Journal on Multimodal User Interfaces (2019) 13:89–97
https://doi.org/10.1007/s12193-019-00297-9

ORIGINAL PAPER

# Investigating the effects of modality switches on driver distraction and interaction efficiency in the car

Florian Roider[1,2] · Sonja Rümelin[1] · Bastian Pfleging[3] · Tom Gross[2]

Received: 3 May 2018 / Accepted: 8 March 2019 / Published online: 19 March 2019

**Abstract**
In the last decade, the number and variety of secondary tasks in modern vehicles has grown exponentially. To address this variety, drivers can choose between alternative input modalities to complete each task in the most adequate way. However, the process of switching between different modalities might cause increased cognitive effort and finally result in a loss of efficiency. Therefore, the effects of switching between input modalities have to be examined in detail. We present a user study with 18 participants that investigates these effects when switching between touch and speech input on task efficiency and driver distraction in a dual-task setup. Our results show that the sequential combination of adequate modalities for subtasks did not affect task completion times and thus reduced the duration of the entire interaction. We argue to promote modality switches and discuss the implications on application areas beyond the automotive context.

**Keywords** Interaction efficiency · Driver distraction · Modality switch costs · Touch input · Speech input · Multimodal interaction

## 1 Introduction

Touch input has emerged as state-of-the-art input modality, not only on our smartphones but also in many other domains. Lately, also speech input is receiving a lot more attention, especially caused by products and services such as Apple Siri, Amazon Alexa, and Microsoft Cortana. More robust recognition algorithms and intelligent interpretation of the user's voice allow designers and developers to integrate speech recognition into many devices. However, it is unlikely that speech recognition will completely replace touch input in the long term. Both modalities have their unique strengths and weaknesses and complement each other in many ways [11]. The availability of both can compensate drawbacks of a single modality by leveraging the individual advantages of each modality [9,13] and enable users to choose the modality that they assess most appropriate for interaction [8].

Over the last decade, touch and speech input have also found their way into our cars to control in-vehicle information systems (IVIS). Drivers find themselves in a dual-task situation, where they perform additional non-driving-related tasks in addition to the primary task of driving the car. The availability of different input modalities allows drivers to use the mode that is less occupied by the primary task and thereby reduce driver distraction [15]. As drivers have to process mainly visual information while driving, many interaction concepts encourage the usage of speech input. It allows drivers to keep the hands on the steering wheel and the eyes on the road while driving. A potential downside of speech input is its short term and sequential nature, which may put a heavy load on human working memory [1]. This causes cognitive distraction, which can influence drivers' visual behavior and result in inattentional blindness [4,18]. Consequently, the best way to keep driver distraction low is to reduce every form of non-driving-related interaction with the vehicle as much as possible. In this regards, increasing interaction efficiency can reduce the amount of time the driver is distracted from the driving task.

However, not all interactions in the car can be efficiently completed by using only speech, due to the fact that one task might be suited for speech input, but a consecutive task is better served using touch (or the other way around). In order

✉ Florian Roider
florian.roider@bmw.de

[1] BMW Group Research, New Technologies Innovation, Parkring 19, 85478 Garching bei München, Germany

[2] Universität Bamberg, Bamberg, Germany

[3] Ludwig-Maximilians-Universität München, Munich, Germany

to complete the entire interaction in the most efficient way, drivers must be allowed to flexibly switch between the modalities. In this regard, many automotive manufacturers allow a more flexible combination of touch and speech input. For example, current IVISs, such as the Daimler MBUX or the BMW iDrive 7, enable drivers to use speech input to directly jump to specific functions, e.g. showing nearby restaurants. Drivers can then continue with speech input, or switch to touch input for a more efficient selection of a restaurant on the map. In this context, it is essential to assess the costs (e.g. time and distraction) arising from the process of switching between input modalities. High switch costs could lead to efficiency losses or increased distraction and thus reduce the benefits of using multiple modalities.

In this paper, we present a user study that investigates the influence of modality switches between touch and speech input. We aim to assess the costs and benefits for switching between touch and speech input, by observing how users perform two tasks in various sequences.

## 2 Related work

A typical way of combining different input modes in multimodal interaction concepts are *temporally cascaded modalities* [10]. In this form, different modalities are sequenced in a particular temporal order. A number of authors have presented concepts that combine modalities such as speech, touch, or gaze in sequential order (e.g. [6,8,13]). The advantage of this approach is that the use of different modalities for different steps in the interaction allows to always use the best suited (e.g. most efficient or most convenient) input modality.

The appropriateness of touch or speech input depends on the type of the task that is to be completed. *Spatial* tasks require interaction with the three axis of translation or orientation (e.g., the movement or rotation of objects). They are best served by manual interaction, such as direct touch input [5,19]. *Verbal* tasks use language or some arbitrary coding to express verbal information (e.g., the name of an object) are best served using speech [10,19]. Since a great number of tasks do not fit clearly into either category, the verbal and spatial labels are best thought of as endpoints of a continuum [19]. However, while certain input modalities might be well suited for certain tasks when they are applied individually that does not necessarily mean that a combination of two modes in close temporal sequence works equally well.

This is why a number of studies have observed costs for switching between modalities: Gondan and colleagues have examined effects for switching between visual and auditory stimuli [3]. They found that reaction times were slower after the modality of the stimulus had changed. Similar costs have been observed when switching between input modalities. Zhang et al. [21] presented an experiment in which they investigated the application of gaze and gesture input for two sequential interaction steps (hover and select). They compared gaze-hover and hand-select with hand-hover and hand-select and observed that the transition from gaze-hover to hand-select took longer than from hand-hover to hand-select. Furthermore, Monsell has conducted several experiments to examine switch costs that emerge from changing from one cognitive task to another [7]: participants conducted pairs of tasks of different types in alternating order. Responses on tasks that occurred immediately after a switch took longer and were usually more error-prone. He also notes that knowledge of the upcoming task and time to prepare for it usually reduced the impact of average switch costs. Moreover, there can be physical switching costs that are for instance modeled in the original Keystroke-Level Model, when moving ("homing") the hands between input devices like keyboard and mouse [2]. Similar costs are reported for moving the hands between different parts of car interfaces [17].

Experiments like these showed that both, the switches between different input and output modalities and switches between different tasks are connected with a loss of efficiency. Therefore, the efficiency of individual input modalities in the context of a multimodal application has to be assessed in regard of the appropriateness of used modalities, as well as the costs that emerge from switching between them. While touch and speech are ubiquitous today, there is so far only little understanding for the costs of switching between modalities. We address this gap with our study where we examine the effects of switching between touch and speech input and, thus, contribute to understanding how to efficiently combine these modalities.

## 3 Experiment

We conducted a user study with 18 participants, to investigate the influence of modality switches between touch and speech input in a *dual-task situation*. The participants completed series of alternating tasks that define an entire interaction sequence (see Fig. 1). The focus of this experiment was the assessment of costs (especially time) that emerge from the process of switching between efficient input modalities rather than a comparison of touch and speech. We propose three hypotheses to address our research question in a differentiated way:

H1 Switching to a more efficient input modality leads to increased task completion times for *individual tasks* (i.e., switch costs).

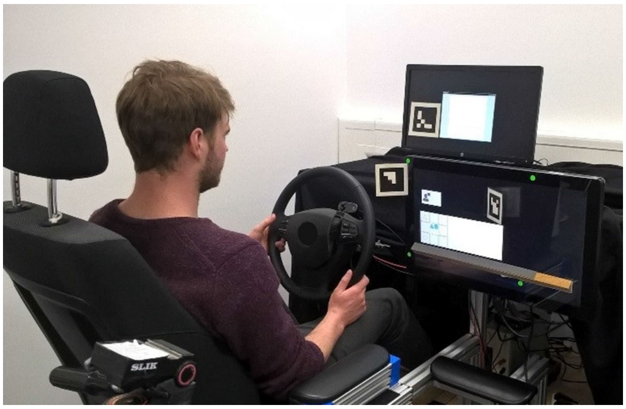

**Fig. 1** The experimental setup for our user study included a primary task in front of the user and a secondary task on the display to the right of the user

The cognitive process of switching the input modality is associated with certain costs. Therefore, task completion times will be longer for tasks that are preceded by a modality switch.

H2 Switching to a more efficient input modality for changing tasks increases the efficiency of the *complete interaction sequence*.
Despite potential costs due to modality switches, the use of more efficient input modalities for tasks will result in a shorter duration for the entire interaction sequence.

H3 Switching to a more efficient input modality for changing tasks improves primary task performance.
The use of more appropriate input modalities reduces cognitive workload, which will have a positive effect on the performance of the primary task.

## 3.1 Participants

A total number of 18 participants (13 males, 5 females) with a mean age of 35.2 years ($SD = 7.7$) participated in the study. All but one participant used touch-enabled screens on a daily basis. Speech interaction was generally less frequently used by participants, but only one participant stated that she had never used speech recognition before.

## 3.2 Study design

We used a within-subject design in this experiment. Each participant conducted ten trials of 90 seconds each. The trials differed in the tasks used, the modalities used, and the sequence of tasks. Table 1 illustrates all trials. They are divided in four blocks, which were permuted between participants to prevent ordering effects.

The *baseline* block contained four trials, where neither the participants' modalities nor their task type changed during the trial run and, thus, repeatedly performed one type of tasks (the first four lines in Table 1). These trails were used to determine the efficiency of touch and speech input for the two tasks, without the influence of any switches. In order to examine the influence of modalitiy switches, there were two *modality switch* runs where the task type stayed the same but the modality changed for each repetition and two *task switch* runs where the modality was fixed for the whole trial, but the task type changed for each repetition. Moreover, there were two *combined switch* runs where both, the modality and the task changed for every repetition: the first of these two runs combined each task with its most suited modality (Move with touch and Describe with speech), while the second one combined each task with the less suitable modality (e.g. speech with the move task). The latter was only included as a matter of completeness but did not reveal interesting insights and will, thus, not be further discussed.

## 3.3 Experimental tasks

Driving a car is an example of a typical dual-task situation. While driving, maneuvering the car is the primary task; other activities such as interacting with in-vehicle systems are secondary tasks [20]. To be able to generalize our findings, we used the Critical Tracking Task [12] as *primary task*, which abstracts the driving task to a continuous visual-manual task. In addition, we used two specific *secondary tasks* to be performed parallel to the primary task. By using abstract tasks, we aim to understand the effects of switching modalities in a more general context.

### 3.3.1 Primary Task: Critical Tracking Task

The Critical Tracking Task (CTT) has been shown to be sensitive to changes in the level of demand of the secondary task by imposing a constant visual-manual load on the user [12]. Figure 2 shows a screenshot of the CTT screen. The user's task is to keep the black vertical line in the middle of the screen, while it is constantly moving away from the dashed line in the center. It gains speed the further it is away from the center. Participants used to two buttons, both on the left side of the steering wheel to move the vertical line to the left or to the right. This way participants could always control the CTT even while making touch inputs with their right hand. Furthermore, the CTT has two advantages over alternative primary tasks that make it very suitable for our study design. First, in comparison to the widespread Lane Change Task, it produces an uniform demand, which is beneficial when splitting up the interaction sequence in short subtasks. This way we ensure that each task is performed with the same demand of the primary task. Second, the CTT is a very simple task, which does not require long training phases for the participants, in contrast to more realistic driving simulators.

**Table 1** This table summarizes the conducted trials in the experiment

| # | Block | Tasks | Modalities | Sequence |
|---|---|---|---|---|
| 1 | Baseline | Move (M) | Touch (T) | **MT–MT–MT**… |
| 2 | Baseline | Move | Speech (S) | MS–MS–MS… |
| 3 | Baseline | Describe (D) | Touch | DT–DT–DT… |
| 4 | Baseline | Describe | Speech | **DS–DS–DS**… |
| 5 | Modality switch | Move | Touch ↔ Speech | MS–**MT**–MS… |
| 6 | Modality switch | Describe | Touch ↔ Speech | DT–**DS**–DT… |
| 7 | Task switch | Move ↔ Describe | Touch | **MT**–DT–**MT**… |
| 8 | Task switch | Move ↔ Describe | Speech | MS–**DS**–MS… |
| 9 | Combined switch | Move ↔ Describe | Touch ↔ Speech | **MT–DS–MT**… |
| 10 | Combined Switch | Move↔Describe | Speech ↔ Touch | MS–DT–MS… |

The trials differ regarding the conducted tasks, Move (M) and Describe (D), as well as the used input modalities, Touch (T) and Speech (S). The execution of a task with a specific input modality is described with both letters, e.g. MT refers to the Move task with touch input

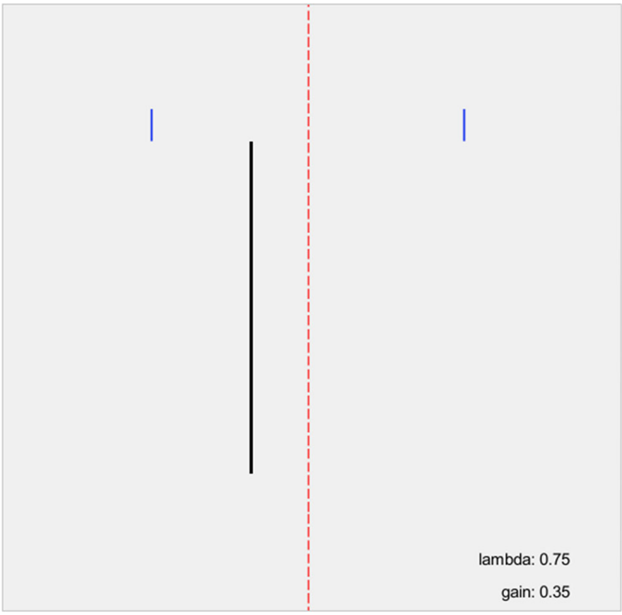

**Fig. 2** Visualization of the Critical Tracking Task: the user's task is to keep the black vertical line in the middle of the screen (dotted line)

This way, we were able to keep explanations and training phases short and avoid learning effects over the duration of the experiment.

### 3.3.2 Spatial Secondary Task: "Move"

The *Move* task is illustrated in Fig. 3. It is a spatial task and thereby potentially well-suited for touch input [5,19]. The goal is to move the two colored shapes from the center area to corresponding placeholders on the outside. A real world equivalent for this abstract task could be the panning of the navigation map to find the desired location. When using *touch input*, participants moved the elements by dragging them with their fingers to the target position on the touch screen. For *speech input*, people selected the elements by naming their shape and then described the location of the target fields to move them (e.g. "Circle to top-right"). Participants were free to move either the left or the right element first.

### 3.3.3 Verbal Secondary Task: "Describe"

*Describe* is a verbal task which exploits the strong descriptive capabilities of speech [10]. Participants have to describe the element that is displayed on top of the screen by characterizing it by three attributes: shape, color, and size. A real world equivalent is the input of an address for a navigation system or any other task where information has to be verbalized. Just like the colored elements, an address can be split into several attributes: the city, the street, and a house number. When using *speech input* for this task, participants could simply name the attributes of the element (e.g., "small yellow circle"). After a description was completed, the next task appeared. When using touch input for this task, the screen displayed a selection of possible attributes as illustrated in Fig. 4. Each column represented the possible selections for one attribute. The participants had to touch the correct buttons in each column. While speech input was used, the interface did not display buttons with possible options, since we did not want the interface to influence what participants say.

## 3.4 Apparatus

The study was conducted in a laboratory setting consisting of a car seat, a steering wheel with buttons, and two displays as shown in Fig. 1. A display at the position of the windshield showed the CTT. A touch-sensitive display to the right of the steering wheel displayed the secondary tasks. The size and position of the interaction area on this display were chosen

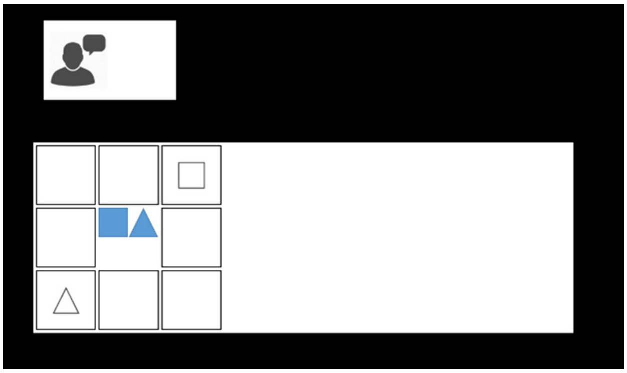

**Fig. 3** Move task: both shapes from the center area have to be moved to corresponding surrounding fields (triangle to bottom-left and square to top-right). The icon on the upper left instructs which modality to use (speech in this case)

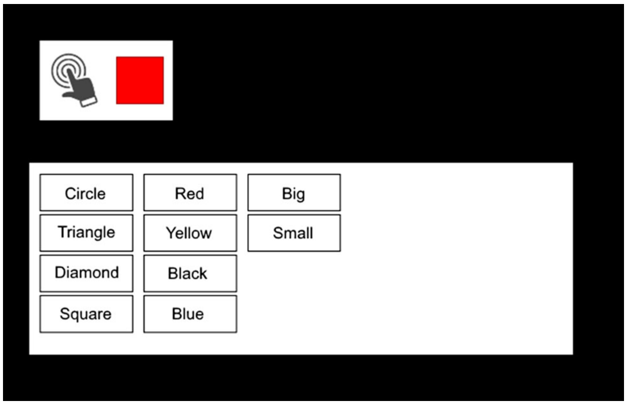

**Fig. 4** Describe task: participants have to characterize the element on top of the screen by shape, color, and size. The touch buttons were only visible when using touch input but were hidden for speech input

according to typical arrangement of touchscreens in premium class cars. We implemented the tasks in a test framework based on HTML5 and JavaScript with full functionality for touch input. Since speech commands for both tasks were relatively long and complicated, but easy to remote control, we decided to use a Wizard-of-Oz approach with a keyboard-based interface that allowed to quickly execute users voice commands. Based on literature [12], we implemented our own version of the CTT in Unity3D, which allowed us to trigger data logging for runs automatically via the framework. We recorded average CTT deviations over the whole duration of each trial as an indicator for primary task performance. The framework recorded the average task completion times (TCT) for a task depending on the input modality and the sequence of tasks in which it occurred.

### 3.5 Procedure

After a short introduction, the participants filled out a consent form and a questionnaire capturing demographic data and adjusted the seat position. The examiner presented the CTT as the primary task and explained its functionality. Participants conducted a trial with the CTT (without a secondary task) to get used to the controls. Then the examiner introduced the secondary tasks and explained the interaction with both modalities. Participants had a few minutes to practice both tasks with both input modalities (without CTT). During the trials, participants were instructed to have their primary focus on the CTT. At the same time, they should try to complete as many secondary tasks as possible in the given time.

## 4 Results

We analyzed data the using SPSS. Kolmogorov–Smirnov tests were used to control that data was distributed normally. Significance levels were corrected and reported according to Greenhouse–Geisser, if Mauchly's test indicated that the assumption of sphericity was violated.

Our hypotheses focus on effects that emerge when switching between the most efficient input modalities for alternating tasks. Therefore, in a first step, we confirm that touch is most efficient modality for *Move* and speech is the most efficient modality for *Describe*. The baseline trials are determined by two independent factors: the task (Move, Describe) and the modality (Touch, Speech). They refer to trials 1–4 in Table 1. A two-way repeated measures ANOVA indicates a strong interaction between *Task* and *Modality* ($F = 295.24$, $p < .001$, $\eta_p^2 = .95$). The results are illustrated in Fig. 5. Follow-up $t$-tests show that touch ($M = 3.99$, $SD = 1.12$) was faster than speech ($M = 6.82$, $SD = 0.72$) for the spatial *Move* task ($t = -10.27$, $p < .001$). The other way around, speech ($M = 4.20$, $SD = 0.41$) was faster than touch ($M = 6.50$, $SD = 1.50$) for the verbal *Describe* task ($t = 7.84$, $p < .001$).

### 4.1 Switch costs for individual tasks

The results from the baseline trials show that touch is more efficient for the move task and speech for the describe task.

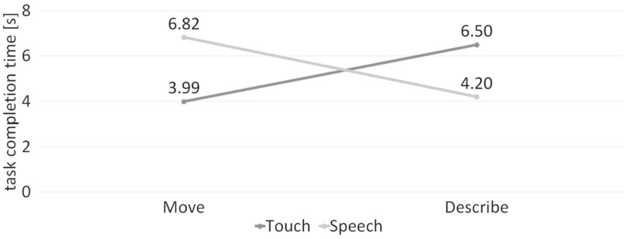

**Fig. 5** The task completion times from baseline runs illustrate the suitability of touch for Move and Speech for Describe

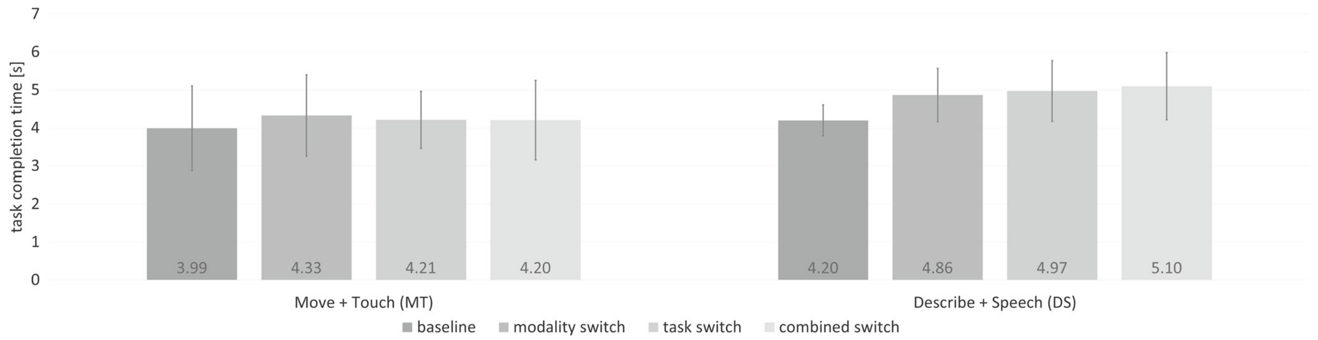

**Fig. 6** Task completion times for Move + Touch and Describe + Speech. Modality switches led to an increase of TCT when the task stayed the same. In contrast, modality switches in the course of a task switch did not induce additional costs compared to only task switches. Error bars indicate the standard deviation

We use the term *tuple* to describe the combination of a task with an input modality. The next step is to determine switch costs by investigating how these efficient tuples perform within different sequences.

The two independent variables for the following analysis are: the tuple (*Move* + *Touch*, *Describe* + *Speech*) and the sequence in which the tuple occurs (baseline, modality switch, task switch, combined switch of task and modality). The according trials in Table 1 are 1, 5, 7, 9 for *Move* + *Touch*, and 4, 6, 8, 9 for *Describe* + *Speech*. Figure 6 shows the task completion times for both tuples in the four different sequences. A two-way repeated measures ANOVA indicates that the task completion time was mainly affected by the tuple ($F(1, 17) = 12.67$, $p < .01$, $\eta_p^2 = .427$), but also by the sequence in which the tuples were performed ($F(3, 51) = 5.73$, $p < .01$, $\eta_p^2 = .252$). Pairwise comparisons (bonferroni-corrected) revealed a significant rise of mean task completion times (+502 ms) during modality switch runs compared to baseline runs ($p < .001$). During task switch trials, the mean completion times also took longer (+500 ms) than in the baseline condition ($p = .001$). Finally, combined task and modality switches also resulted in significantly increased completion times (+556 ms) compared to the baseline ($p = .010$). The processes of switching between modalities, switching between tasks, and switching task and modality at the same time all led to an increase of task completion times compared to baseline runs. Surprisingly, the costs for modality switches and task switches did not add up when both occurred at the same time. TCT in combined task and modality switch runs were not longer compared to TCT in in task switch runs ($p > .05$).

## 4.2 Interaction sequence efficiency

In the next step, we focus on the efficiency for completing an interaction sequence that consists of different tasks. Therefore, we compare the summed task completion times for completing both tasks depending on the used modalities: only speech input, only touch input, or task-depending input of both (Move + Touch and Describe + Speech). This refers to trials 7, 8, and 9 in Table 1. Figure 7 illustrates the summed task completion times for both tasks. The average time to complete both tasks was greatest using only speech input ($M = 11.78$, $SD = 1.80$). Using only touch input was faster ($M = 10.36$, $SD = 1.67$). Finally, the combined usage of touch and speech was the most efficient input form to complete both tasks ($M = 9.30$, $SD = 1.76$). A one-way repeated measures ANOVA showed a significant main effect ($F(2, 34) = 19.99$, $p < .01$, $\eta_p^2 = .540$). All pairwise comparisons (bonferroni-corrected) were significant. The combined use of both input modalities reduced the interaction time by 21% (2.48 s) compared to only speech input ($p < .001$) and by 10% (1.03 s) compared to only touch input ($p < .01$).

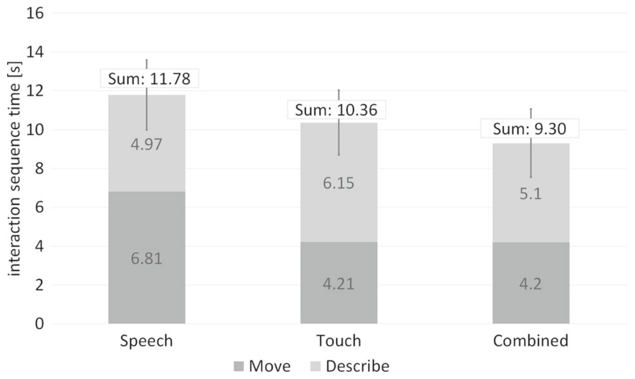

**Fig. 7** The usage of both input modalities for alternating tasks was faster than only speech or only touch. Error bars indicated the standard deviation

## 4.3 Primary task performance

CTT values range from −100 (the vertical bar is at the left border) to 100 (the vertical bar is at the right border). The CTT performance is measured as the mean absolute

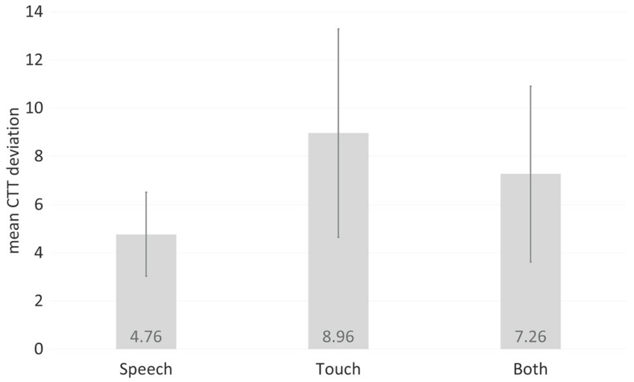

**Fig. 8** Mean CTT deviations for input with only touch, only speech, or touch and speech input when alternating between tasks. Error bars indicated the standard deviation

position of the vertical bar during one trial. Again, we examine CTT performance depending on *used modalities* (only speech input, only touch input, or task-depending input of both). Figure 8 shows the average CTT values for those conditions with only touch ($M = 8.96, SD = 4.20$), only speech ($M = 4.76, SD = 1.69$) and the alternating use of touch and speech ($M = 7.26, SD = 3.54$). A one-way repeated measures ANOVA shows that the *used modalities* had a significant effect on average CTT deviation, $F(2, 34) = 16.64$, $p < .01$, $\eta_p^2 = .495$. Pairwise tests (bonferroni-corrected) show that the combination of touch and speech was less distracting than only touch ($p < .05$), but also more distracting than only speech ($p < .01$).

## 5 Discussion

The basic assumption for the hypotheses of this study is that some tasks are better suited for either speech or touch input [19]. The results from the baseline trials justify this basic assumption: Touch was the more efficient input modality for Move, whereas speech input was more efficient for Describe.

*H1* addresses the costs of modality switches. We claimed that switching to a more efficient modality leads to increased completion times for individual tasks, due to the process of switching. Our results showed that modality switches caused an increase of task completion time compared to baseline runs. However, in an interaction sequence that requires the user to change between different tasks—which means that there are already costs for task switches—additional modality switches could be performed without a loss of efficiency. The process of switching between efficient modalities did not induce any additional time for task completion and we reject *H1*.

For *H2*, we assumed that switching to a more efficient input modality increases the total efficiency of an interaction sequence with different tasks. A simple example for such an interaction sequence is illustrated in Fig. 7. The results showed that touch is more efficient for *Move* and speech for *Describe*. Moreover, in the previous section we concluded that switches between both modalities can be made without loss of efficiency. Consequently, switching between modalities compared to using only speech or only touch increases the overall efficiency and we accept *H2*. While this result was to be expected, the important insight is that the benefit regarding efficiency of the entire sequence compensates for any loss of efficiency due to the process of switching between subtasks.

In *H3* we claimed that switching to a more efficient input modality could reduce cognitive workload and therefore increase the performance of the primary task. Switching between touch and speech was more distracting than only speech input, but also less distracting than only touch input. This can be explained since participants spent about half of the duration of the trial with either modality. This implies that the process of switching the input modality did not result in additional distraction. The CTT deviation was mainly influenced by the used modalities and not by the modality switch itself. While CTT results did not directly show that the efficient use of touch and speech reduces distraction from the primary task there might still be an indirect effect. Distraction might not be smaller when switching between modalities, but the time the driver is distracted from the primary task will be definitely shorter. Still, the best primary task performance was achieved using only speech and we reject *H3*.

The small effect of the modality switches in our setup might be also influenced by the fact, that average switch costs are usually reduced when knowledge of the upcoming task exists [7,8]. In our study design, participants had knowledge about upcoming tasks and modalities. We assume that this knowledge had a similarly positive effect on TCT, according to knowledge of the upcoming task. Conversely, not knowing the modality for an upcoming interaction step might result in increased TCT, as the user has to determine the requested input modality first. A user interface could require the user to switch to a more efficient modality, without increasing TCT, as long as the user has prior knowledge of this switch.

### 5.1 Limitations

With respect to our results it is important to respect some limiting factors that our particular experimental apparatus entailed. First, the CTT is no replacement for a real driving task. Pressing buttons on the steering wheel does not equal the use of a steering wheel to control a real car. Therefore, CTT values cannot be directly translated to absolute driving performance. However, the CTT can be used to indicate relative patterns regarding the distraction of secondary tasks [12]. Second, *Move* and *Describe* are abstract tasks, which are used as reasonable stand-ins for subtasks within a particular interaction sequence. It remains to be determined in future work how far the results apply to different types of secondary tasks or to other input modalities, such as gestures or steering wheel controls. Third, the alternation between two task represents only an approximation to real-world interaction steps. In-car interactions are often composed of more than just two tasks, but still shorter than 90 seconds. We chose this duration for runs in order to produce meaningful CTT values [12]. Alternating between two tasks for this duration allowed us to focus on the effects of the switching process based on a larger number of samples. Finally, as with all experiments using a Wizard-of-Oz approach, the performance of the assistant must be critically reflected with regard to the influence on the results. We minimized the wizard's influence with the help of extensive practice by the wizard during pre-studies, combined with a quick and simple hotkey interface for speech commands.

### 5.2 Implications for multimodal applications

The design of the abstract tasks was based on typical in-vehicle tasks. Yet, by abstraction of concrete contents and by largely forgoing design elements, such as color themes, special shapes, or animations, we kept design-related effects as small as possible. For this reason, we claim that our findings are applicable for a wider range of applications beyond the automotive context, which aim to apply different modalities in dual task situations, such as using the smartphone while walking, or operating a computer while surveilling an assembly line.

Developers in these domains should consider optimizing interaction steps for the most efficient modalities, in order to further exploit the strengths of each modality. This might also require to promote modality switches to the user. For example, once a user starts interacting with a certain input modality, interactive systems often adapt to this modality and foster the consecutive use of the same modality. This can be observed in automotive infotainment systems, but also on mobile devices such as smartphones. When touching on a button to enter an address in the navigation system or search for a name in my contacts the usual behavior of the system is that a keyboard pops up in order to facilitate touching for the user. We argue that such systems should strongly promote a switch to speech input instead, e.g. by enabling speech recognition automatically and accordingly notify the user whenever any textual input is required. In this regard, it has been shown that visual cues can significantly increase the use of speech input [16]. Our results showed that speech input is the superior input mode for textual input and that the switch from the touch based selection to speech based text input does not induce a loss of efficiency. Although speech has shown to be most efficient for these tasks, many users simply still do not like speech-based systems [14]. Especially for these users, a system that promotes speech input not for all tasks, but only when it provides a significant benefit, could reduce distraction from the primary task. On the other hand, use cases which are well suited for touch input, such as map panning or selection of few items should be further developed for touch input.

## 6 Conclusion

In this paper, we presented a user study that examines costs and benefits when switching between touch and speech input in a dual task situation. Switching input modalities between different tasks increased efficiency compared to both unimodal conditions. The process of switching itself did not increase task completion times when it occurred in the course of a task switch. The performance in the primary task indicates that distraction in the multimodal condition is mainly determined by the used modalities and not by the process of switching between them. If efficiency and suitability of interaction is prioritized over reducing distraction (e.g. with higher levels of automation that allow drivers to shift their attention away from the driving task), developers of multimodal systems should consider promoting modality switches for individual task in order to exploit the benefits of individual modalities. A challenge for future work will be to investigate how drivers can be effectively guided to use suited modalities, without restricting them in their interaction possibilities.

## References

1. Bradford JH, James H (1995) The human factors of speech-based interfaces. ACM SIGCHI Bull 27(2):61–67. https://doi.org/10.1145/202511.202527
2. Card SK, Moran TP, Newell A (1980) The keystroke-level model for user performance time with interactive systems. Commun ACM 23(7):396–410. https://doi.org/10.1145/358886.358895
3. Gondan M, Lange K, Rösler F, Röder B (2004) The redundant target effect is affected by modality switch costs. Psychon Bull Rev 11(2):307–313
4. Harbluk JL, Noy YI, Trbovich PL, Eizenman M (2007) An on-road assessment of cognitive distraction: impacts on drivers visual behavior and braking performance. Accid Anal Prev 39(2):372–379. https://doi.org/10.1016/j.aap.2006.08.013
5. Koons DB, Sparrell CJ, Thorisson KR (1993) Integrating simultaneous input from speech, gaze, and hand gestures. In: Maybury MT (ed) Intelligent multimedia interfaces. AAAI Press, Menlo Park, pp 257–276
6. Mitrevska M, Moniri MM, Nesselrath R, Schwartz T, Feld M, Korber Y, Deru M, Muller C (2015) SiAM—situation-adaptive multimodal interaction for innovative mobility concepts of the future. In: Proceedings of the international conference on intelligent environments—IE '15. IEEE, pp 180–183. https://doi.org/10.1109/IE.2015.39
7. Monsell S (2003) Task switching. Trends Cogn Sci 7(3):134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

8. Müller C, Weinberg G, Vetro A (2011) Multimodal input in the car, today and tomorrow. IEEE Multimed 18(1):98–103. https://doi.org/10.1109/MMUL.2011.14
9. Ohn-Bar E, Trivedi MM (2014) Hand gesture recognition in real time for automotive interfaces: a multimodal vision-based approach and evaluations. IEEE Trans Intell Transp Syst 15(6):2368–2377. https://doi.org/10.1109/TITS.2014.2337331
10. Oviatt S (2012) Multimodal interfaces. In: Jacko JA (ed) The human–computer interaction handbook: fundamentals, evolving technologies and emerging applications, 14th edn. CRC Press, Boca Raton, pp 286–304
11. Oviatt S, Coulston R, Lunsford R (2004) When do we interact multimodally? Cognitive load and multimodal communication patterns. In: Proceedings of the 2004 international conference on multimodal interfaces—ICMI '04, pp 129–136. http://doi.acm.org/10.1145/1027933.1027957
12. Petzoldt T, Bellem H, Krems JF (2014) The critical tracking task: a potentially useful method to assess driver distraction? Hum Fact J Hum Fact Ergonom Soc 56(4):789–808. https://doi.org/10.1177/0018720813501864
13. Pfleging B, Schneegass S, Schmidt A (2012) Multimodal interaction in the car—combining speech and gestures on the steering wheel. In: Proceedings of the 4th international conference on automotive user interfaces and interactive vehicular applications—AutomotiveUI '12. ACM, New York, pp 0–7. https://doi.org/10.1145/2390256.2390282
14. Pickering CA, Burnham KJ, Richardson MJ (2007) A research study of hand gesture recognition technologies and applications for human vehicle interaction. In: Institution of engineering and technology conference on automotive electronics, pp 1–15
15. Pickering CA, Burnham KKJ, Richardson MJM (2007) A review of automotive human machine interface technologies and techniques to reduce driver distraction. In: 2nd IET international conference on system safety. IEEE, pp 223–228. https://doi.org/10.1049/cp:20070468
16. Roider F, Rümelin S, Gross T (2018) Using visual cues to leverage the use of speech input in the vehicle. In: International conference on persuasive technology. Springer, Cham, pp 120–131. https://doi.org/10.1007/978-3-319-78978-1_10
17. Schneegaß S, Pfleging B, Kern D, Schmidt A (2011) Support for modeling interaction with automotive user interfaces. In: Proceedings of the 3rd international conference on automotive user interfaces and interactive vehicular applications—AutomotiveUI '11, p 71. https://doi.org/10.1145/2381416.2381428
18. Strayer DL, Watson JM, Drews FA (2011) Cognitive distraction while multitasking in the automobile. In: Federmeier KD (ed) Psychology of learning and motivation, vol 54. Elsevier, Amsterdam. https://doi.org/10.1016/B978-0-12-385527-5.00002-4
19. Wickens CD, Sandry DL, Vidulich M (1983) Compatibility and resource competition between modalities of input, central processing, and output. Hum Fact 25(2):227–248. https://doi.org/10.1177/001872088302500209
20. Wierwille WW (1993) Demands on driver resources associated with introducing advanced technology into the vehicle. Transp Res Part C Emerg Technol 1(2):133–142. https://doi.org/10.1016/0968-090X(93)90010-D
21. Zhang Y, Stellmach S, Sellen A (2015) The costs and benefits of combining gaze and hand gestures for remote. In: Abascal J, Barbosa S, Fetter M, Gross T, Palanque P, Winckler M (eds) Human–computer interaction—INTERACT 2015. INTERACT 2015. Lecture notes in computer science. Springer, Cham. https://doi.org/10.1007/978-3-319-22698-9

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.