation of the items, during the retention interval, or both. Experimental participants are asked to ignore any sound they hear and to concentrate on the memory task, which means memorizing the material and rehearsing it during the retention interval. In a baseline condition, the memory task is executed without background noise, i.e. in silence. The dependent variable is the number of errors in serial recall under each sound condition.

Not all memory tasks are susceptible to the irrelevant sound effect. It leads to significant differences only in memory tasks demanding the retention of serial order information. But exactly this component is very important in many everyday mental activities [14].

4.1. Overview

Since it was not the goal of the study to conduct an irrelevant sound experiment in the narrower sense, we used a modified version of the standard paradigm to investigate the influence of task-relevant speech and non-speech sounds on short-term memory performance and time needed to understand (decode) the audio message. More detailed, we used the framework of the serial recall short-term memory task and inserted a choice reaction task in the retention interval to study the effect of different types of auditory output (speech and non-speech sounds) on response times on the one hand and its effect on memory performance on the other hand.

The memory task itself was identical to the ISE paradigm. But instead of instructing our participants to ignore the sounds during the retention interval, we told them to try to confirm a presented sound by pushing the corresponding button as quickly as possible but nevertheless keeping the digits in mind. More precisely, they were instructed to treat the memory task (i.e. the rehearsal of the series of numbers) with highest priority and regard the pressing of buttons as only of secondary importance. In this way a dual task setting was established reflecting real-life situations in which subjects had to maintain situation-relevant information in memory while at the same time receiving auditory system messages. The sounds presented to the participants as stimuli were either earcons, auditory icons, long speech messages, or keywords in order to allow for comparisons between speech and non-speech sounds.

The data on response times and number of errors in the choice reaction task should not be interpreted as corresponding only to situations in which buttons have to be pressed after hearing a signal. The reaction data in this experiment are collected to provide information on the time needed for a person to perceive an auditory event, process it while being mentally engaged in another task and react to it. In this respect, pressing the button serves as a means of reporting the end of cognitive processing, i.e. decoding the semantic information of an auditory event.

Due to the fact that the task of pressing a button is the same in every experimental condition, the amount of distraction produced by selecting and pushing the button is held constant. Together with the instruction to allocate most attentional resources to keeping the digit sequence in mind, these preconditions allow us to attribute changes in performance in the memory task to properties of the sounds used to indicate the target button. Thus, we were able to compare the influence of speech, earcons, and auditory icons on short-term memory directly and to answer the question whether or not earcons or auditory icons lead to effects similar to the ISE.

4.2. Participants

The study involved fifteen participants aged between 24 and 50 years (mean age: 28 years), which were monetarily compensated for their participation. None of them reported any hearing-related problems.

4.3. Design and Tasks

The different auditory system messages were compared in a within-subjects design. The independent variable was type of auditory system output with five levels: long speech message, keywords only, earcons, auditory icons, silence.

The serial recall task was equivalent to the standard ISE paradigm, as described above. Subjects had to memorize a sequence of 9 digits in the correct order and to reproduce this sequence after a retention interval. This procedure was held constant in all experimental conditions. The dependent variable was the number of errors in serial reproduction.

In the standard ISE paradigm, the experimental conditions are defined by presenting sounds versus silence during the retention interval. In the present study, the sound condition in the choice reaction task differentiated between experimental and control condition. Earcons, auditory icons, long verbal messages, or single keywords indicated in the experimental conditions, which one of the buttons on a board equipped with push buttons was to press. For each of the four groups of sounds four different stimuli were available for playback. Accordingly, four push buttons with small pictograms in front of them (referring to the different four stimuli per group) were mounted on this board. As the experiment was conducted in the context of in-vehicle interaction research, the pictograms showed symbols of a seat belt, a battery, a speedometer, and a gas pump. Speech and non-speech sounds were designed to represent these symbols in different degrees of abstraction. A description of the corresponding four different sounds for each condition is given in section 4.5. Silence, i.e. no auditory output served as baseline for the serial recall task. In this condition, subjects were given visual cues (the corresponding pictograms) for the push-button task.

For this choice reaction task (CRT), subjects were told to press the correct of the four buttons as fast as possible, without neglecting the memory task. The dependent variables were the number of correctly pressed buttons and the respective response times. Response times and errors in the silence condition (the baseline condition for the short-term memory task) were not collected, as the comparisons of interest in the CRT are not between auditory and visual presentation of stimuli, but between auditory stimuli only. Thus, there is no special baseline condition for the CRT, speech and non-speech auditory conditions will be analyzed in planned comparisons.

4.4. Procedure

The procedure of each experimental trial was as follows: The participants were sitting in front of the computer screen wearing headphones. A fixation cross was displayed in the middle of the computer screen for two seconds. Afterwards a sequence of nine digits was presented, one after another. The numbers were randomly drawn without replacement from the set of integers ranging from 1 to 9. They were displayed for 800 msec each with an interstimulus interval of 200 msec, which are typical presentation durations for serial recall tasks. Subjects were instructed to memorize the exact sequence and rehearse it carefully.

Subsequently, an arrow appearing in the middle of the screen initiated the retention interval by pointing towards the keyboard for the pushbutton choice reaction task. Depending on the experimental condition, subjects then either heard an earcon, auditory icon, a long verbal message, or a keyword indicating which button to press. No visual cue was given. In the case of the long speech message, subjects were told that only the first word of this speech message (which served as keyword) is relevant and that they may ignore the rest of the text. In the baseline silence condition no auditory cue, but a visual cue designated the target item.

This choice reaction task was repeated once in every conditon, thus confronting the participants with two pushbutton tasks in one retention interval. The second choice reaction task within each trial was of the same auditory condition like the first one (earcons, auditory icons, long speech message, keywords or silence, respectively), but always involved pressing a different button. Altogether, the duration of the retention phase added up to 15 sec.

Following this task, a 3 × 3 number pad with clickable digits appeared on the screen. The digit layout corresponded to the arrangement keys on a PC's numeric keypad (without zero). Using the mouse, the digits 1 to 9 had to be entered in the remembered order.

At the beginning of the experiment, subjects received two practice trials. Afterwards, the participants had to absolve five experimental blocks of two trials each. Within one block, the auditory condition remained the same. The succession of the five blocks was individually randomized for each subject.

As some of the auditory conditions represented the pictograms on the board more directly than others, the effort to cognitively link these elements was considerably variable between the conditions. For instance, there is almost no learning effort necessary to link the symbol of a speedometer with the word "speed", whereas the degree of cognitive effort is quite higher when mapping the image of a speedometer to non-speech sounds, especially in the case of arbitrarily assigned tones like earcons. To partly compensate for this, subjects practiced the auditory representation of each pictogram before starting each experimental block. In the case of non-speech sounds, they were provided with a random order presentation of sounds of the respective condition, thereby ensuring that no identical sounds succeeded one another. This was done until they reached 100% correct recognition (correct button presses) in three consecutive presentations of that sound. The training was shortened for the two speech conditions: One correct reproduction was sufficient for fulfilling the task as a pretest with three subjects has shown.

4.5. Apparatus and Stimuli

The program used to present visual and auditory stimuli and collect experimental data ran on a laptop with a 1.2 GHz Intel Pentium M® CPU (chipset Intel 855GM®), 12" TFT screen, and SigmaTel C-Major Audio on-board sound card. The position of the laptop was adjusted to allow for a good view onto the screen. In front of the laptop the board for the choice reaction task was located, which is described in more detail in the "design and tasks" section.

Auditory stimuli were presented using Sennheiser HD 25-1 closed-back professional studio monitoring headphones specially designed to offer high attenuation of background noise in order to minimize distraction by environmental sounds during the experiment.

Since the study took place in an in-vehicle research context, the four events that were represented by stimuli for each group of sounds were: *low battery, speed limit, seat belt, out of fuel*. As described above, the task of pressing a button to confirm the audio message is not meant to correspond to a real-life situation, but serves as a means of reporting the understanding of the meaning of the respective auditory output. All sounds and speech messages have been equally leveled. Non-speech

stimuli were chosen to sound very distinct within the respective group.

- Earcons: The sounds were taken with permission from Stephen Brewster's homepage [15].
  - Battery: ~400 msec effectively + reverberation (adds to ~1sec), violin-like pizzicato motif, three very quick notes.
  - Speed: ~900 msec, piano-like open fifth with tambourine hit in attack phase, single touch.
  - Seat belt: ~780 msec, e-piano/organ-like arpeggio chord, ninth character.
  - Fuel: ~1.4 sec, e-piano minor chord, single touch.
- Auditory icons:
  - Battery: ~370 msec, sound of sci-fi high voltage induced sparks.
  - Speed: ~220 msec, falling pitch like exaggerated departing race car.
  - Seat belt: ~420 msec, metallic clicks, lock-like.
  - Fuel: ~350 msec, quickly pouring liquid

Although the duration of the earcons is longer than those of the auditory icons, their individual character has been established well below 200 ms (for further information on the earcons used, see [15], [16]).

The key principle for the design of the longer speech messages used in this experiment was to keep them comparable to the other conditions in the choice reaction task. Therefore, they made use of the same keyword as the respective keyword condition and this keyword was located at the beginning of each sentence. Thus, the message must not be listened to completely to understand its meaning. This was done to avoid trivially longer response times in the long speech condition. Consequently, a large part of the speech message was in a certain sense irrelevant, as it does not provide the listener with further necessary information.

The keywords and longer speech messages are translated here and specified in the following together with the words and sentences used in the experiment in German. (Note: In German, the keyword was presented at the beginning of each sentence.)

- Keywords:
  - Battery: ~550 msec, "*battery*" ("*Batterie*")
  - Speed: ~920 msec, "*speed*" („*Geschwindigkeit*")
  - Seat belt: ~280 msec, "*seat belt*" ("*Gurt*")
  - Fuel: ~440 msec, "*fuel*" ("*Tanken*")
- Long speech messages:
  - Battery: ~3.7 sec, "Press *battery* to check the state of the battery." ("*Batterie* drücken, um den Ladezustand der Batterie zu überprüfen.")
  - Speed: ~4.2 sec, "Press *speed* to check the speed of the vehicle." ("*Geschwindigkeit* drücken, um die Geschwindigkeit des Fahrzeugs zu überprüfen.")
  - Seat belt: ~3.7 sec, "Press *seat belt* to check the seat belt fastener." ("*Gurt* drücken, um den Verschluss des Gurtes zu überprüfen.")
  - Fuel: ~3.6 sec, "Press *fuel* to check the fill level of the tank." ("*Tanken* drücken, um die Tankfüllung des Fahrzeugs zu überprüfen.")

4.6. Hypotheses

The main hypotheses of this study concern the effect of speech versus non-speech sounds on short-term memory.

- Based on the examination of prior research in the field of irrelevant sound effects mentioned above, we assume that the long verbal messages will have a detrimental effect on short-term memory performance compared to the baseline silence condition.
- As the disruption of the rehearsal process is markedly shorter in the auditory icon and the earcon condition, it can be hypothesized that they will have a noticeable weaker influence on errors made in serial reproduction and will therefore not differ significantly from the baseline condition in this respect, i.e. earcons and auditory icons should not lead to ISE-like effects.
- Of special interest is the keyword condition: If speech itself causes disruptions in short-term memory performance, even short speech messages should cause higher disruptions in short-term memory than non-speech audio. However, given the results of similar effects of speech and non-speech sounds in ISE studies, we assume that if speech messages are as short as keywords, they will not differ in their effect on serial recall from equally short non-speech sounds.

For the choice reaction task, the following hypotheses are brought forward:

- In both the keyword and the long speech condition, one single keyword informs the participants which button to press. As the keyword is located at the beginning of a sentence in the longer speech message, it seems plausible that both speech conditions lead to identical response times. Furthermore, both speech conditions should lead to the lowest button pressing related error rate. This seems plausible because the speech conditions provide the most unambiguous output in this experiment.
- Although every auditory condition was trained before the experimental trials, it seems likely that reacting on abstract auditory events needs more time, because more cognitive processing is needed to understand the meaning of the stimulus. It also seems likely that the same consideration holds for error rate. Thus, we assume that earcons will show the longest response times and possibly lead to more errors, as the tones do not inhere any intrinsic relation to the pictograms they represent.

5. RESULTS

To analyze performance in the short-term memory task, errors in the reproduction of the digit sequence were added up and averaged across all subjects in each experimental condition. Response times and errors in button pressing were also averaged this way.

The results indicate that it takes most time to react to earcons, whereas no differences were found between both speech conditions and auditory icons. Regarding all conditions, subjects did not make many errors in button pressing and no differences between the groups of sounds were found here. Concerning the short-term memory task, long speech messages have a significantly detrimental effect on short-term memory performance, whereas non-speech sounds as well as the presentation of keywords only did not influence recall in comparison to the baseline silence condition.

5.1. Serial recall task

In the serial recall task, the results of all sound conditions were compared against performance under silence during the retention interval. Accumulating both trials of each condition, subjects made on average 7.13 errors in the keyword condition, 7.20 errors in the earcon condition and 7.33 errors in the auditory icon condition in the reproduction of the digit sequence. The results are summarized in Table 1.

Table 1: *Mean sum of serial position errors (and standard error) in the short-term memory task.*

	Auditory Condition				
	Long Speech	Keyword	Earcon	Auditory Icon	Silence
mean	9.67	7.13	7.20	7.33	6.93
SE	0.83	0.87	0.86	1.00	0.74

At first sight, these error rates do not differ strongly from the baseline silence condition, where 6.93 errors were made on average. Paired t-tests calculated for pairwise comparisons confirm this result. As all tests were calculated against the same baseline (a priori planned comparisons) and thus leading to a fewer number of tests than experimental conditions, the significance levels needed not to be adjusted (see [17], chap. 8). No statistically meaningful differences were found between silence and keywords ($t(14)=0.213$; $p=0.834$), silence and earcons ($t(14)=0.654$; $p=0.524$), or silence and auditory icons ($t(14)=0.477$; $p=0.641$). A different image emerges when looking at the long speech condition (see Figure 1): With 9.67 errors on average, an evidently larger effect of the type of feedback sound on memory performance can be stated. The difference of 2.74 errors to the baseline condition is significant ($t(14)=2.541$; $p=0.024$).

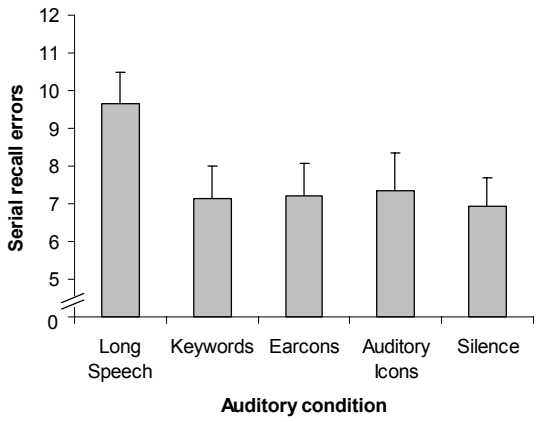

Figure 1: *Mean sum of serial position errors in dependence of auditory condition in the choice reaction task during the retention interval. Error bars represent standard error of the mean.*

5.2. Choice reaction task

In the choice reaction task, the silence condition did not serve as a baseline, because the push-button performance under visual versus auditory feedback is not the topic of this study. Thus, there was no special baseline condition. However, not all possible pairwise comparisons are of interest. Long speech messages and keywords are compared to test the hypothesis that verbal output after the significant part of the message does not lead to longer response times. Furthermore, based on the hypotheses outlined above, both non-speech sound conditions are compared against each other and against the keyword condition. Bonferroni's correction was used to adjust for multiple comparisons. Four comparisons were calculated, so alpha was lowered for each test to 0.0125.

For the analysis of the response time data, only correct button pushes were included. The number of errors in button pushing is analyzed separately. Table 2 summarizes the results.

Table 2: *Mean values (and standard error) of response times (RT) [msec] and relative frequency of errors in button pushing in the choice reaction task.*

		Auditory Condition			
		Long Speech	Keyword	Earcon	Auditory Icon
RT	mean	1223	1211	1689	1135
	SE	82.82	70.15	98.13	65.33
errors	mean	0.03	0.03	0.13	0.05
	SE	0.02	0.02	0.05	0.04

A gaze at Figure 2 reveals that the effect of earcons is outstanding, as they lead to remarkably longer response times. No significant difference was found between response times caused by longer speech messages vs. keywords ($t(14)=0.134$; $p=0.895$). Among the remaining comparisons, only the effect of earcons was found to be reliably different from auditory icons ($t(14)=5.514$; $p=0.000$) and from keywords ($t(14)=-4.847$; $p=0.000$). No difference was detected between auditory icons and keywords ($t(14)=0.967$; $p=0.350$).

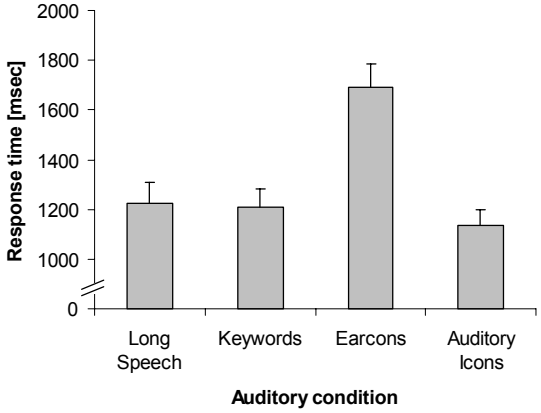

Figure 2: *Response times in the choice reaction task in dependence of auditory condition. Depicted are mean values and standard error of the mean.*

Analyzing the effect of type of auditory output on errors in the choice reaction task, no significant differences were found. Altogether, only few errors were made. The same pairwise comparisons as for the analysis of the response time data have been calculated with one exception: In the long speech condition and in the keyword condition, subjects made exactly the same amount of errors. Thus, no test was calculated for this comparison. Earcons seem to lead to somewhat more errors (13%) compared to keywords (3%) However, there is only a

tendency towards different means: t(14)=-1.871, p=0.082. This tendency is further weakened by the fact that multiple comparisons have been carried out. The differences between earcons and auditory icons (13% vs. 5%) or auditory icons and keywords (5% vs. 3%) also did not reach statistical significance (t(14)=1.234; p=0.238 and t(14)=-0.564; p=0.582).

6. DISCUSSION

In this study we compared the influence of speech and non-speech sounds on short-term memory. Auditory icons and earcons served as non-speech sounds and speech was represented by longer verbal messages and keywords. Besides the memory aspect, we examined the effects of these auditory stimuli on response times and accuracy. The work has been conducted in the context of research on in-car information systems, but there is reason to assume that the results can be transferred to other multiple task scenarios in which informative sounds must not disrupt short-term memory processing.

6.1. Short-term memory aspects

The experiment was designed to answer the questions if earcons or auditory icons can cause detrimental effects on short-term memory and if there is a difference between short and long speech messages in this respect.

As expected from research on the irrelevant speech effect, the long speech condition led to a remarkable decrease in short term memory performance. This finding is especially intriguing, because not the whole speech message was task-relevant. The meaningful part, i.e. the keyword, was located at the beginning of the sentence and subjects were told that only this first word is important and they may ignore the rest. Consequently, even if only a part of a longer speech message is important, users will get distracted by the irrelevant part.

Both non-speech sound conditions did not produce effects on serial recall significantly different from the silence baseline condition. This may be caused by the fact, that the non-speech sounds were very short and therefore not as distracting as the long verbal message. The question remains open whether longer earcons or auditory icons would have caused an effect similar to longer speech. Whereas this is of theoretical interest, it is not topic of applied research, because one of the fundamental sound design principles is to use brief sounds [16]. As no sound designer would reasonably think of creating sounds of approximately equal length like average speech messages, no experimental condition to test the influence of long non-speech sounds was implemented.

Considering the results in the keyword condition, it seems likely that the duration of sounds is a critical parameter: Like the non-speech sounds, the keywords did not negatively influence serial recall as compared to silence. For this reason, the possible hypothesis that speech always leads to disruption in short term memory for serial order can be ruled out.

Taken together this study shows that longer speech messages disturb serial recall, whereas non-speech audio and speech messages in keyword form do not. This raises an interesting question for further investigation: What is the relation between length of utterance and short-term memory performance? Can a general statement be made on the number of words or syllables in a speech message and magnitude of disruption? In the presented experiment, we tested one-word speech messages (leading to no disruption) and complete sentences (leading to highly noticeable decreases in performance). But we did not vary the number of words or syllables systematically. However, it is possible that one-word messages constitute an optimum regarding the disruption of short-term memory.

6.2. Response times and accuracy

Data obtained in the choice reaction task are analyzed in terms of time needed to decode semantic the information of auditorily presented material, therefore response times served to measure how fast subjects understand the meaning of the stimuli. Errors in the choices made upon reacting on auditory events are interpreted as accuracy of the recognition process.

As expected there were no differences found in response times within the group of speech messages because the meaningful part of both speech message types was represented by a keyword, which was located at the beginning of a sentence in the long speech condition. The overall short response times in the speech conditions are most likely caused by the fact that it was possible to summarize the complete message in one single keyword. As soon as a long system message must be listened to from beginning to end in order to understand its meaning (i.e. the content cannot be summarized in one word), response times as short as to those found in this study cannot be expected. In those cases a clear advantage of non-speech sounds can be assumed due to their higher expressive capability [5].

The results support the hypothesis that it takes more time to understand arbitrarily assigned auditory signals: We found a significant difference in terms of response time between earcons and auditory icons and between earcons and speech. Interestingly, there is no difference between keyword speech and auditory icons. This means that the abstraction in mapping natural sounds to system parameters does not hinder rapid understanding of the corresponding event. Some additional aspects must be considered when interpreting these results. First, the slower response times of earcons proved to be significant, but this does not necessarily answer the question whether the magnitude of this difference is meaningful for the design of sounds used to convey non-critical information. In certain cases, the potential of auditory icons of leading to faster stimulus recognition does not compensate for their lack of universality. Factors like the capability of earcons of representing hierarchies [16] or allowing for easy expansions of systems and still keep a consistent sound assignment might be more relevant than faster stimulus recognition in the range of some hundred milliseconds. Second, the time needed to understand the meaning of an earcon is clearly training-dependent. Although we tried to establish familiarity by repeatedly confronting our participants with the sounds in the practice phase before every experimental block, it cannot be ruled out completely that more practice would have further reduced the differences between earcons and keywords or earcons and auditory icons.

An analysis of response errors delivers further information on that issue. No significant differences were found between the experimental conditions. Whereas on a descriptive level it looks like earcons lead to more button pressing related errors, the difference to the other conditions is not statistically reliable. This result can be explained most easily by assuming successful training of the auditory stimuli before every experimental block. Interpreting data this way it seems likely that indeed the participants received an adequate amount of training and that correspondingly the difference in response times between earcons and the other auditory conditions would not have been diminished completely by further practice trials.

A primary goal of this study was to exclusively investigate the effects of auditory output on response time and accuracy. Therefore, only auditory feedback was provided; no visual representation of system events was set up. Based on research on multimodal feedback [18], it seems highly reasonable that by combining visual and auditory output improvements in overall choice reaction task performance can be achieved. However, the suitability of the different types of auditory output to provide non-critical information in the context of driving and in-vehicle devices depends on their capability of supporting auditory-only interaction, because the amount of visual distraction imposed on the driver especially by those devices not primarily relevant for driving should be kept to minimum.

One of the most intriguing findings is that although earcons lead to longer response times, which is a clear sign of higher cognitive processing effort, this higher cognitive load does not interfere with short-term memory demands. In terms of Wicken's [19] multiple resource model, independent cognitive resources seem to be involved in the processes of manually answering to the auditory stimuli and rehearsing serial order information. That explains why earcons can consume more cognitive resources in extracting the meaning of an event (leading to slower response times) and at the same time lead to baseline-like performance in the serial recall task.

7. CONCLUDING REMARKS

This work reports on the investigation of the suitability of speech and non-speech sounds to inform a car driver of relevant but not critical events concerning the state of the vehicle. As reporting non-critical events to the car driver must not interfere with cognitive planning demands of the strategic level of driving [6], short-term memory results from this study indicate that long auditory system messages presented verbally are not suited well for auditory output from a human factors point of view. In a serial recall task, long speech messages in form of complete sentences caused a remarkable decrease in short-term memory performance in comparison to very short speech messages (keywords only) and to non-speech sounds (earcons, auditory icons). Results on speed and accuracy of responses to auditory stimuli show that it is somewhat more difficult to react fast to earcons in comparison to auditory icons and keywords. The latter two proved to be comparable in leading to lower response times and errors in the choice reaction task.

Taken together the results permit to derive some general recommendations for the design of non-distractive auditory output. If using speech, the duration of output should be kept as short as possible. This may seem trivial, but this aspect is definitely neglected in many systems nowadays available. Considering both short-term memory and response aspects, keywords and auditory icons seem most appropriate. However, although earcons led to significantly slower responses, the magnitude of the difference may be negligible for non-time-critical information.

Short keywords or punchy auditory icons are especially recommended to provide information in the vehicle for the auditory channel. The use of earcons should be considered as an alternative to auditory icons depending on key factors of the to-be designed system: If fast and intuitive responses are relevant, auditory icons are advantageous. If the system must allow for consistent and scalable sound design throughout the interface, earcons are the better alternative, because not every interaction and event can be mapped to a natural sound.

8. REFERENCES

[1] M. Blattner, D. Sumikawa, and R. Greenberg, "Earcons and icons: Their structure and common design principles," *Human-Computer Interaction*, vol. 4, pp. 11-44, 1989.

[2] S.A. Brewster, P.C. Wright, and A.D.N Edwards, "A detailed investigation into the effectiveness of earcons," in *Proc. ICAD'92*, 1992, pp. 471-498.

[3] S.A. Brewster, P.C. Wright, and A.D.N. Edwards, "An evaluation of earcons for use in auditory human-computer interfaces," in *Proc. INTERCHI'93*, 1993, pp. 222-227.

[4] W. Gaver, "The SonicFinder: An interface that uses auditory icons," *Human-Computer Interaction*, vol. 4, no. 1, pp. 67-94, 1989.

[5] S.A. Brewster, "Nonspeech auditory output," in *The Human-Computer Interaction Handbook*, J.A. Jacko and A. Sears, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 2002, pp.220-239.

[6] J.A. Michon, "A critical view of driver behavior models: What do we know, what should we do?" in *Human behavior and traffic safety*, L. Evans and R.C. Schwing, Eds. New York: Plenum Press, 1985, pp. 487-525.

[7] W. Ellermeier and J. Hellbrück, "Is level irrelevant in 'irrelevant speech'? Effects of loudness, signal-to-noise ratio, and binaural masking," *Journal of Experimental Psychology*, vol. 24, pp. 1406-1414, 1998.

[8] H.A. Colle and A. Welsh, "Acoustic masking in primary memory," *Journal of Verbal Learning and Verbal Behavior*, vol. 15, pp. 17-31, 1976.

[9] P. Salamé and A.D. Baddeley, "Disruption of short-term memory by unattended speech: Implications for the structure of working memory," *Journal of Verbal Learning and Verbal Behavior*, vol. 21, pp.150-164, 1982.

[10] D.M. Jones, C. Miles, and J. Page, "Disruption of proof-reading by irrelevant speech: Effects of attention, arousal or memory?" *Applied Cognitive Psychology*, vol. 4, pp. 89-108, 1990.

[11] P. Salamé and A.D. Baddeley, "Effects of background music on phonological short-term memory," *Quart. Journal of Exp. Psychology*, vol. 41A, pp. 107-122, 1989.

[12] D.M. Jones and W.J. Macken, "Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory," *JEP: Learning, Memory, and Cognition*, vol. 19, pp. 369-381, 1993.

[13] S.P. Banbury, W.J. Macken, S. Trembley, and D.M. Jones, "Auditory distraction and short-term memory: Phenomena and practical implications," *Human Factors*, vol. 43, no. 1, pp. 12-29, 2001.

[14] S.E. Gathercole and A.D. Baddeley, *Working memory and language*. Hove: Lawrence Erlbaum Associates, 1993.

[15] S. A. Brewster. (2005, April). Principles for improving interaction in telephone-based interfaces. [Online]. Available: http://www.dcs.gla.ac.uk/~stephen/research/telephone/simulator.shtml

[16] G. Leplâtre and S.A. Brewster, "Designing non-speech sounds to support navigation in mobile phone menus," in *Proceedings of BCS HCI 2000*, 2000, pp. 190-199.

[17] G. Keppel, *Design and analysis*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[18] H.S. Vitense, J.A. Jacko, and V.K. Emery, "Multimodal feedback: An assessment of performance and mental workload," *Ergonomics*, vol. 46, pp. 68-87, 2003.

[19] C.D. Wickens and J.G. Hollands, *Engineering psychology and human performance*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2000.