VoxVisio – Combining Gaze and Speech for Accessible HCI

David Rozado, Alexander McNeill and Daniel Mazur
Otago Polytechnic

ABSTRACT

In this work we introduce an open source software called VoxVisio intended to help motor impaired subjects to efficiently interact with a computer hands-free by combining gaze and speech interaction. The resulting multimodal system allows interaction with a user interface by means of gaze pointing for target selection and subsequent speech commands for target specific action. VoxVisio allows customized maps of speech commands to specific interaction commands such as: left mouse click, right mouse click or page scroll. With VoxVisio, the user gazes at the object in the user interface with which it wants to interact and then triggers a target-specific action by carrying out a specific speech command. Initial assessment of this novel interaction modality suggests that the proposed interaction paradigm holds the potential to improve the performance of traditional accessibility options such as speech only interaction and gaze only interaction. We make the VoxVisio software freely available to the community so the output of our research can help the target audience.

INTRODUCTION

Users with a motor disability are often unable or severely handicapped to use the mouse and/or the keyboard (Bates & Istance, 2003). Accessibility solutions for motor disabilities include: speech recognition, gaze tracking, alternative keyboard layouts, Brain Computer Interfaces (BCI), mechanical switch activated interfaces and other alternative pointing and clicking devices (Cook & Polgar, 2014). Unfortunately, the performance of such technologies in terms of the information transfer rate between the human and the computer is often sub-optimal since they tend to lag behind the accuracy and latency metrics achieved by standard keyboard and mouse interaction(Yuan et al., 2013).

Speech interaction although still struggling with accuracy issues, is very appropriate for continuous speech input such as when dictating an email or word document but highly impractical for pointing and selection tasks which are ubiquitous in standard GUI based interfaces. Gaze interaction is a very promising interaction modality for subjects with motor disability since eye movements are relatively spared in spinal cord traumatic injuries and diseases such as motor neuron disease or muscular dystrophy (Istance et al., 1996). However, this modality of interaction has its own drawbacks such as an accuracy limit of about 0.5 degrees of visual angle (Rozado, 2013) which renders interaction with small icons impractical. Additionally, humans are used to employing gaze for pointing at targets of interest but not for using gaze as an interaction effector (Rozado, Moreno, San Agustin, Rodriguez, & Varona, 2013). Thereby, gaze only interfaces are plagued with false positives activation due to the user just gazing at an object to gather information about it rather than out of a desire to interact with it (Hales, Rozado, & Mardanbegi, 2013). Additionally, the long acquisition times required by dwell interaction to acquire a target have been described as frustrating by the users (Rauterberg, Menozzi, & Wesson, 2003). As such, the usage of switches to asynchronously signal click actions has been suggested as a complement to gaze-only interaction(Grauman, Betke, Lombardi, Gips, & Bradski, 2003). The usage of a switch however, is not always feasible for users with very advanced degrees of motor disabilities such as patients with locked in syndrome. Furthermore, mechanical switch based interaction does not scale well, given that the degree of motor impairments in potential users limits the amount and type of switches they can operate (Grauman et al., 2003). Hence, gaze only or single-switch assisted gaze interaction are still considerable slower and more error prone than standard mouse based interaction.

The accuracy limitations of gaze interaction has often been circumvented in the past by the usage of customized software with large interaction elements which are gaze responsive. This however severely limits the range of software available for users dependent on gaze for interaction with computers. An alternative

to customized gaze responsive software, has been to make a pointer on the screen such as the traditional mouse cursor follow the gaze of the user. Unfortunately, this approach has been described as inconvenient for the user since the drift between actual gaze position and inferred gaze position is distracting for the user.

An innovative interaction paradigm to make gaze interaction possible with traditional GUIs while still affording the selection of small targets by using a zoom mechanism was proposed by (Pomplun, Ivanovic, Reingold, & Shen, 2001). This solution has been recently implemented by commercial developers of gaze interaction software such as Tobii with its Windows Gaze Control Software, see Figure 1. This selection modality consists of two parts. In the first step, the user selects a desired interaction task by looking at an icon in a docked taskbar on the edge of the monitor that displays different available interaction tasks (right click, left click, scroll, etc). In the second step, the user gazes at the desired portion of the screen on which it wants to execute the previously selected task. A dynamic zoom towards the area surrounding the region of interest gazed at is then triggered by a mechanical switch or by using a dwell time threshold activation. The user can steer the dynamic zoom towards the desired target by gaze to allow for fine grained target selection. The process ends with the execution of the previously gaze selected task upon the final destination of the zooming trajectory. Such two-step gaze selection mechanism makes it possible for the user to control a standard Windows desktop operating system. Unfortunately, the requirement to gaze select a command in the taskbar before interacting with a target makes such a system still slower than traditional interaction methods. In this work, we propose to eliminate the need to gaze select a command prior to engaging a target by rather simply gazing at the desired target and use speech commands as a multi-switch system for target specific interaction.

To sum up, in this work we introduce VoxVisio, a software for multimodal interaction consisting of combining gaze with speech. VoxVisio allows customization of the mapping between speech commands and control commands such as left mouse click, right mouse click, scroll, page down, enter, etc through its user interface. The combination of speech commands with gaze targeting creates an innovative multi-modal interaction method where interface items being gazed at react seamlessly to specific speech commands.

VOXVISIO

For interested readers, video demos of the VoxVisio software being used to interact with a computer hands-free using the combination of gaze and speech can be found at: https://www.youtube.com/watch?v=nHpkNqTqpCA
The VoxVisio software can be downloaded at: https://github.com/AlexanderMcNeill/voxvisio.

Hardware requirements

A Tobii Eye X eye tracker was used for gaze tracking during development. Gaze samples were sampled at 30 frames/s. Gaze estimation accuracy was around 0.5 degrees of visual angle at 60 cm from the screen. A SpeechWare USB 3-in-1 TableMike microphone was used to monitor the subject voice. The microphone was positioned roughly at 15 cm from the subject.

Software requirements

Dragon NaturallySpeaking and Windows Speech Recognition were used for voice interaction.

Using Zooming to facilitate Gaze Selection

To circumvent the accuracy limitations of gaze estimation VoxVisio implements an automatic zoom function to facilitate acquisition of small targets. The zooming mechanism is automatically triggered by the software around the vicinity of the target with which the subject wishes to interact after the utterance of a speech command. Throughout the zooming process, the user can steer by gaze the direction of the zoom for fine grained control of target acquisition. The process finalizes after the zoom is completed when the system generates the speech pre-selected command in the final destination of the zoomed in area.

VoxVisio software

VoxVisio is our multimodal system that enables users to interact with a computer through a mixture of voice and gaze interaction. For example, looking at a folder and saying "open" will open the gaze specified folder. This mode of Windows control makes it possible for the user to control a standard Windows desktop operating system with a two step selection

method (gazing at a target and speech commands for triggering target the specific behavior) which reduces the risks of unwanted clicks, a pervasive issue in gaze only interaction. The software provides two states for computer interaction, these states being command mode and dictation mode. The user is able to switch between the two states by focusing on their associated hotspots that are displayed on the side of the screen, see Figure 2a.

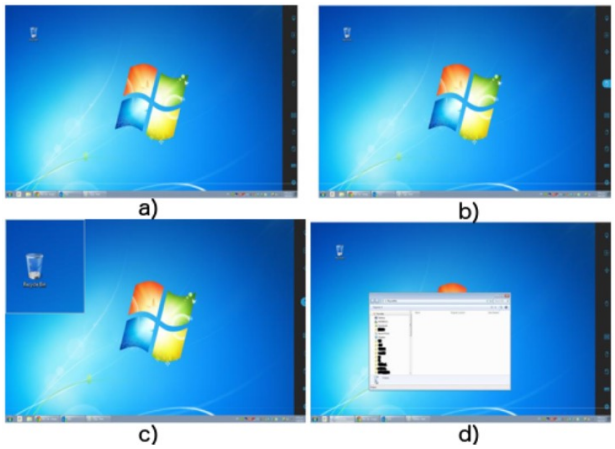

Figure 1: Tobii Gaze Selection. In Figure 1a the user can inspect via gaze any part of the screen without risk of accidentally triggering any command. In Figure 1b the user has, via gaze, activated the left click action command on the right task bar. In Figure 1c the user gazes at its desired target location, in this example the recycle bin. After a dwell-time activation threshold has been exceeded, a zooming mechanism superimposed on the screen is triggered. This zooming mechanism allows the user to dynamically steer gaze towards the target. Finally, in Subfigure 1d the action has been completed and the desired target icon has been opened.

In command mode the user is able perform common computer actions such as clicking, copying, deleting, and scrolling. To perform these actions, the system takes advantage of keyboard shortcuts, so when the user says a command, the system simply fires off the key and mouse events associated with the command at the gazed location. For commands that need fine grained mouse targeting the system first zooms roughly where the user wants to interact, then, after a short delay fires the click event where they are looking in the interface superimposed zoom window, see Figure 2a. This gives the user fine grained target acquisition accuracy. In dictation mode, Figure 2c, the user is able to dictate text using voice recognition and has access to a small set of commands for editing and grammar. The system by default uses the Windows Speech Recognition engine for dictation but, due to engine inaccurate issues, we have also give users the option of selecting Dragon Naturally Speaking for dictation. Due to the association between commands and a set of key/mouse events, the system has provided a friendly Graphical User Interface for users to customize and map specific commands to different keyboard or mouse events, see Figure 2d for an illustration of VoxVisio GUI. A schematic of the Software arquitecture can be found in Figure 3.

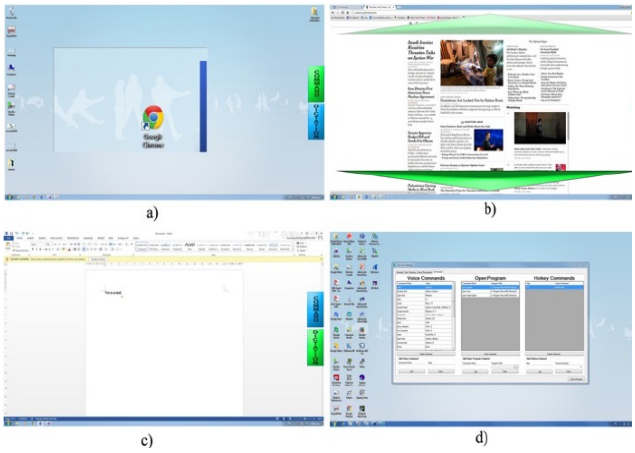

Figure 2: a) Zooming when using the "click" command b) Example of scrolling using hotspots. c) Example of dictation mode. d) Example of the GUI used for adding/editing commands.

DISCUSSION

In this work, we have presented the VoxVisio interaction paradigm which we believe has the potential to outperform traditional accessibility options, such as speech only interaction and gaze-only interaction. We shall investigated this proposition in future work by empirically comparing the VoxVisio system to

gaze only interaction, voice only interaction and the traditional baseline of mouse and keyboard-based interaction.

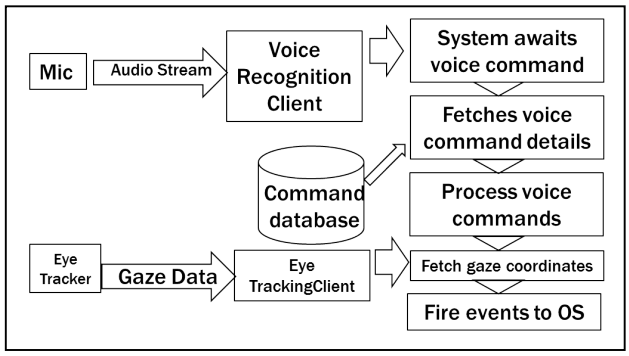

Figure 3: VoxVisio software architecture

We have made our software open source so anyone can benefit from it, assuming they can afford the cost of the eye tracker. Fortunately, these devices have massively come down in price recently and there are already companies offering an eye tracker for just $99. Another benefit of the open source nature of the project is that those with the required technical skills can push the boundaries of accessibility technologies further by enhancing VoxVisio.

REFERENCES

Bates, R., & Istance, H. O. (2003). Why are eye mice unpopular? A detailed comparison of head and eye controlled assistive technology pointing devices. *Universal Access in the Information Society*, *2*(3), 280–290. http://doi.org/10.1007/s10209-003-0053-y

Cook, A. M., & Polgar, J. M. (2014). *Assistive Technologies: Principles and Practice*. Elsevier Health Sciences.

Grauman, K., Betke, M., Lombardi, J., Gips, J., & Bradski, G. R. (2003). Communication via eye blinks and eyebrow raises: video-based human-computer interfaces. *Universal Access in the Information Society*, *2*(4), 359–373. http://doi.org/10.1007/s10209-003-0062-x

Hales, J., Rozado, D., & Mardanbegi, D. (2013). Interacting with Objects in the Environment by Gaze and Hand Gestures. Presented at the 17th European Conference on Eye Movements. ECEM 2013, Lund, Sweden.

Istance, H. O., Spinner, C., & Howarth, P. A. (1996). Providing Motor Impaired Users with Access to Standard Graphical User Interface (GUI) Software via Eye-based Interaction. In *Proceedings of the 1st European Conference on Disability, Virtual Reality and Associated Technologies (ECDVRAT '96* (pp. 109–116).

Pomplun, M., Ivanovic, N., Reingold, E. M., & Shen, J. (2001). Empirical Evaluation of a Novel Gaze-Controlled Zooming Interface. In *In*.

Rauterberg, M., Menozzi, M., & Wesson, J. (2003). *Human-computer Interaction, INTERACT '03: IFIP TC13 International Conference on Human-Computer Interaction, 1st-5th September 2003, Zurich, Switzerland*. IOS Press.

Rozado, D. (2013). Mouse and Keyboard Cursor Warping to Accelerate and Reduce the Effort of Routine HCI Input Tasks. *IEEE Transactions on Human-Machine Systems*, *43*(5), 487–493. http://doi.org/10.1109/THMS.2013.2281852

Rozado, D., Moreno, T., San Agustin, J., Rodriguez, F., & Varona, P. (2013). Controlling a Smartphone Using Gaze Gestures as the Input Mechanism. *Human–Computer Interaction*, (just-accepted).

Yuan, P., Gao, X., Allison, B., Wang, Y., Bin, G., & Gao, S. (2013). A study of the existing problems of estimating the information transfer rate in online brain–computer interfaces. *Journal of Neural Engineering*, *10*(2), 026014. http://doi.org/10.1088/1741-2560/10/2/026014