Using Visual Cues to Leverage the Use of Speech Input in the Vehicle

Florian Roider[1,2], Sonja Rümelin[1], Tom Gross[2]

[1]BMW Group Research, New Technologies, Innovations, Munich, Germany
[2]Human-Computer Interaction Group, University of Bamberg, Bamberg, Germany
florian.roider@bmw.de

Abstract. Touch and speech input often exist side-by-side in multimodal systems. Speech input has a number of advantages over touch, which are especially relevant in safety critical environments such as driving. However, information on large screens tempts drivers to use touch input for interaction. They lack an effective trigger, which reminds them that speech input might be the better choice. This work investigates the efficacy of visual cues to leverage the use of speech input while driving. We conducted a driving simulator experiment with 45 participants that examined the influence of visual cues, task type, driving scenario, and audio signals on the driver's choice of modality, glance behavior and subjective ratings. The results indicate that visual cues can effectively promote speech input, without increasing visual distraction, or restricting the driver's freedom to choose. We propose that our results can be applied to other applications such as smartphones or smart home applications.

Keywords: visual cues, prompts, speech input, triggers, persuasive

1 Introduction

Touch input has become the state of the art input modality for interaction with many devices over the last decade. More recently, speech input is about to emerge as a full-fledged alternative to touch input, supported by the success of voice based systems such as Amazon Alexa, Apple's Siri or the Google assistant. Besides mobile devices or smart home applications, touch and speech have evolved as the dominating input modalities in the automotive domain. The latest models of many manufacturers integrate large touch based screens and intelligent speech based systems, but there is no or only little interplay between both modalities at the moment. Touch is usually the primary input mode, while speech input is mostly a less used alternative path for specific use cases that work independently of the touch interaction.

There is a number of advantages of speech compared to touch input that support the driver's safety. Speech input reduces visual distraction, it allows drivers to keep both hands on the steering wheel, and it offers a fast and convenient way to achieve many tasks in the vehicle, especially those that require the driver to enter text in any forms (e.g. when giving destinations, searching for contacts, or composing text messages). However, for some tasks, especially those that require the user to express spatial information, touch input is suited much better [14]. Furthermore, it has been shown that speech input is not free of distraction either and that situational influences can impair the suitablity of speech input [10, 12]. Finally, speech input still faces some technical challenges such as understanding heavy dialect or recognition in noisy conditions.

In order to cope with such problems it makes sense to integrate both input modalities in the car. The challenge is to find a seamless and efficient interplay between alternative input modalities, so that users can actually benefit from the many possibilities they have. The "user should be made aware of alternative interaction options without being overloaded by instructions that distract from the task" [9]. In this work, we address the question if visual cues provide an effective, but unobtrusive way to leverage speech input while driving.

2 Related Work

Fogg describes the likelihood of influencing peoples' behavior as product of three factors [4]. Besides sufficient motivation and the ability to perform a target behavior, effective triggers are necessary. There are three types of triggers: *sparks* motivate behavior, *facilitators* make behavior easier, and *signals* simply remind people to perform a behavior [4]. In the case of speech input while driving, reduced distraction and increased safety provide a strong motivation. Furthermore, we assume that people have the ability and know-how to use speech input. In this case, visual cues are signals that just remind people to use speech input now. But they can also serve as facilitators, that make the target behavior easier to do. By displaying possible voice commands they help reducing the effort for formulating words ourselves, reduce the thinking effort and thus increases the likeliness that speech input is used.

2.1 Effects of the Prompt Modality

Why do people rather interact via touch instead of speaking to current cars in regard of these benefits? A psychological explanation is the cognitive mapping of visual stimuli to manual responses [13, 14]. Large touch-sensitive screens in current vehicles provide visual stimuli that provoke direct touch input. The other way around, auditory stimuli are most compatibly mapped to speech responses[13, 14]. Accordingly, one way to remind users to use speech input is to prompt them with auditory cues, such as spoken prompts or earcons. Yet, visual cues have some major advantages over spoken or auditory cues: Visual cues are faster. Users can benefit from preattentive processes that support rapid pattern recognition and thereby absorb information at one glance [8]. Furthermore, auditory prompts are short term and sequential by nature and thus make heavy demands on human working memory [1]. Visual cues, in contrast, do not have this temporal relation and can be displayed permanently. At the same time, they are less disruptive than acoustic prompts. Playing a sound or spoken prompt whenever the user should use speech interaction can be very annoying. Parush compared spoken and visual prompts for speech dialog interaction in multitasking situations such as driving [8]. They found that speech interaction with spoken prompts took longer than with visual prompts, whereas the driving performance was better with spoken prompts. Their study also showed that the difficulty of the tracking task affected these results. They conclude that multitask situations must not always have spoken prompts. Especially novice users can profit from visual cues for speech interaction [15]. In multimodal systems, this allows to display the names of possible selections to suggest or explicitly indicate what users can say.

2.2 Implicit vs. Explicit Prompts

Explicit prompts stand in contrast to implicit prompts that help to direct user input in a more reserved way. Yankelovic proposes that those are not two distinct categories but spoken prompts rather fall along a continuum from implicit to explicit [15]. The most explicit form of prompts are directive prompts. They tell user the exact words they should say. Descriptive icons such as microphones or speech bubbles are one potential way to notify users to begin speaking [9]. Kamm concludes that directive prompts can faclitate the "ease of use" of voice interfaces [6].

Explicitly telling people what to do can potentially result in the exact opposite behavior. Prompts that are perceived as restricting to ones's freedom (to choose the input modality) can arouse reactance [3]. Reactance is an unpleasant motivational arousal that serves as a motivator to restore ones freedom e.g. by not following what the system suggests [11]. The extent to which a message is perceived as threatening to one's freedom finally influences peoples' behavior to follow or not follow the advice of the message [11].

2.3 Summary

Although research has shown that speech interaction leads to a safer and more efficient interaction, there are many situations where drivers do not decide to use their voice intuitively. We assume that this could be changed by providing a suitable trigger. Visual cues have some advantages over auditory cues that make them a promising means for triggering speech input. They can range from implicit hints to very explicit directive prompts. The latter ones are potentially more effective, yet they might draw too much of the driver's attention, or arouse reactance so that user will eventually not follow the system's advice.

3 Method

We conducted a user experiment that investigated the efficacy of visual cues to leverage speech input while driving. In order to address this research question in a differentiated way, we propose five hypotheses:

H1: Visual cues increase the amount of speech interactions.
H2: Explicit visual cues result in higher speech rates than implicit ones.
H3: Additional audio signals result in higher speech rates than only visual cues.
H4: Explicit visual cues cause higher visual distraction than implicit ones.
H5: Explicit visual cues induce a higher threat to freedom than implicit cues.

3.1 Participants

45 participants, 17 females and 28 males, with a mean age of 30.2 years ranging between 21 and 58 years took part in the study. All of them were either native German speakers or had excellent knowledge of the German language and none of the participants had motor impairments of the upper limbs, which would have shifted their decisions towards either touch or speech input. Participants' self-reported data showed about the same openness to touch and speech input with a slight advantage for touch. Tendencies to use rather speech or touch input while driving was balanced over all participants.

3.2 Experimental Design

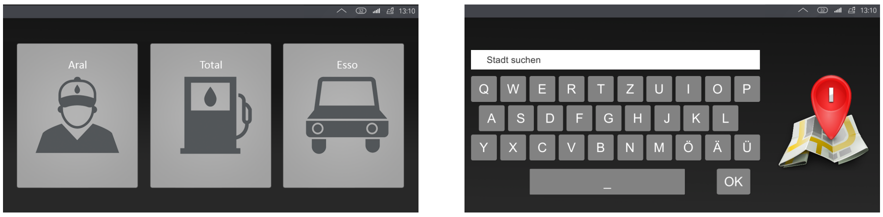
The experiment used a within-subject design. Each participant completed 64 tasks that were displayed on a secondary display while they were driving. For every task, participants had to decide whether to use speech or touch input. Tasks varied in presence and explicitness of visual cues (none, implicit cues, explicit cues, implicit and explicit cues). In order to create a greater generalizability of our results, we additionally included two task types (selection, text input) and two driving scenarios (easy, difficult) and varied the presence of an additional audio signal (none, audio). Each specific configuration occurred twice to each participant. All tasks were counterbalanced in order to prevent ordering effects.

In both driving scenarios, participants followed a leading vehicle on a highway with three lanes and slight curves. In the easy scenario, the leading vehicle moved with 100 km/h, it stayed on the rightmost lane and did not overtake. There was only few traffic. In the difficult scenario, there was more traffic. The leading vehicle moved at 130 km/h and it used all three lanes to overtake slower cars. The audio signal was the standard earcon of a current BMW 7 series for pressing the push-to-talk button on the steering wheel. Task types and visual cues will be explained in detail in the following sections.

3.3 Experimental Tasks

We used two task types in our experiment. The *selection task* is well suited to be solved with touch input, while the *text input task* is better solved using speech. By including these very different task types we aim to achieve a better generalizability of our results for a broader range of tasks. The speech recognizer was active as soon as a task appeared.

The goal of the *selection task* is to make a selection out of three elements. It is illustrated in Figure 1a. The task displayed either three gas stations or



(a) Selection task

(b) Text input task

Fig. 1: The selection task displayed three big elements that displayed gas stations (example in the figure) or restaurants. The text input task displayed an input field and a virtual keyboard to search a destination city (example in the figure) or a contact name.

three restaurants. Participants were instructed, which elements to select for the gas stations ("Total") and the restaurants ("Seehaus"). Selections were made by saying the name of the instructed element or by touching the according tile whenever this screen would appear.

In the *text input task*, participants had to enter a short text in form of a contact name or a destination. It is illustrated in Figure 1b. They were instructed to enter "Lisa" for contacts and "Jena" for the destination by either saying the requested entry or by typing it on the keyboard. Both instructed texts have four letters. We assume that current intelligent text input systems propose a small selection of possible words about three letters. They only require the user to tap a forth time to select out the correct proposition.

3.4 Visual Cues

In a preceding brainstorming session, we identified interface elements in touch based systems that users associate with the use of speech input. Identified elements were split in two groups: implicit cues and explicit cues.

Implicit cues are more subtle adaptations that refer to speech input without explicitly telling the user what to do. In the experiment, three types of adaptions were made when implicit cues were used. First, the highlighting of touch elements such as buttons was reduced. Touchable areas are often highlighted in brighter colors, which creates a visual stimulus that makes users more likely to touch them. Second, more emphasis was put in visible text on the screen by highlighting possible commands with quotation marks, making it easier for users to remember potential commands. Third, text was rephrased to be rather conversational and therefore promote a spoken answer. For example, instead of "Search city" the text input task displayed "Which city?".

Explicit cues, in contrast, are more noticeable and directly prompt the user to use speech input. Again, there were three adaptations made in conditions with explicit cues. First, a notification banner was displayed on the top of the screen to catch users' attention. Second, on the banner, there was a microphone symbol

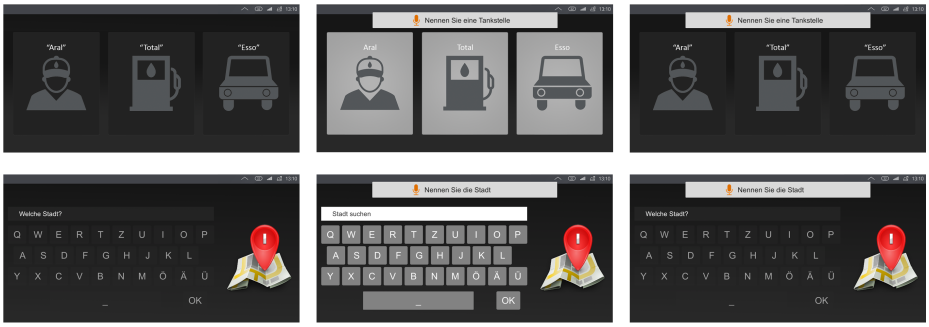


Fig. 2: Both tasks with increasing levels of visual cues. From left to right: implicit, explicit, implicit and explicit.

orange color. Third, there was a short text displayed, that prompted users to name the desired selection or text.

Figure 2 displays the application of implicit and explicit cues on the two experimental tasks. The most left picture illustrates the task with implicit cues (Imp). The next one shows the explicit cues (Exp). Finally, the most right picture integrates both, implicit and explicit cues (ImpExp). Together with the basic version of each task (see Figure 1), both rows illustrate four rising levels within the continuum from implicit (left) to explicit adaptions (right).

3.5 Apparatus

The experiment was conducted in a static high-fidelity driving simulator illustrated in Figure 3. The driving scene was projected on a 180 degree canvas in front of the vehicle mock-up. There were two displays in the cockpit: the instrument cluster displayed a speedometer and rounds per minute, the central information display in the dashboard showed the experimental tasks. The latter was a 10.1 inch *Faytech capacitive touch display*[1] with a resolution of 1280x800 pixels. The experimental tasks were integrated in a special application implemented in *Unity3D*. Speech recognition was achieved using the built-in speech engine in *Unity3D* which uses the Windows speech recognition engine in combination with a *Rode SmartLav+*[2] clip on microphone. The users' glance behavior was recorded with *Dikablis Essential*[3] eye tracking glasses in combination with infrared markers.

[1] https://www.faytech.com/de/katalog/product/101-capacitive-touch-monitor-ft10wtmbcap/

[2] http://de.rode.com/microphones/smartlav

[3] http://www.ergoneers.com/eye-tracking

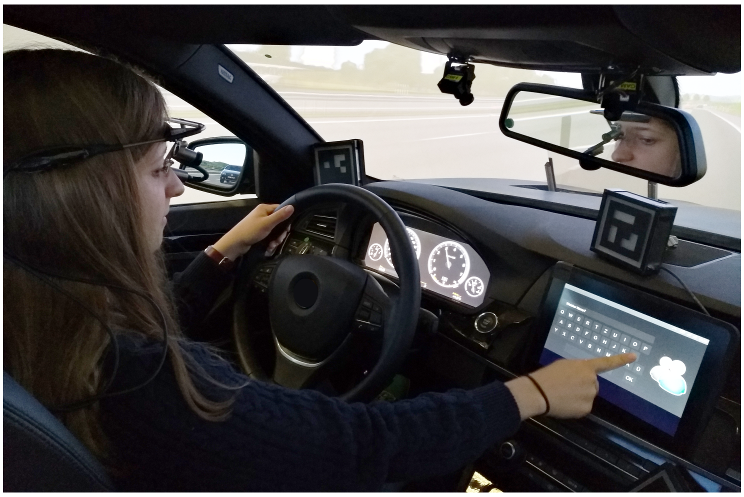


Fig. 3: The cockpit in the experimental setup. The experimental tasks were displayed on the central display and participants decided whether to use touch or speech for interaction. Glance behavior was recorded using a head-mounted eye tracker.

3.6 Procedure

Participants completed a short form covering demographic data before they were introduced to the experimental tasks. They were shown all tasks (selection-gas stations, selection-restaurant, text-contacts, text-destination) in the basic version, without any visual cues and without an audio signal (as illustrated in Figure 1). They were instructed to memorize the correct selection for each of the four tasks. Participants were *not* told that there will be additional visual cues, but the examiner emphasized that participants *always* have the choice to use either touch or speech input. Tasks appeared automatically on the central display after a random wait time between 10 and 15 seconds. This varying wait time avoided that participants got used to a certain rhythm, and ensured that there was sufficient time between tasks, so that each task was handled independent of the previous one. As soon as the task was displayed participants looked at the screen to identify the task, decided whether to use touch or speech, and made their input. Tasks disappeared after the selection or text input was completed and participants turned back to driving until the next task appeared. After all tasks had been completed, participants were shown all possible combinations of task type, visual cue and audio signal on a laptop display without driving. This way they could concentrate on the illustration of tasks. For each specific illustration, they rated the suitability of touch and speech input, as well as the threat to freedom. The order in which tasks appeared was counterbalanced.

3.7 Data

There was a total number of 2880 choices (45 participants*4 visual cues*2 scenarios*2 task types*2 audio signals*2 choices per configuration) and 90 choices for one specific configuration. The Eye Tracking system recorded the total glance time per task, which is the average duration that a participant looked on the display while a task was active. Finally, there were participants' self-reported assessments about the perceived threat to freedom that is caused by a specific illustration of a task. They are based on the ratings of four items, each on a 5-point Likert scale from -2 (strongly disagree) to 2 (strongly agree) [3].

4 Results

4.1 Choice of Input Modality

The choice of input modality was encoded in a binary variable (0 = touch input, 1 = speech input). Figure 4 illustrates the percentage of speech inputs depending on the the visual cue, the driving scenario, the task, and on the occurrence of an audio signal. The percentage of speech input grew with increasing level of the visual cues. The maximum increase was 16% for the selection task and 15% for the text task. This effect can be observed for both task types and both driving scenarios. The results of a Friedman test show that the visual cues had a significant influence on the choice of input modality ($\chi^2 = 13,904, p = .003, r = 2.07$). Additional Wilcoxon signed-rank tests were used to compare implicit cues to those conditions with explicit cues. They show that only explicit cues ($Mdn = 0.69$) did not result in significantly higher percentage of speech input than implicit cues ($Mdn = 0.56$). Instead, implicit and explicit cues ($Mdn = 0.69$) led to a significant rise of the speech rates compared to implicit cues ($Z = -2.48, p = .013, r = 0.37$). The level of significance was corrected according to Bonferroni.

Additionally, a logistic regression was performed to analyze the influence of all factors on the participants' modality choice. The results show that both the logistic regression model $\chi^2(4) = 350.00, p < .001$, as well as the individual coefficients (except the audio signal) were statistically significant. The model correctly classified 66.8% of the cases. Increasing the explicitness of visual cues by one level rises the relative probability to choose speech input by 17.4%. In the difficult driving scenario the relative probability to choose speech is 54.2% higher than in the easy one. Finally, the task-type had the greatest impact. Choosing speech for text input was 290.0% more likely than for the selection task. R^2 (Nagelwerke R square) is 0.155, which indicates a strong effect [2].

In line with these findings, a Wilcoxon signed-rank test between conditions with acoustic signal ($Mdn = 0.59$) and without acoustic signal ($Mdn = 0.56$) showed no significant differences ($Z = -0.87, ns.$).

We can summarize that the task type was the most decisive coefficient, followed by the driving scenario. Visual cues play a smaller, yet decisive role in

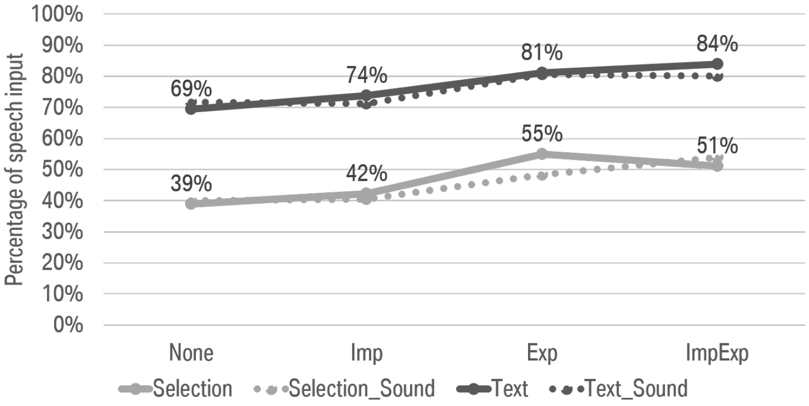


Fig. 4: The percentage of speech input depending on the visual cues, the task-type and the audio signal (dotted line).

influencing the participants' decision. Moreover, the model shows that the probability to choose speech input rises with the level of visual cues. The additional audio signal did not influence the participants' decisions.

4.2 Glance Duration on the Display

The average glance duration in both driving scenarios was between 0.83 and 0.89 seconds for the selection task and between 1.41 and 1.83 seconds for the text entry task. Figure 5 illustrates the total glance times for both tasks individually. We observe a light tendency that glance times decrease with increasing level of visual for the text input task, while the selection task remains constant. Glance data was not normally distributed. Results of a Friedman test showed that the average glance duration on the display was not significantly affected by the visual cues ($\chi^2 = 2.32, ns.$). Wilcoxon tests confirmed that the total glance duration without visual cues did not significantly differ between implicit cues ($Mdn = 1.03$) compared to Explicit cues ($Mdn = 1.13$) or implicit and explicit cues ($Mdn = 0.99$). The duration that participants looked on the display did not depend on the type or presence of visual cues.

4.3 Perceived Threat to Freedom

The perceived threat to freedom was measured with the mentioned four questions. Ratings were very diverse because some participants did not feel any restriction, while others reported that they felt like being influenced or urged to behave in certain way. The data was not normally distributed. A Friedman test indicted that the average ratings were not significantly affected by the visual cues ($\chi^2 = 5.49, ns.$). Accordingly, results of Wilcoxon tests showed that explicit cues ($Mdn = 0.00$) or implicit and explicit cues ($Mdn = 0.19$) were not associated with a higher threat to freedom compared to implicit cues ($Mdn = 0.00$).

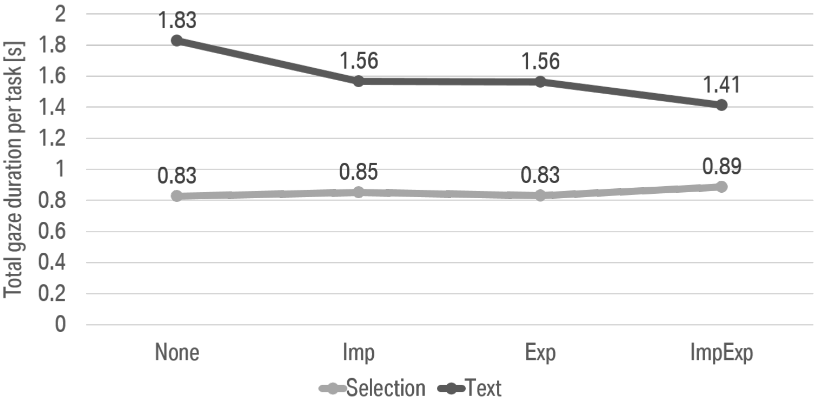


Fig. 5: Total glance time (TGT) on the display while a task was active. Visual cues did not significantly affect TGT.

5 Discussion

The first hypothesis H1 proposes that visual cues increase the amount of speech input used. Our results show that the user's choice of input modality was mainly determined by the task type and the driving scenario. Still, the visual cues had a significant influence, which can be classified as a strong effect based on the estimated effect size of the Friedman test [2]. Furthermore, the results of the logistic regression model and Figure 4 indicate that implicit cues can already increase speech usage and we can accept H1.

H2 claimed that explicit cues would be more effective to promote speech input than implicit ones. The logistic regression model supports this thesis, but additional pairwise comparisons showed that the increase of speech rates for explicit cues compared to implicit cues was not significant. Based on these findings we do not accept H2. However, *extending* implicit cues by explicit ones (ImpExp) led to a significant increase of speech interaction. This suggests that the effects of implicit and explicit cues complement one another. The combination of both led to overall highest speech input rates.

H3 proposed that additional audio signals increase the amount of speech input used. However, the results show that they did not have a statistically significant influence on the participants' decisions. This was surprising, given the fact that speech input is mostly prompted using audio signals. At the same time, it is in line with the disadvantages of the temporal and short term nature of audio [1]. The audio signals were played the moment the task appeared on the screen, but participants often needed a couple of seconds to control the vehicle before attending to the task. The trigger existed, but it was not well-timed, which is one possible reason why achieving the target behavior fails [5].

H4 proposed that explicit cues cause increased visual distraction compared to implicit ones. The results did not reveal significant differences between the four levels of visual cues. A deeper look into glance data shows different trends for glance behavior depending on the task in Figure 5. For the text input task, the usage of speech input rises with increasing level of visual cues while the average glance times for both, speech and touch decreases. This means that not the actual glance times per task changed, but rather the amount of (less visually distracting) speech inputs rose, which led to an overall decrease of the total glance time. The fact that this affects only the text input task shows that glance times for touch and speech input for the selection task were similar, since the higher percentage of speech selections did not reduce the overall glance duration. In summary, explicit cues did not result in longer or more glances on the display, but they reduced the overall visual distraction by increasing the amount of speech usage. For these reasons, we do not accept H4.

H5 assumed that explicit visual cues induce a higher threat to freedom than implicit ones, which increases the likeliness to show reactance and that participants will not follow the systems' advice. The average ratings from participants' self-assessed threat to freedom did not differ significantly between conditions. This indicates that the design of our visual cues did not have a big influence on the perceived freedom to choose themselves. A limitating factor might be that participants were explicitly told that they can always decide freely. Moreover, previous work in this field notes under-reporting as potential problem for participants' self-reported data. This might also contribute the missing variance in this case [7]. Therefore, we do not accept H5.

6 Conclusion

In this experiment, we explored the influence of visual cues on the users' choice whether to use speech or touch input. Our results show that visual cues can significantly contribute to leverage speech input while driving. This effect can be observed across different task types and different driving scenarios. At the same time, visual cues did not cause increased visual distraction. In contrast, there is a tendency that the overall glance time away from the street can be reduced for text input tasks by using explicit visual cues. We conclude that visual cues are an effective means to influence the user's choice of input modality and thereby to support users by emphasizing suited input modalities. The system can guide users in an unobtrusive way so that they can benefit from the whole range of input modalities, without concerning themselves with the decision. Our study showed that visual cues increased the amount of speech input used, decreased visual distraction on the road, and thereby contributed to the driver's safety.

While this experiment was conducted in the automotive domain, our results can be applied in other domains that offer speech along with other input modalities such as smartphones or smart home devices. This experiment served as a first try to test the potential of simple graphical adaptions. Future experiments should explore the full potential of visual cues, to see if incorporating animations or context sensitivity can further increase their persuasive influence.

References

1. James H. Bradford and James H. The human factors of speech-based interfaces. *ACM SIGCHI Bulletin*, 27(2):61–67, 1995.
2. Jacob Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992.
3. James Price Dillard and Lijiang Shen. On the Nature of Reactance and its Role in Persuasive Health Communication. *Communication Monographs*, 72(2):144–168, 2005.
4. B. J. Fogg. A Behavior Model for Persuasive Design. In *Proceedings of the 4th International Conference on Persuasive Technology*, page 1, New York, New York, USA, 2009. ACM Press.
5. B. J. Fogg. Creating Persuasive Technologies : An Eight-Step Design Process. In *Proceedings of the 4th International Conference on Persuasive Technology*, volume 91, pages 1–6, 2009.
6. Candace Kamm. User interfaces for voice applications. *Proceedings of the National Academy of Sciences*, 92(22):10031–10037, 1995.
7. Brenda Miranda, Chimwemwe Jere, Olayan Alharbi, Sri Lakshmi, Yasser Khouja, and Samir Chatterjee. Examining the efficacy of a persuasive technology package in reducing texting and driving behavior. In *PERSUASIVE 2013LNCS*, volume 7822, pages 137–148, 2013.
8. Avi Parush. Speech-Based Interaction in Multitask Conditions: Impact of Prompt Modality. *Human Factors*, 47(3):591–597, oct 2005.
9. Leah M. Reeves, Jean-Claude Martin, Michael McTear, TV Raman, Kay M. Stanney, Hui Su, Qian Ying Wang, Jennifer Lai, James A. Larson, Sharon Oviatt, T. S. Balaji, Stéphanie Buisine, Penny Collings, Phil Cohen, and Ben Kraal. Guidelines for multimodal user interface design. *Communications of the ACM*, 47(1):57–59, 2004.
10. Florian Roider, Sonja Rümelin, Bastian Pfleging, and Tom Gross. The Effects of Situational Demands on Gaze, Speech and Gesture Input in the Vehicle. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 94–102, New York, USA, 2017. ACM Press.
11. Christina Steindl, Eva Jonas, Sandra Sittenthaler, Eva Traut-Mattausch, and Jeff Greenberg. Understanding psychological reactance: New developments and findings. *Journal of Psychology*, 223(4):205–214, 2015.
12. David L Strayer, Frank A Drews, and Dennis J Crouch. A comparison of the cell phone driver and the drunk driver. *Human factors*, 48(2):381–91, 2006.
13. Warren H. Teichner and Marjorie J. Krebs. Laws of visual choice reaction time. *Psychological Review*, 81(1):75–98, 1974.
14. Christopher D. Wickens, Diane L. Sandry, and Michael Vidulich. Compatibility and resource competition between modalities of input, central processing, and output. *Human factors*, 25(2):227–248, 1983.
15. Nicole Yankelovich. How Do Users Know What to Say? *Interactions*, 3(6):32–43, 1996.