Just Look:The Benefits of Gaze-Activated Voice Input in the Car

Florian Roider
BMW Group Research, New Technologies, Innovations
Munich, Germany
florian.roider@bmw.de

Lars Reisig
BMW Group Research, New Technologies, Innovations
Munich, Germany
lars.reisig@bmw.de

Tom Gross
Human-Computer Interaction Group, University of Bamberg
Bamberg, Germany
tom.gross@uni-bamberg.de

ACM.
AutomotiveUI '18 Adjunct,, September 23–25, 2018, Toronto, ON, Canada
ACM 978-1-4503-5947-4/18/09.
https://doi.org/10.1145/3239092.3265968

Abstract
Voice interaction provides a natural and efficient form of communication with our cars. Current vehicles require the driver to push a button or to utter an artificial keyword before they can use speech input. This limits the potential naturalness and efficiency of voice input. In human communication, we usually use eye contact to express our intention to communicate with others. We conducted a user study with 25 participants that investigated gaze as a means to activate speech input while being occupied with a primary task. Our results indicated a strong dependency on the task. For tasks that refer to information on the screen, gaze activation was superior to push-to-talk and keyword, but it was less valuable if the task had no relation to screen content. We conclude that gaze cannot replace other modes for activation, but it can boost efficiency and user experience for display related tasks.

Author Keywords
Gaze; intelligent environment; multimodal user interface; voice.

CCS Concepts
•Human-centered computing → **User studies; Usability testing;** *Interaction techniques;*

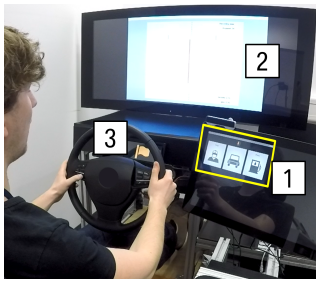

Figure 1: The prototype had a gaze-aware screen area (1) to the right of the steering wheel. Size and location of this area was derived from displays in current vehicles. During the experiment, users also had a primary task (2) and a small display in front of them for instructions (3).

Introduction

Voice input provides the potential for safe, efficient and natural in-car interaction (e.g. [1, 3, 7]). Therefore, effort is put into the improvement of automatic speech recognition (ASR) systems, to increase recognition accuracy and to integrate more functionality. Yet, before users can actually make a voice command, all current voice recognition systems require an explicit activation to express that the user's command was directed to the system. The most widespread option in current vehicles is the use of a push-to-talk (PTT) button on the steering wheel. However, the activation via a haptic button is an artificial action that is in contrast to the naturalness and intelligence of (future) voice interaction systems and does not live up to the potential of voice interaction. In this regard, the activation via keyword (e.g. "Hey Mercedes,..."[1]) is a much more natural option that also appears in human communication (e.g. "Hey Bob,..."). However, compared to a simple button press, the uttering of keyword takes relatively long and might be annoying to the user for multiple requests. In summary, there is a trade-off between speed (PTT) and naturalness (Keyword) for currently used voice activation techniques.

Related Work

In interhuman communication, people use a different mode to quickly express their focus of attention. It has been shown that eye-contact most reliably indicates the target of a users' attention [8]. Maglio et al. further investigated gaze patterns during the communication with multiple intelligent devices. They found that users nearly always looked at a device before making a request [4]. Based on that, Oh et al. investigated the use of gaze to activate the automatic speech recognition to enable more natural human-computer interaction. They compared this look-to-talk (LTT) approach, compared to a keyword-based implementation and PTT and concluded that LTT was a promising approach (under ideal conditions, i.e. good accuracy, short latency) [5]. For applications in the driving context, it must be considered that drivers are in a dual-task situation with a visually demanding primary task. Their eye-gaze is mainly focused on the street, even when talking to a passengers next to them. Thus, the results about the potential of LTT in a conversational setting cannot directly be transferred to the automotive domain. Moreover, we have to take additional aspects into account besides the naturalness of interaction, such as distraction and efficiency.

In this paper, we present a comparison of PTT, Keyword and LTT for the activation of a speech recognition system while being occupied with a visually demanding primary task. We analyze visual distraction, efficiency, and user experience.

Prototype

We implemented a prototype that supported all three activation modes in an automotive setup. It is displayed in Figure 1. A directional microphone was used for recording voice commands. The software prototype ran on a Windows 10 machine and was implemented in Unity3D, which uses the integrated speech engine in Windows. A Tobii 4C[2] eye-tracker was mounted behind the steering wheel to track the user's gaze .

The prototype supported two types of secondary tasks. They are displayed in Figure 2. In the *navigation* task, users chose one out of three gas stations (e.g. "Navigate to Esso"), which were displayed on the gaze-aware screen area. In the *phone* task, users made a phone call to a person (e.g. "Call Lisa"). Before making a voice request, users had to activate the ASR by either pressing the PTT button (on the right side of the steering wheel), by saying the keyword,

[1] https://www.mercedes-benz.com

[2] https://tobiigaming.com/product/tobii-eye-tracker-4c/

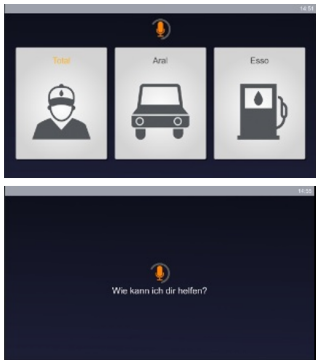

Figure 2: In the *navigation task* users chose one out of three gas stations (top). In the *phone task* users made a phone call (bottom).

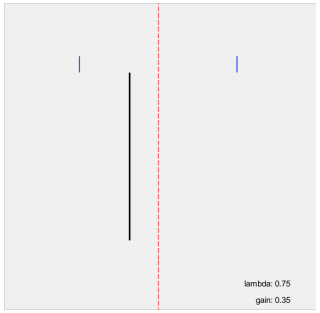

Figure 3: The aim of the CTT is to control a vertical line that is constantly fading away from the center position. Participants use two buttons on the steering wheel to move the line left and right.

or by looking at the screen area for 300 ms. This duration approximately represents the duration of a short eye fixation [2]. An animated microphone icon showed the state of the ASR and served as a visual representation to support making eye-contact with the screen. Upon activation, the icon animation started and the system played an earcon ("pling"). It stayed active for at least four seconds, so that users could look back on the street after the activation. Upon recognition of the voice command there was another earcon ("plong") and the task disappeared.

Experiment

A total of 25 participants (9 female) with mean age of 30.24 years ($SD = 11.53$) took part in the experiment. We used a within-subject design with repeated measures. There were two independent variables: three modes of speech activation (PTT, Keyword, LTT) and two task-types (navigation, phone). Participants completed one trial for each combination of speech activation and task-type. The order of appearance of task-types and activation modes was counterbalanced over all participants.

We used the critical tracking task (CTT) as a primary task in order to create a constant visual demand on the participants [6]. It was displayed on a screen in front of the participants (see Figure 1). Another small display was placed behind the steering wheel. It was used to instruct the name of the person to call for the phone task, so that participants did not have to look to the screen area. In the navigation task, the instruction was made directly on the screen area to the right. Users were asked to pick the gas station whose name was written in orange.

The experiment started with the adjustment of the seat position and the calibration of the eye-tracker. The examiner explained the CTT and the two task types with the different activation techniques. All participants practiced the CTT

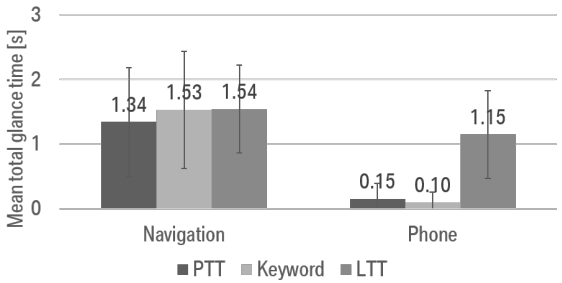

Figure 4: Mean total glance times hardly differed in the navigation task. In the phone task, LTT led to increased visual distraction.

and the secondary tasks and quickly familiarized with it. Before each trial, they were informed about the upcoming task and activation technique. One trial consisted of ten voice commands with ten second breaks between a successful command and the instruction of the next task. We recorded the participants' total glance time on the screen, the task completion times to asses the efficiency of each mode and participants completed the User Experience Questionnaire (UEQ).

Results and Discussion

Visual Distraction

The total glance time (TGT) is the aggregated duration of glances on the gaze-aware screen area while a task was active. In the navigation task, TGT hardly differed between PTT, Keyword, and LTT (see Figure 4). In all three conditions the participants had to glance at least once at the display to see the selectable elements. The fact that LTT did not result in a longer TGT away from the primary task shows that the naturally occurring glances (that also appear in the PTT or Keyword condition) can be efficiently used

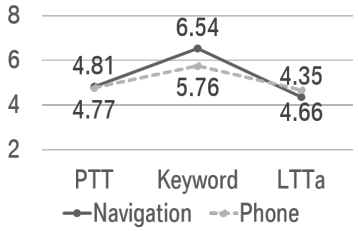

Figure 5: Mean task completion times.

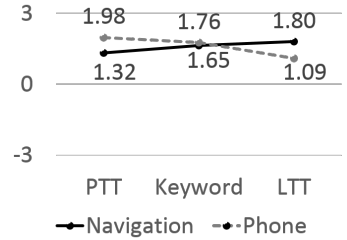

Figure 6: Mean UEQ ratings regarding the pragmatic quality.

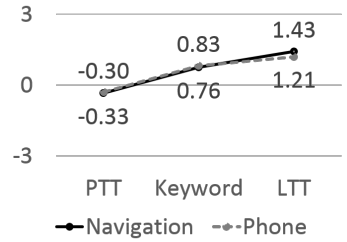

Figure 7: Mean UEQ ratings regarding the hedonic quality.

for activation. In contrast, in the phone task the activation mode had a significant influence on TGT (Friedman test: $F(2) = 39.56, p < .01$). LTT was significantly longer than PTT ($p < .01$) and Keyword ($p < .01$). PTT and Keyword did not need glances on the screen, because the task does not refer to on-screen information. LTT however needs a short glance at the screen to activate system. Although TGT for LTT is relatively short ($M = 1.15$) these glances are not naturally occurring, but rather an additional, artificial movement with the eyes. Despite the effects on glance behavior, the mean CTT scores for both tasks did not significantly differ across activation techniques.

Efficiency

In both tasks, the shortest task completion time (TCT) was achieved with LTT, followed by PTT, and keyword. They are illustrated in Figure 5. A repeated measures ANOVA showed that the activation mode had a significant effect on TCT ($F(2, 48) = 32.82, p < .01, \eta^2 = .58$). TCT with the keyword was significantly slower than LTT (p < .01) or PTT (p < .01), which is probably due to the time that is needed to speak and process the keyword. PTT and LTT allow a much faster and more direct activation.

User Experience

The dimensions of the UEQ can be summarized regarding the pragmatic and the hedonic quality of the interaction. In Figure 6, we see that the pragmatic quality was depending on the task type. For the phone task, LTT had a lower pragmatic value than the Keyword ($p < .05$) and PTT ($p < .01$), but for the navigation task the mean LTT rating was higher than the other modes ($ns.$). The pragmatic value of the keyword activation was between PTT and LTT for both tasks. In the navigation task, LTT made use of the naturally occurring glances to activate ASR without additional effort of the user. Thus, it is perceived as a very pragmatic solution. In contrast, in the navigation task, LTT required users to make one extra glance at the screen area. This is perceived as a small, yet extra effort and thus less pragmatic.

For the hedonic quality (Figure 7), the activation mode had the same effect on both tasks ($F(2, 48) = 43.80, p < .01, \eta^2 = .65$). LTT had the highest ratings, followed by Keyword. PTT was rated significantly worse (both $p < .01$). Hedonic quality represents attributes like stimulation and novelty. While the haptic button press for PTT has been well known in cars for years and keyword activation can be found in many current consumer electronic devices (though relatively new in cars), LTT was a new experience for almost all participants. Although it has some obvious pragmatic disadvantages for the phone task, it is noticeable that this did not impair the high perceived hedonic quality of the approach.

Conclusion

Our results show that the benefit of LTT is strongly depending on the task type. For tasks that do not require glances at the screen, the benefit of LTT is small. There are short task times and a good hedonic quality at the cost of increased visual distraction and lower pragmatic value than PTT or Keyword. However, if users have to glance at the screen for a specific task, LTT has the potential to increase efficiency and user experience (pragmatic and hedonic) of the interaction, without creating additional visual distraction. We conclude that LTT can not replace other activation modes for voice input in current vehicles, but it could be a valuable addition that allows fast and user-friendly interaction with on-screen content. Future work will have to address the application in more realistic scenarios and the proneness for inadvertent activation. With the coming automation in future vehicles and the decrease of importance of visual distraction, the value of gaze-activated voice input could further grow.

REFERENCES

1. Ignacio Alvarez, A Martin, Jerone Dunbar, J Taiber, D M Wilson, and Juan E Gilbert. 2011. Designing Driver-Centric Natural Voice User Interfaces. In *Adjunct Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 42–49.

2. Robert J. K. Jacob and Robert J. K. 1991. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems* 9, 3 (apr 1991), 152–169. DOI: http://dx.doi.org/10.1145/123078.128728

3. Jannette Maciej and Mark Vollrath. 2009. Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis & Prevention* 41, 5 (sep 2009), 924–930. DOI: http://dx.doi.org/10.1016/J.AAP.2009.05.007

4. Paul P Maglio, Teenie Matlock, Christopher S Campbell, Shumin Zhai, and Barton a Smith. 2000. Gaze and Speech in Attentive User Interfaces. In *Proceedings of the Third International Conference on Multimodal Interfaces*, Vol. 1948. 1–7. DOI: http://dx.doi.org/10.1007/3-540-40063-X_1

5. Alice Oh, Harold Fox, Max Van Kleek, Aaron Adler, Krzysztof Gajos, Louis-Philippe Morency, and Trevor Darrell. 2002. Evaluating look-to-talk. In *Extended Abstracts on Human Factors in Computing Systems*. ACM Press, New York, New York, USA, 650. DOI: http://dx.doi.org/10.1145/506443.506528

6. Tibor Petzoldt, Hanna Bellem, and Josef F. Krems. 2014. The Critical Tracking Task: A Potentially Useful Method to Assess Driver Distraction? *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56, 4 (2014), 789–808. DOI: http://dx.doi.org/10.1177/0018720813501864

7. C.a. Carl a. Pickering, K.J. Keith J. Burnham, and Michael J. M.J. Richardson. 2007. A Review of Automotive Human Machine Interface Technologies and Techniques to Reduce Driver Distraction. In *2nd IET International Conference on System Safety*. IEEE, 223–228. DOI: http://dx.doi.org/10.1049/cp:20070468

8. Roel Vertegaal and Jeffrey S. Shell. 2008. Attentive User Interfaces: The Surveillance and Sousveillance of Gaze-Aware Objects. *Social Science Information* 47, 3 (sep 2008), 275–298. DOI: http://dx.doi.org/10.1177/0539018408092574