TECHNISCHE UNIVERSITÄT CHEMNITZ

Faculty of Mechanical Engineering

Professorship of Ergonomics and Innovation Management

Univ.-Prof. Dr. Angelika Bullinger-Hoffmann

Professorship of Machine Tool Design and Forming Technology

Prof. Dr.-Ing. Matthias Putz

Do I look at what I'm saying?

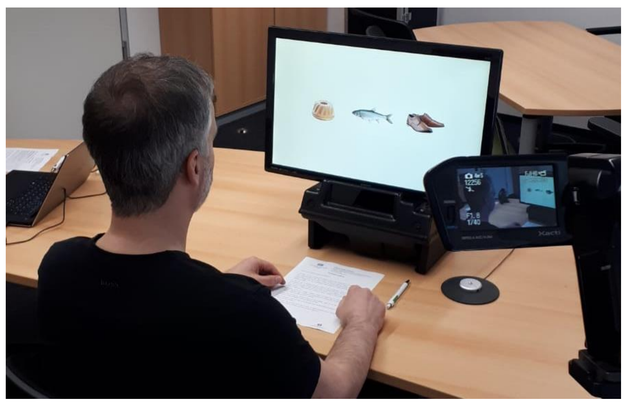
Danny Rüffert, Alexander Kögel, Jennifer Brade, Daniel Liebscher, Philipp Klimant, Frank Dittrich

Introduction

The use of VR and AR technologies demands new forms of human-technology interaction, since classical input devices such as keyboard or mouse do not meet the requirements of a usable application in a virtual context. Natural forms of interaction such as speech and gaze have the potential to close this gap. However, the technologies are often flawed when used individually. For example, speech recognition systems have problems with dialects, ambiguities and deixis. The combination of speech and gaze as multimodal input can reduce errors because the focus of attention (gaze) is related to speech. The authors ask themselves the question to what extent the gaze behavior changes with a repeated speech selection of the same object, if the speech recognition does not understand the first input correctly.

Experiment

To answer the research question we conducted a within-subject design were the participants saw different objects and had to select one of them via simulated speech recognition by enunciating the color (sequence 1 and 2) or the object itself (sequence 3). Randomly the speech recognition system did not understand the speech input and the participants had to repeat their selection. For the study a LC Technologies eye tracker, with a sampling rate of 120 Hz and a accuracy of 0.4 degree in conjunction with an 24 inch monitor at a distance of approx. 65 cm was used. This monitor was used as the presentation screen while the investigator controlled the experiment from a laptop. NYAN 3 was used as the eye tracking software. The experimental setup is presented in figure 1 (left). The test started with a tutorial, which introduced the participants to the task and the two different audio outputs of the systems (correct recognition and the request to repeat the speech input). Afterwards, the participants solved three sequences with different object counts (two to six) and types. For the sequences 1 and 2, different geometrical objects in five different colors were used, all having a contrast ratio of 7,9:1 in respect to the background. In sequence 1, 2 to 3 objects per image were presented and 4 to 6 in sequence 2. For sequence 3, different pictures of real objects, which should be nameable by an 2-year old child following the SBE-2-KT test, were presented. After each sequence the participants also rated the effort of the task.



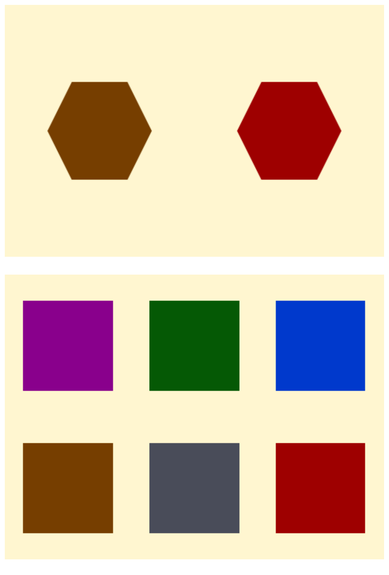


Figure 1: Experimental setup with sequence 3 shown on screen (left) and examples of objects presented during sequence 1 (top right) and sequence 2 (bottom right)

Results

Overall 5 woman and 5 men attended the study, all of them were employees of the University of Technology Chemnitz. None of the them had an eye disease, but 60% wear glasses, which were not worn during the study. Because of the small study sample, no significant tests were calculated. To measure the effort of the task, the effort scale by Nachreiner et al. was used, which measures the effort on an continuous, one-dimensional, graphical scale from 0 to 220. The effort of the three sequences was ranked nearly equal and showed that the task was not exhausting for the participants. Still, the repeated request of the system to repeat the input was experienced as strenuous. For the evaluation, we analyzed the areas of interest on the presented screens in the timeframe the participants enunciated the color or the object itself. We then calculated which object was primary considered and counted the times of matches between speech-selected and gazed-at objects. This procedure was replicated for the repeated speech selection and for all sequences – the mean values are presented in figure 2. *First choice* refers to the first selection while *second choice* represents the repeated speech selection results.

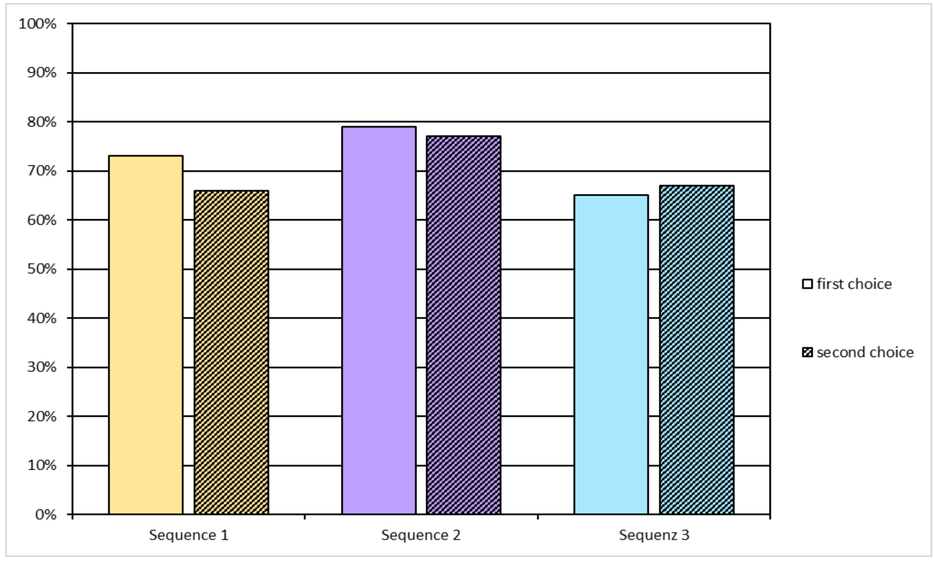


Figure 2: Means of the match between speech selected object and object being gazed at

Conclusion

Although no statistical tests have been performed, tendencies can be identified. When naming simple things (color), the match decreases with repetition. For complex objects, the consistency of the repeated selection increases. With a larger sample, the trends should be statistically verified. Allover there is a high match rate between the object being gazed at and the speech selected object in general. Therefore, a multimodal input system which supports the speech recognition with eye tracking results during the pronunciation to counteract ambiguity of the speech input is highly promising. To develop an user friendly multimodal input system, we focus on repeating our study with an head-mounted see-through display with implemented eye tracker.

Acknowledgements

This project is co-financed with tax money based on the state budget, passed by the representatives of the Saxon Landtag.

Europäische Union

Europa fördert Sachsen. ESF Europäischer Sozialfonds

Diese Maßnahme wird mitfinanziert durch Steuermittel auf Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

Contact: Alexander Kögel
phone: +49 (0)371 – 531 33020
e-mail: alexander.kögel@mb.tu-chemnitz.de

IWP INSTITUTE OF MACHINE TOOLS AND PRODUCTION PROCESSES

ARBEITSWISSENSCHAFT UND INNOVATIONSMANAGEMENT