Hands-Free Web Browsing: Enriching the User Experience with Gaze and Voice Modality

Korok Sengupta
University of Koblenz-Landau
Koblenz, Germany
koroksengupta@uni-koblenz.de

Min Ke
University of Koblenz-Landau
Koblenz, Germany
minke@uni-koblenz.de

Raphael Menges
University of Koblenz-Landau
Koblenz, Germany
raphaelmenges@uni-koblenz.de

Chandan Kumar
University of Koblenz-Landau
Koblenz, Germany
kumar@uni-koblenz.de

Steffen Staab*
University of Koblenz-Landau
Koblenz, Germany
staab@uni-koblenz.de

ABSTRACT

Hands-free browsers provide an effective tool for Web interaction and accessibility, overcoming the need for conventional input sources. Current approaches to hands-free interaction are primarily categorized in either voice or gaze-based modality. In this work, we investigate how these two modalities could be integrated to provide a better hands-free experience for end-users. We demonstrate a multimodal browsing approach combining eye gaze and voice inputs for optimized interaction, and to suffice user preferences with unimodal benefits. The initial assessment with five participants indicates improved performance for the multimodal prototype in comparison to single modalities for hands-free Web browsing.

CCS CONCEPTS

• **Human-centered computing → Interaction techniques**; **Pointing**;

KEYWORDS

Hands-free interaction, multimodal interfaces, eye tracking, voice input, speech commands, Web accessibility.

ACM Reference Format:
Korok Sengupta, Min Ke, Raphael Menges, Chandan Kumar, and Steffen Staab. 2018. Hands-Free Web Browsing: Enriching the User Experience with Gaze and Voice Modality. In *ETRA '18: 2018 Symposium on Eye Tracking Research and Applications, June 14–17, 2018, Warsaw, Poland.* ACM, New York, NY, USA, Article 4, 3 pages. https://doi.org/10.1145/3204493.3208338

1 INTRODUCTION

Web browsers are one of the most important tools to enable digital information access and communication. Such a tool should not depend solely on traditional input methods like mouse, keyboard or touch that require fine motor control. Instead, the interface should allow for other, more natural, inputs that facilitate interaction abilities of the user. Voice and eye gaze input are frequent candidates for this.

Voice input has often been used synonymously with the term hands-free interaction with respect to its popularity and integration with most modern technology and services. Several existing tools like Click by Voice (CV)[1], Windows Speech Recognition (WSR) System [Brown 2008], HandsFreeChrome (HFC)[2], make Web interaction and navigation possible. However, the use and effectiveness of voice commands are subjective, depending on the accuracy of the technology, privacy and personal preferences [Karl et al. 1993].

Eye tracking, another hands-free input modality, is primarily used as the gaze-based communication medium for people lacking fine motor skills. Gaze-based mouse and keyboard emulation approaches[3] are often being used for application control, which also supports Web browser access. Moreover, GazeTheWeb [Menges et al. 2017] (GTW), Text2.0 [Biedert et al. 2010], WeyeB [Porta and Ravelli 2009] are advance developments, which adapt the Web environment for direct gaze access. While gaze as an input has the advantage of natural positioning, it suffers from accuracy and ambiguity issues (Midas Touch) [Jacob and Stellmach 2016], causing frustration and poor user experience.

Limitations of individual modality impact the performance and experience of interaction for end-users. Specially the Web environment encompasses various interactive elements, and each task requires several browsers or page specific complex interactions. Research on multimodal interactions has already shown positive indications that gaze and voice input could help overcoming limitations of accuracy and ambiguity [Mantravadi 2009; Van der Kamp and Sundstedt 2011]. Thus, it makes the investigation of the unison of gaze and voice modality for enhanced hands-free browsing experience imperative. To the best of our knowledge, there is no prior work on exploring the feasibility of gaze and voice modality for Web interaction and accessibility.

2 ANALYSIS OF GAZE AND VOICE MODALITY

To understand how modalities can be combined to realize their advantage for better Web browsing experience, we first investigated the unimodal techniques to analyze when it performs better,

*Also with University of Southampton, Web and Internet Science Research Group.

ETRA '18, June 14–17, 2018, Warsaw, Poland

ACM ISBN 978-1-4503-5706-7/18/06.
https://doi.org/10.1145/3204493.3208338

[1] https://github.com/mdbridge/click-by-voice
[2] https://www.handsfreechrome.com
[3] https://www.tobiidynavox.com/software/windows-software/windows-control

and when it is the optimum for browsing interactions. In this regard, we conducted an observational study (with 10 participants) of hands-free browsing operations. HFC and GTW were used as the representative of browsing tools for voice and eye gaze modalities respectively.

There are several implications of the input technology, specific interactions, user preferences and the environment. Certain interactions that do not involve spatial context could be performed much effectively by voice commands, e.g., Web browser specific activities such as opening a new tab, bookmarking. Gaze modality needs multiple selections (by dwelling on menu objects) to accomplish such activities, yielding more time and effort. Similarly, gaze-based URL or search string entry requires dwell-based typing of each character, resulting in a tiresome experience for the users, hence they would rather prefer text entry by speech input. However, it was noted on multiple occasions, speech input cannot interpret the user's vocabulary correctly, resulting in frustration of the user.

Gaze modality appears to provide a better experience with interactions that require spatial context, e.g., the user could intuitively scroll through the desired positions following gaze orientation. In contrast, with voice, users need numerous commands to move up, down, top, bottom, left, right. Some of the users complained that they felt more like reading commands than focusing on the actual page content.

Moreover, there are additional factors concerning the environment and fatigue hampering the user experience, for example, 7 of 10 participants could not pause a video using 'pause' voice command, probably because of background noise from the video. Similarly, the eye tracking accuracy is generally affected by head position, movement and ambient lighting.

Besides the aforementioned user preferences and issues with individual modalities, page specific interactions like clicking a hyperlink induce accuracy and ambiguity issues with both gaze and voice modalities. Due to the accuracy limitation of eye tracking, the user could not click precisely in dense page environment. Voice input is not aware of the spatial context; hence it's challenging to identify which page element the user intended to interact with. In HFC, users need to use a 'map' command that embeds numbering to each interactable element. Users need to dictate a specific number to resolve the ambiguity. The additional effort required to map elements was not considered as natural interaction by the users.

3 MULTIMODAL BROWSING FRAMEWORK

We developed a multimodal browsing framework combining gaze and voice modality to address the above-mentioned issues. The framework incorporates a two-fold integration: unified optimization where gaze and voice could complement each other, and the possibility for users to choose singular modality as per the context.

- Unified Interactions: Users look at the object they intend to interact with [Miniotas et al. 2006]. We employ the phenomena for optimized Web page interaction, where gaze provides the spatial context of attention, and voice plays an important role in confirming users' intention. For the purpose, all the elements of a Web page like links, check boxes, videos, buttons, etc. were extracted. When the user expresses the intention by issuing voice commands, we match the semantics of all extracted elements (within the gaze focus area) with the voice command, to perform the desired interaction.
- Solitary Interactions: For a better overall experience, either voice or gaze mode could be chosen for interactions with respect to performance (e.g., menu operations/text entry by speech commands, scrolling by gaze), depending on the environment (e.g., gaze in noisy conditions). Also, irrespective of the performance and environment, it could be simply chosen as a "fall-back option" when one mode does not work (e.g., not able to enter a particular word correctly by speech could be frustrating, and eye typing might be used instead). Furthermore, voice or gaze input could be demanding and cause "fatigue", where the proposed multimodal environment provides an option for the user to switch between the two hands-free options based on their mood and preferences.

Implementation. The implementation of the multimodal framework is based on GTW, an open source gaze-controlled browser encompassing CEF[4] and a custom graphical browser interface [Kumar et al. 2017]. Therefore, the look and feel of the multimodal system is centered towards the GTW eye-tracking interface. We integrate the unimodal voice commands and merge specific voice commands with gaze fixation location for optimized interaction. The voice commands are first recorded in an audio buffer which is sent to the Google Speech API[5]. Contents of the "transcript" received from the Google Speech API is then mapped to the defined command, using the Levenshtein distance as distance measurement.

Evaluation. A pilot trial with 5 participants, aged between 22 to 28 years (mean = 26.2 years; SD = 2.384) was conducted. SMI RED-n eye tracker with a sampling rate of 60Hz was used for gaze interpretation. The eye tracker was attached to a 24-inch monitor. The inbuilt microphone of the laptop that was connected to the 24-inch monitor was used to record voice commands. The participants were asked to perform browsing operation including search, navigate (for both text and video content), and bookmark pages. The average task completion time was *19.9 seconds* for the *multimodal browser*, *103.5 seconds* for *GTW*, and *27.0 seconds* for *HFC*. The initial results and feedback are promising for multimodal browsing framework. Moreover, results also highlight an improved performance for unified interactions, i.e., link selection activity performs 70% better for multimodal browser in comparison to unimodal approach.

4 CONCLUSION

In this paper, we discussed a multimodal framework and initial prototype for hands-free Web browsing, where the users could effectively perform various browsing operation using gaze and voice input. In future, we aim to enhance the features and experience, by integrating low-level semantics of Web pages, and technical advancements of offering continuous voice stream input. We also plan to conduct thorough evaluation and analysis how end-users adapt to the multimodal concept with respect to performance, feasibility, individual preferences, and environment.

[4]https://bitbucket.org/chromiumembedded/cef
[5]https://cloud.google.com/speech/

ACKNOWLEDGMENTS

This work is financially supported by MAMEM[6], EU Horizon 2020 research and innovation program: grant agreement number: 644780.

REFERENCES

Ralf Biedert, Georg Buscher, Sven Schwarz, Manuel Möller, Andreas Dengel, and Thomas Lottermann. 2010. The Text 2.0 Framework: Writing Web-based Gaze-controlled Realtime Applications Quickly and Easily. In *Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction (EGIHMI '10)*. ACM, New York, NY, USA, 114–117. https://doi.org/10.1145/2002333.2002351

Robert Brown. 2008. Exploring new speech recognition and synthesis APIs in Windows Vista. Talking Windows. *MSDN* (Mar 2008).

Rob Jacob and Sophie Stellmach. 2016. What You Look at is What You Get: Gaze-based User Interfaces. *interactions* 23, 5 (Aug. 2016), 62–65. https://doi.org/10.1145/2978577

Lewis R. Karl, Michael Pettey, and Ben Shneiderman. 1993. Speech versus mouse commands for word processing: an empirical evaluation. *International Journal of Man-Machine Studies* 39, 4 (1993), 667 – 687. https://doi.org/10.1006/imms.1993.1078

Chandan Kumar, Raphael Menges, Daniel Müller, and Steffen Staab. 2017. Chromium Based Framework to Include Gaze Interaction in Web Browser. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 219–223. https://doi.org/10.1145/3041021.3054730

Chandra Sekhar Mantravadi. 2009. *Adaptive Multimodal Integration of Speech and Gaze*. Ph.D. Dissertation. New Brunswick, NJ, USA. Advisor(s) Wilder, Joseph and Tremaine, Marilyn. AAI3387134.

Raphael Menges, Chandan Kumar, Daniel Müller, and Korok Sengupta. 2017. GazeTheWeb: A Gaze-Controlled Web Browser. In *Proceedings of the 14th Web for All Conference*.

Darius Miniotas, Oleg Špakov, Ivan Tugoy, and I. Scott MacKenzie. 2006. Speech-augmented Eye Gaze Interaction with Small Closely Spaced Targets. In *Proceedings of the 2006 Symposium on Eye Tracking Research &Amp; Applications (ETRA '06)*. ACM, New York, NY, USA, 67–72. https://doi.org/10.1145/1117309.1117345

Marco Porta and Alessia Ravelli. 2009. WeyeB, an Eye-controlled Web Browser for Hands-free Navigation. In *Proceedings of the 2Nd Conference on Human System Interactions (HSI'09)*. IEEE Press, Piscataway, NJ, USA, 207–212. http://dl.acm.org/citation.cfm?id=1689359.1689396

Jan Van der Kamp and Veronica Sundstedt. 2011. Gaze and Voice Controlled Drawing. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications (NGCA '11)*. ACM, New York, NY, USA, Article 9, 8 pages. https://doi.org/10.1145/1983302.1983311

[6]http://www.mamem.eu