Gaze & Tongue: A Subtle, Hands-Free Interaction for Head-Worn Devices

Tan Gemicioglu
tgemici@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

R. Michael Winters
mikewinters@microsoft.com
Microsoft Research
Redmond, WA, USA

Yu-Te Wang
yutewang@microsoft.com
Microsoft Research
Redmond, WA, USA

Thomas M. Gable
thomas.gable@microsoft.com
Microsoft
Seattle, WA, USA

Ann Paradiso
annpar@microsoft.com
Microsoft Research
Redmond, WA, USA

Ivan J. Tashev
ivantash@microsoft.com
Microsoft Research
Redmond, WA, USA

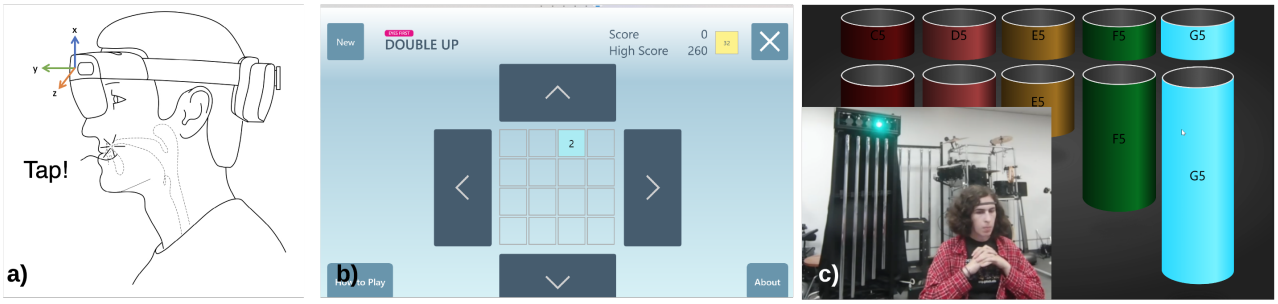

Figure 1: (a) Tongue gestures being used to control a head-worn device. (b) A hands-free game using Gaze & Tongue. (c) Tubular bell instrument controlled using Gaze & Tongue.

ABSTRACT

Gaze tracking allows hands-free and voice-free interaction with computers, and has gained more use recently in virtual and augmented reality headsets. However, it traditionally uses dwell time for selection tasks, which suffers from the Midas Touch problem. Tongue gestures are subtle, accessible and can be sensed non-intrusively using an IMU at the back of the ear, PPG and EEG. We demonstrate a novel interaction method combining gaze tracking with tongue gestures for gaze-based selection faster than dwell time and multiple selection options. We showcase its usage as a point-and-click interface in three hands-free games and a musical instrument.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Gestural input**; *Interaction techniques.*

KEYWORDS

hands-free, non-intrusive, tongue gestures, tongue interface, eye tracking, BCI

ACM Reference Format:
Tan Gemicioglu, R. Michael Winters, Yu-Te Wang, Thomas M. Gable, Ann Paradiso, and Ivan J. Tashev. 2023. Gaze & Tongue: A Subtle, Hands-Free Interaction for Head-Worn Devices. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3544549.3583930

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3583930

1 INTRODUCTION

Head-worn devices are used in a wide range of settings where users' hands are occupied or otherwise impaired in movement. These include virtual and augmented reality (VR/AR) headsets, earphones, headphones and glasses. A variety of people rely on hands-free interaction, for example, people with permanent impairments such as muscular dystrophy and stroke, and temporary impairments such as driving, cooking, and exercising. Some of the most popular means of providing hands-free interaction include speech recognition, and more recently, eye-tracking paradigms such as gaze and dwell.

However, input using speech or eyes is not ideal in certain contexts. Speech recognition is inapplicable in noisy environments, which degrade the acoustic signal quality, or in public, where the voice can be easily overheard. Eye tracking has gained more usage in recent years [6]. However, prominent input paradigms, such as

"dwell", suffer from the Midas Touch problem [9], where users can unintentionally select options while looking around the screen.

In this work, we validated a novel method of providing hands-free, voice-free input in head-worn devices, one which makes use of silent and near-invisible tongue gestures. Although most previous works required custom hardware with highly invasive sensors near or even inside the mouth, we demonstrate that tongue gestures can be recognized from sensors already present in existing head-worn devices. In particular, we found that inertial measurement units (IMUs) were the most useful among such sensors, but other sensors like photoplethysmography (PPG) and electroencephalography (EEG) can provide valuable supplementary signals. We explored how tongue gestures could be combined with eye-tracking to perform simple point-and-click actions, hands-free. We found that this multimodal solution overcame the Midas Touch problem, and delivered input significantly faster (400ms) than existing eye-only input methods (~900ms). We demonstrate this input method in various hands-free games and use it to perform an excerpt from Beethoven's 9th symphony.

2 PREVIOUS WORK

Past work on eye-tracking has used different control schemes where the dwell-based interaction can be turned on/off [5] or selection is performed using different multimodal gestures [3]. Moreover, gesture-based interaction is faster than dwell-based interactions and can have fewer error rates [2]. Many interfaces using alternative selection methods require hand movement, but some papers have proposed facial and mouth-based interaction methods as accessible, hands-free selection approaches. For example, Zhao et al. used teeth clicks to enable typing with gaze tracking [16]. Surakka et al.'s frowning and Tuisku et al.'s Wireless Face Interface use facial muscle activations as a selection method [14, 15]. Meanwhile, ClenchClick has applied teeth clenching as a selection method in augmented reality, showing the promise of such methods in head-worn devices [12].

The tongue offers a versatile alternative to such interfaces as it allows for multiple gestures and provides haptic feedback for its completion through the walls of the mouth [1]. Tongue gestures can also be performed with the mouth closed, making them nearly invisible to an observer and more appropriate for public settings than past methods. Where early work on tongue interfaces used invasive sensors within the mouth [11], recent work has shown that tongue movement can be sensed non-intrusively using sensors around the face. For example, IMUs near the ear can detect motion from the styloglossus muscle to give information on tongue and jaw movements [7, 13]. Further, the glossokinetic potential generated by the tongue allows tongue movement to be distinguished in EEG signals [8, 10]. However, past interfaces have used invasive, custom sensors and used open-mouth interaction methods such as silent speech, which are not as private as closed-mouth tongue gestures. To address these limitations, we explored whether closed-mouth tongue gestures could be recognized using sensors in commercial head-worn devices. We found that multimodal sensing could be used to recognize up to 8 closed-mouth tongue gestures. As such, our work demonstrates that closed-mouth tongue gesture recognition can be achieved affordably with off-the-shelf devices.

3 METHODOLOGY

We combined sensor data from two off-the-shelf head-worn devices and asked a sample of 16 participants across two locations to perform a series of eight tongue gestures. Using a Random Forest classifier, we found that up to 8 tongue gestures could be recognized with over 90% accuracy. Further, the recognition was fast, computing classification results in less than 10ms and only requiring 400ms time windows from the sampled sensors. Our final models used a subset of sensors including electroencephalography (EEG), inertial measurement units (IMUs) and photoplethysmography (PPG).

3.1 Hardware

Our Gaze & Tongue interface consists of a Muse 2 EEG headset and a Tobii 4C desktop eye tracker. In initial experimentation and data collection, we used the Muse 2 with the HP Reverb G2 Omnicept Edition VR headset [4], which contained an embedded Tobii eye tracker in the headset. While this was more appropriate for our goal of experimenting with head-worn devices and gave access to a broader range of sensors, we noticed that we could get greater than 90% accuracy using only sensors on the Muse 2 headset, in particular with the IMU located at the back of the ear. Some of the sensors we expected to make use of on the HP Reverb G2, such as the mouth camera, could not be used since participants were wearing face masks as COVID safety precautions. Moreover, using only one headset significantly reduced don and doff time, which is important for our live demonstration to allow more people to try on the system. For the Muse 2 headset, the IMU data was sampled at 52Hz, EEG at 256Hz, and PPG at 64Hz. These were streamed to the computer over Bluetooth via BlueMuse and recorded using the Lab Streaming Layer (LSL) protocol.

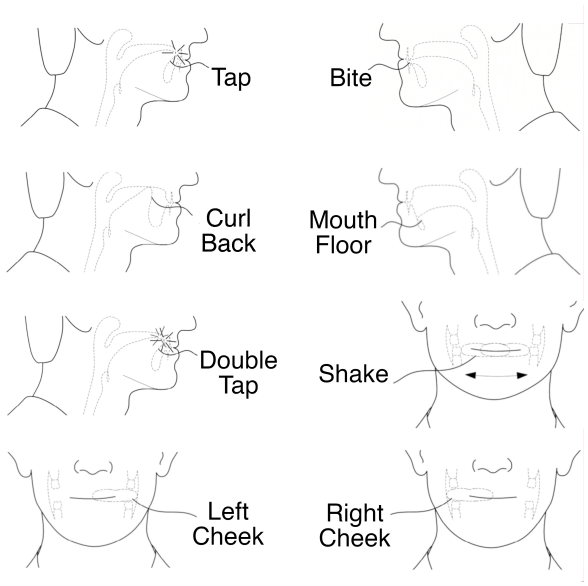

Figure 2: 8 tongue gestures developed for use with the interface. A subset is used for the demonstration.

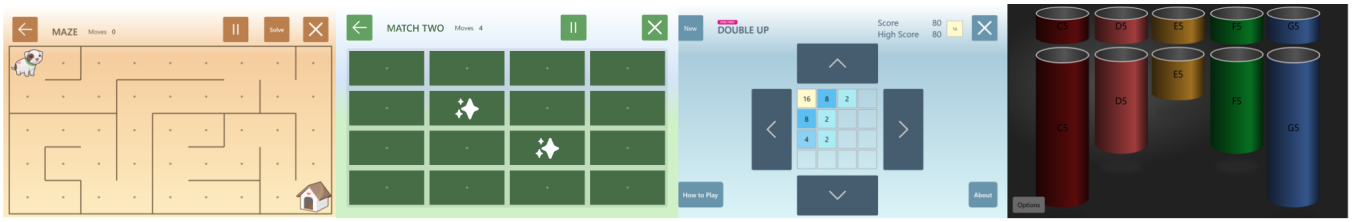

Figure 3: Applications used in the demonstration. (a) Eyes First - Maze. (b) Eyes First - Match Two. (c) Eyes First - Double Up. (d) Tubular.

3.2 Gestures & Interaction

The full list of gestures available for use in our interface is shown in 2. User-independent models for these gestures were developed with a total of 4,800 trials per gesture, across 16 participants with 300 trials per gesture each. For our Gaze & Tongue interface, we chose to use only the "Single Tap" gesture, as it was the simplest, for most of our interactions in our demonstration. We also included one hands-free game called *Eyes First - Double Up* (further described in Section 4) that used four gestures from this list as an example of multi-gesture control schemes. This provides greater versatility than gaze & dwell because it allows for similar functionality to a "right click" or "menu" without requiring multiple selections.

3.3 Real-Time Recognition

The pipeline for obtaining data from the sensors and performing real-time recognition is shown in Figure 4. Data is streamed to a Python recognizer using LSL. We use a moving window of 400 milliseconds to detect gestures, faster than most dwell-based interaction schemes. The Python script performs real-time classification by running Principal Component Analysis on IMU and PPG and Independent Component Analysis on EEG data, followed by a Random Forest model for gesture classification. The script executes a click or key press whenever a gesture is recognized, which is then registered by hands-free applications built using the Universal Windows Platform (UWP). To reduce false positives, we implemented a "cooldown" period of 0.5 seconds between each successful gesture recognition and required the gesture to be detected two frames in a row before executing the action.

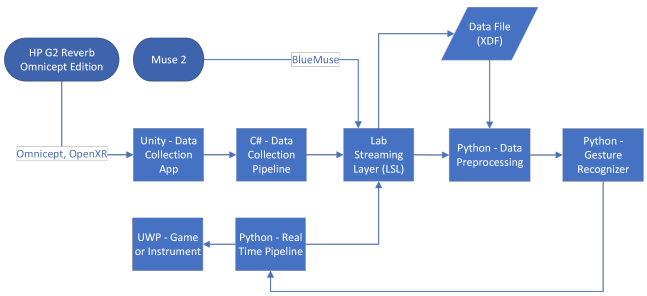

Figure 4: Pipeline for gathering data from devices and performing real-time recognition.

4 POINT-AND-CLICK INTERACTIONS WITH GAZE & TONGUE

We demonstrate our Gaze & Tongue interface in three hands-free games and a musical instrument, the *Tubular Bells*. These applications have built-in eye tracking functionality and are available publicly on the Microsoft Store except for the musical instrument. The applications are shown in Figure 3. For our live demonstration, we plan to alternate between these applications to show different use scenarios of the system. However, we plan to use the musical instrument for most of the live demonstration as we believe it will be the most enjoyable application for observers and participants.

The first game, *Eyes First - Maze* is a simple maze that the player navigates by pointing at a tile with gaze and clicking by tapping their teeth with their tongue, using the "Single Tap" gesture. The second game, *Eyes First - Match Two* is a matching game where players gaze at the card they would like to select and perform the "Single Tap" gesture with their tongue to select it. This game includes greater distance between selections, with the potential for false positives while moving between cards. The third game, *Eyes First - Double Up* is similar to the popular game 2048, where blocks are combined together to increase the player's score. The game controls are different than the first two, as it demonstrates what a tongue-only interaction might look like by mapping four gestures onto the different swipe directions. We chose to use a "Single Tap", "Left Cheek", "Right Cheek", "Mouth Floor" gesture configuration to spatially map the directions to the sides of the mouth.

The final application, *Tubular* is a musical instrument with a single chord, and users can play the instrument via "Single Tap" gestures while gazing at the key they would like to press. Of note is the fact that we only need a single target for each key, whereas a dwell-based system would need two targets for repeated presses. We chose to include a musical instrument in our demonstration as it shows the temporal precision of our system. While the sounds of the instrument are computer-generated by default, the app also has the functionality to connect to a physical instrument called the Galactic Bell Star, which we include in our video. However, the physical instrument would be very expensive and challenging to transport to the conference venue due to its size. As such, it will not be included in the live, in-person demonstration. For the musical piece, we chose to use an excerpt from Beethoven's 9th Symphony, Ode to Joy, as it is a well-known piece of music that is recognizable, does not have chords, and can be played easily within a single octave, even by beginners.

5 CONCLUSION

We demonstrated Gaze & Tongue, a novel interaction method combining gaze tracking with tongue gestures. We proposed a nearly invisible, hands-free, voice-free gesture modality that can be integrated into head-worn devices alongside gaze tracking systems to solve the Midas Touch problem. We built an interface allowing point-and-click functionality using IMU, PPG and EEG, non-intrusive sensors that are already present in head-worn devices. Finally, we showed the application of the interface to three hands-free games and a musical instrument. Future work in this area would increase the ecological validity of Gaze & Tongue by testing it in more realistic scenarios such as when participants are driving, and collecting more resting-state data to handle non-gesture movement with greater accuracy. We hope that Gaze & Tongue will show new paths towards hands-free interaction methods that do not depend on voice or dwell, allowing hands-free interaction in a wider range of environments.

REFERENCES

[1] Victor Chen, Xuhai Xu, Richard Li, Yuanchun Shi, Shwetak Patel, and Yuntao Wang. 2021. Understanding the Design Space of Mouth Microgestures. In *Designing Interactive Systems Conference 2021 (DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1068–1081. https://doi.org/10.1145/3461778.3462004

[2] Morten Lund Dybdal, Javier San Agustin, and John Paulin Hansen. 2012. Gaze input for mobile devices by dwell and gestures. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 225–228. https://doi.org/10.1145/2168556.2168601

[3] Carlos Elmadjian and Carlos H Morimoto. 2021. GazeBar: Exploiting the Midas Touch in Gaze Interaction. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 248, 7 pages. https://doi.org/10.1145/3411763.3451703

[4] Tan Gemicioglu, Mike Winters, Yu-Te Wang, and Ivan Tashev. 2022. Tongue Gestures for Hands-Free Interaction in Head Worn Displays. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Cambridge, United Kingdom) *(UbiComp/ISWC '22 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 3 pages. https://doi.org/10.1145/3544793.3560363

[5] Howell Istance, Richard Bates, Aulikki Hyrskykari, and Stephen Vickers. 2008. Snap Clutch, a Moded Approach to Solving the Midas Touch Problem. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (Savannah, Georgia) *(ETRA '08)*. Association for Computing Machinery, New York, NY, USA, 221–228. https://doi.org/10.1145/1344471.1344523

[6] Robert J.K. Jacob and Keith S. Karn. 2003. Commentary on Section 4 - Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. In *The Mind's Eye*, J. Hyönä, R. Radach, and H. Deubel (Eds.). North-Holland, Amsterdam, 573–605. https://doi.org/10.1016/B978-044451020-4/50031-1

[7] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: "Hearing" Your Silent Speech Commands In Ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (July 2022), 57:1–57:28. https://doi.org/10.1145/3534613

[8] Rasmus Leck Kæseler, Lotte N. S. Andreasen Struijk, and Mads Jochumsen. 2020. Detection and classification of tongue movements from single-trial EEG. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. 376–379. https://doi.org/10.1109/BIBE50027.2020.00068 ISSN: 2471-7819.

[9] Päivi Majaranta and Andreas Bulling. 2014. *Eye Tracking and Eye-Based Human–Computer Interaction*. Springer London, London, 39–65. https://doi.org/10.1007/978-1-4471-6392-3_3

[10] Yunjun Nam, Bonkon Koo, Andrzej Cichocki, and Seungjin Choi. 2016. Glossokinetic Potentials for a Tongue-Machine Interface: How Can We Trace Tongue Movements with Electrodes? *IEEE Systems, Man, and Cybernetics Magazine* 2, 1 (Jan. 2016), 6–13. https://doi.org/10.1109/MSMC.2015.2490674 Conference Name: IEEE Systems, Man, and Cybernetics Magazine.

[11] T. Scott Saponas, Daniel Kelly, Babak A. Parviz, and Desney S. Tan. 2009. Optically sensing tongue gestures for computer input. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology (UIST '09)*. Association for Computing Machinery, New York, NY, USA, 177–180. https://doi.org/10.1145/1622176.1622209

[12] Xiyuan Shen, Yukang Yan, Chun Yu, and Yuanchun Shi. 2022. ClenchClick: Hands-Free Target Selection Method Leveraging Teeth-Clench for Augmented Reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (Sept. 2022), 139:1–139:26. https://doi.org/10.1145/3550327

[13] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (Sept. 2022), 140:1–140:26. https://doi.org/10.1145/3550281

[14] Veikko Surakka, Marko Illi, and Poika Isokoski. 2004. Gazing and frowning as a new human–computer interaction technique. *ACM Transactions on Applied Perception* 1, 1 (July 2004), 40–56. https://doi.org/10.1145/1008722.1008726

[15] Outi Tuisku, Veikko Surakka, Toni Vanhala, Ville Rantanen, and Jukka Lekkala. 2012. Wireless Face Interface: Using voluntary gaze direction and facial muscle activations for human–computer interaction. *Interacting with Computers* 24, 1 (Jan. 2012), 1–9. https://doi.org/10.1016/j.intcom.2011.10.002

[16] Xiaoyu (Amy) Zhao, Elias D. Guestrin, Dimitry Sayenko, Tyler Simpson, Michel Gauthier, and Milos R. Popovic. 2012. Typing with eye-gaze and tooth-clicks. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 341–344. https://doi.org/10.1145/2168556.2168632