TongueTap: Multimodal Tongue Gesture Recognition with 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

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

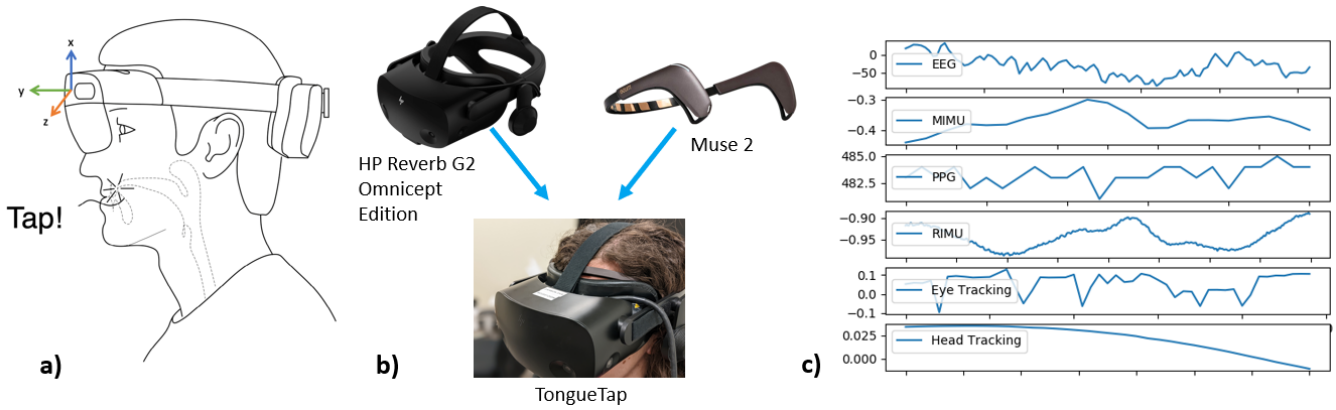

Figure 1: (a) Tongue gestures being used to control a head-worn device. (b) The TongueTap hardware setup, combining two off-the-shelf devices. (c) Data streams from six sensing modalities during a tongue gesture.

ABSTRACT

Mouth-based interfaces are a promising new approach enabling silent, hands-free and eyes-free interaction with wearable devices. However, interfaces sensing mouth movements are traditionally custom-designed and placed near or within the mouth. TongueTap synchronizes multimodal EEG, PPG, IMU, eye tracking and head tracking data from two commercial headsets to facilitate tongue gesture recognition using only off-the-shelf devices on the upper face. We classified eight closed-mouth tongue gestures with 94% accuracy, offering an invisible and inaudible method for discreet control of head-worn devices. Moreover, we found that the IMU alone differentiates eight gestures with 80% accuracy and a subset of four gestures with 92% accuracy. We built a dataset of 48,000 gesture trials across 16 participants, allowing TongueTap to perform user-independent classification. Our findings suggest tongue gestures can be a viable interaction technique for VR/AR headsets and earables without requiring novel hardware.

ICMI '23, October 9–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10.
https://doi.org/10.1145/3577190.3614120

CCS CONCEPTS

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

KEYWORDS

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

ACM Reference Format:
Tan Gemicioglu, R. Michael Winters, Yu-Te Wang, Thomas M. Gable, and Ivan J. Tashev. 2023. TongueTap: Multimodal Tongue Gesture Recognition with Head-Worn Devices. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614120

1 INTRODUCTION

Head-worn devices are increasingly ubiquitous in our lives due to the growing usage of headphones and virtual or augmented reality (VR/AR) headsets. With the rising prevalence of such devices, new interaction methods have become necessary to control the devices without requiring an external controller. These interactions commonly rely on the hands, such as pressing a button on the headphones or hand-based gesture control for augmented reality headsets. Hands-free interaction methods such as speech recognition and eye tracking provide an alternative for use cases where the user's hands may be permanently or situationally impaired. A

wide range of neuromotor disorders including Amyotrophic Lateral Sclerosis (ALS), muscular dystrophy and stroke greatly reduce the ability to move the hand voluntarily. Meanwhile, head-worn devices are used in settings such as warehouses [50], manufacturing [8] and surgeries [21] where users' hands are occupied and cannot be used for interactions.

However, speech recognition, the most common hands-free interaction method, is unusable when the environment is noisy or privacy is necessary. Gaze tracking requires continuous attention to sustain the interaction, making it difficult to control and distracting from other tasks a user may be performing. Gaze and dwell, the most common approach for gaze-based interaction is slow and has a high error rate, especially for novices [14, 39]. As a result, both speech recognition and gaze tracking are inaccessible to a wide range of users, particularly when their interactions need to be discreet and ephemeral.

Mouth-based interaction methods are hands-free, voice-free and eyes-free, offering a deeply enabling approach to interacting with head-worn devices. Past work on mouth-based interaction methods have involved custom hardware that is around the jaw [44] and neck [54] or within the mouth [16, 53]. For mouth-based interaction to be used in everyday devices, both the device and the interaction must be discreet, necessitating sensors that can be embedded in existing form factors. While recent studies have investigated sensors around the ear [5] and eyes [66], these studies have been focused on a single sensing modality with custom hardware, limiting the reproducibility and accessibility of research on mouth-based interaction. Moreover, due to the emphasis on silent speech commands [12] for these interfaces, there has been a bias towards modalities that can capture multi-organ movement from the lips to the larynx due to vocal articulation. As a result, there is a gap in multimodality and a lack of mouth interactions that are invisible for daily use.

We created a tongue gesture interface (TongueTap), by combining sensors in two commercial off-the-shelf headsets. Using this interface, we demonstrate that even sensors far from the mouth can recognize tongue movement. Compared to silent speech interfaces, interfaces using tongue gestures minimally engage the lips and jaws, and can be performed with the mouth closed, generating limited visual movement. Closed-mouth tongue gestures allow privacy during ephemeral interactions such as increasing volume in earables or closing a notification in augmented reality (AR). The availability of the two devices we used in this study may help make it easier for researchers to reproduce and experiment with their own mouth-based interaction methods using the same devices. We evaluate the performance of eight different gestures and six sensing modalities via data gathered simultaneously for comparability and make the dataset publicly available.

We evaluated TongueTap in a series of offline tests comparing accuracy across different gestures, sensing modalities, data amounts and moving window sizes across sixteen participants spread over two study locations. We compared our eight selected gestures to two controls, blinking and sticking out the tongue, as a comparison for gaze and facial interaction researchers. We developed a pipeline for real-time user-independent classification of tongue gestures, demonstrating it in different desktop applications. We also collected informal qualitative feedback and NASA Task Load Index (TLX)[22] questionnaires for each gesture.

The key contributions of this paper are:

(1) **Tongue Gesture Recognition** using commercial head-worn devices, increasing the accessibility and reproducibility of research on mouth-based interaction methods. To the best of our knowledge, this is the first tongue interface designed for off-the-shelf use. To facilitate such use, we've made our data open-access at https://zenodo.org/record/8247217.
(2) **Gesture Recognition Experiments** using 8 closed-mouth tongue gestures and two baseline conditions. We report our recognition accuracy on user-dependent and independent models, and present our findings for the ideal window sizes, gesture subsets, effects of pre-training, and a NASA-TLX.
(3) **Sensing Modality Experiments** that reveal the most descriptive sensors and sensor groups (Table 1). Our findings– such as that 80% of the accuracy was due to the Inertial Measurement Units (IMUs)—are useful for the design of future head-worn tongue gesture platforms.

2 RELATED WORK

Wearable hands-free interaction approaches have diversified significantly over the years with advanced speech recognition and techniques such as eye tracking [3], facial gestures [41, 60] and brain-computer interfaces (BCIs) [63]. For developing TongueTap, we primarily drew from past research in gaze and teeth interactions, facial muscle sensing in earables, and the glossokinetic potential, an electrophysiological motion artifact by the tongue.

2.1 Hands-free Interaction

Speech recognition and voice commands have an extensive history in facilitating hands-free interaction with head-worn devices [15] and have improved alongside speech recognition technologies. The limitations of speech recognition in different settings has led to a need for hands-free and voice-free interaction approaches.

Eye tracking emerged quickly as one such approach [3]. Gaze and dwell, an interaction method relying on fixating the eye on a single point, has remained the most common method of gaze-based interaction [62]. While expert users of the method can achieve up to 300ms with adjustable dwell times [39], dwell-based gaze interactions have suffered from high error rates and cognitive load with limited speed [14]. Eye gestures have been used at up to 250ms in head-worn displays [7, 13]. However, occupying the eyes with gestures is often undesirable as they draw too much of the user's attention.

BCIs have attempted to control devices without requiring movement [63]. Steady-state visually evoked potentials (SSVEP), showing rhythmic visual stimuli to the user, has been used with VR headsets to classify visual targets [30, 40]. While useful for paralyzed users, SSVEP has the same problems with eye tracking due to constant visual attention. BCIs have been effective with movement, and Bleichner et al. have shown attempted mouth movements to be decodable even with paralyzed users, providing support for the viability of mouth-based interactions [6].

2.2 Mouth-based Interaction

The mouth has been a target of physiological sensing for various research aims. Human activity recognition researchers have focused

Device	Modality	Location	Frequency	Details
Muse 2	IMU	Behind left ear	52Hz	6-axis, 3 from accelerometer and 3 from gyroscope
	EEG	2 across forehead, 2 behind ears	256Hz	5-axis: 2 temporal, 2 frontal electrodes and 1 amplified auxiliary channel
	PPG	Forehead	64Hz	3-axis, 2 IR and 1 red LED
HP Reverb G2 Omnicept Edition	IMU	Forehead, middle of face gasket	998Hz	6-axis, 3 from accelerometer and 3 from gyroscope, HP reports 512Hz instead of 998Hz
	PPG*	Forehead, middle of face gasket	0.2Hz	No direct access, reported as Heart Rate and Heart Rate Variability
	Eye Tracking	Between eyes in VR headset	120Hz	21-axis gaze tracking and pupillometry for both eyes
	Head Tracking	Calculated from cameras around headset	54Hz	19-axis linear and angular position/velocity/acceleration accessed through OpenXR
	Mouth Camera*	In front of nose, bottom of headset	90Hz	400x400 grayscale pixels from IR camera
	Cognitive Load*	Calculated from eye tracking and PPG	1Hz	1 minute after start, cognitive load and confidence [2]

Table 1: Sensors and calculated measures from the Muse 2 and Reverb G2 Omnicept Edition. Modalities marked with * were not used for classification for reasons explained in Section 3.1.

on detecting daily activities such as chewing, drinking and speaking [4]. Facial and mouth expressions have been sensed for use in virtual reality and teleconferencing [9, 35, 36].

Much interest in mouth-based sensing and interaction has focused on silent speech, an interaction method enabling speech communication when an audible signal cannot be used [12]. Silent speech interfaces have been targeted as a strategic interaction method for enabling fast, hands-free communication using sensors within and around the mouth. These have allowed communication with head-worn displays [5], interactions with voice assistants [26, 29] and text entry [28] by developing recognition models with large vocabularies for sensors around or inside the mouth.

Some silent speech interfaces have relied on non-contact approaches using lip reading [49, 58], infrared imaging [65] and acoustics [17]. While these are useful for interacting with mobile devices such as smartphones, head-worn displays and earables already have contact points where sensors can detect mouth movements, allowing greater flexibility in sensing approaches while keeping sensors invisible. The potential uses have resulted in a push towards non-intrusive silent speech for interacting with head-worn displays, via infared camera in HMDSpeller [1] and via acoustics in EchoSpeech [66]. Particularly exciting about EchoSpeech is the discreet form factor of the sensors, showing that they could be integrated into future head-worn displays without changing device shape.

After a silent speech interface detecting ear canal deformation by Sahni et al., there's been interest in making silent speech interactions for earables [52]. EarCommand achieved 32-word silent speech recognition with earphones and MuteIt characterized silent speech recognition using the jaw as a secondary articulator [23, 27, 56]. Roddiger et al. note that mouth-based interactions with earables have successfully detected gestures from the jaw, teeth and tongue with surprising accuracy [51]. Such earables have made use of muscles connecting the muscles around the mouth, including the tongue, to the styloid process near the ear. The styloglossal muscle has made tongue sensing possible through sensors in the upper face, which TongueTap also makes use of, for mouth gestures rather than silent speech.

Mouth gestures differ from silent speech commands by allowing a wider range of more ephemeral and short-term interactions for daily, quick usage. Many mouth gesture interfaces have involved teeth clicks [59, 61] and jaw clenching [27], but we attempted to minimize jaw and teeth movement as such gestures are audible and visible to an observer. Mouth gestures can be more discreet than silent speech by keeping input intraoral [16]. They can make mobile input easy for wearable devices without necessitating silent speech commands [42]. Chen et al. mapped the space of mouth gesture design in more detail, finding that users prefer short and direct gestures while avoiding natural motions like smiling [10]. Chen et al. further note that mouth gestures can provide haptic feedback for themselves through various surfaces around the mouth, making closed-mouth tongue gestures an appealing intraoral interaction method.

2.3 Tongue Interfaces

Many tongue interfaces have used intrusive methods that require a retainer or magnet inside the mouth [37, 43, 52, 53]. While this approach provides reliable signals, as demonstrated by SilentSpeller's 1164-word vocabulary, it comes at the cost of making users uncomfortable and limiting interaction duration [28].

Non-intrusive approaches have tried to replace such tongue interfaces using electromyography (EMG) signals from around the cheeks, neck and jaw, [44, 54, 64] or pressure sensors on the cheek [11]. Such interfaces still occupy the lower face, making tongue interactions very inconvenient for daily use. Instead, non-contact methods have used cameras [38, 48] and Doppler radar [20]. These methods require the tongue gestures to be detected through external movements, making them less viable for discreet, closed-mouth tongue gestures. A tongue interface that stands out from among such interfaces is TYTH, which only uses electroencephalography (EEG) and EMG sensors around the ear to detect tongue gestures [47]. TYTH uses the hypoglossal cranial nerve and the styloglossus and hyoglossus muscles, the same muscles allowing earable silent speech interfaces and a primary target for TongueTap. However, TYTH requires a custom headset and was still highly visible to observers due to the gestures chosen.

Some tongue interfaces have made use of the glossokinetic potential, an electrophysiological motion artifact caused by tongue movement that is commonly observed in EEG studies. Nam et al. have explored the glossokinetic potential for their tongue gesture interfaces controlling robots and electric wheelchairs [45, 46]. Kæseler et al. have investigated the glossokinetic potential as a movement-based brain-computer interface, achieving the discreetness of brain-computer interfaces with a much more reliable movement-related potential than typically possible [25, 34]. This was the second signal targeted by TongueTap in addition to the styloglossus muscle outlined in the previous section, although it showed an underwhelming result compared to the IMU. By including all of these modalities in a single study, we hope to provide a more comprehensive comparison of the different sensing modalities for tongue gestures.

3 DESIGN

3.1 Hardware Selection

We primarily selected hardware to include a range of sensors, with an emphasis on motion and electrophysiology based on past performance of IMUs in earables [41, 56] and EEG/EMG in tongue gestures [26, 47, 54]. While IMUs are available in some earables and in the correct position for tongue sensing, no commercial headphones contain EEG/EMG sensors at the time of writing. Instead, we sought to select a VR/AR headset capable of all such sensors, but the location of IMUs and the lack of reliable EEG or EMG sensors in commercial VR/AR headsets made it difficult. We thought the HP Reverb G2 Omnicept Edition (OE), the VR headset with the widest range of sensors among the headsets we looked at, would be sufficient for our goals as its documentation mentioned facial EMG, yet these sensors were not included in the headset. We combined the Reverb G2 OE with an EEG headset such that wearing both at the same time wouldn't be too uncomfortable for the study duration. We note that despite using a VR headset and EEG headset, we believe the most meaningful use cases of tongue gestures are for earables and AR. The headsets we selected are equivalent to what sensor placement in earables and AR headsets could be.

The hardware for TongueTap consists of an HP Reverb G2 OE VR headset [2] and a Muse 2 EEG headband [33]. The sensors contained by these devices are described in more detail in Table 1. Notably, both headsets contained IMUs and photoplethysmography (PPG) sensors. We excluded the calculated measures of the Reverb G2 OE as their frequency was too low, with the heart rate and variability at 0.2Hz and cognitive load at 1Hz. Moreover, we excluded the mouth camera, originally one of the most promising sensors, due to challenges with the Omnicept software used for data collection making it impossible to obtain the images. As we later elaborate in Section 8.1, the Muse 2 EEG headband may have limited our EEG results due to the five dry electrodes being on the forehead and noisier than gel electrodes.

The two headsets can be fitted to a user by extending, then contracting the Muse 2 on the user's forehead and repeating the same process with the Reverb G2, finalized by tightening the head strap to the top of the user's head. The combined hardware puts the Muse 2's forehead sensors slightly above the top of the Reverb G2's face gasket, as shown in Figure 1b.

Gesture Name	Description
Single Tap	Tap front upper teeth once with tongue
Double Tap	Tap front upper teeth twice in a row with tongue
Shake	Swing tongue left and right repeatedly
Left Cheek	Tap left cheek with tongue
Right Cheek	Tap right cheek with tongue
Mouth Floor	Touch bottom of mouth, behind lower teeth with tongue
Curl Back	Curl tongue up and towards the back of the palate
Bite	Gently bite on tongue with front teeth

Table 2: Eight discreet, closed mouth tongue gestures and how they are performed.

3.2 Gesture Design

In selecting gestures, we made sure that all of the gestures could be performed with the mouth closed so that there were neither auditory nor visual cues to a third-party observer. As Chen et al. have already conducted a gesture elicitation study for mouth gestures, we relied on their findings in choosing our gestures [10]. However, we deviated from their "best" gestures as we also sought to have a spatial mapping of the gestures around the mouth while ensuring they would be easy to recognize by machine learning models [18]. For example, we sought to have a gesture pointing up, which became curling the tongue above and backward, and another pointing left and right, which was performed as a tap on the left and right cheeks. The eight gestures selected are shown in Table 2. Notably, only three of the gestures require any jaw movement while others only engage the tongue. All the gestures are silent, contained within the mouth and use the teeth, cheeks and palate for haptic feedback.

We had a total of 10 gestures for our study. In addition to the eight tongue gestures described in Table 2, we selected two control gestures, "Blink" and "Stick Out" to benchmark our performance. The "Blink" serves as a point of comparison for gaze tracking and BCI researchers while helping verify signal quality and timestamping by using the high-amplitude EEG signals and eye tracking measurements generated during the gesture. Meanwhile, the "Stick Out" gesture is an open-mouth gesture where the tongue is stuck out to make usage obvious because the eight closed-mouth gestures were sometimes too discreet to be noticed by the experimenters. The "Stick Out" gesture is also comparable to lip-based gestures such as those used in LipIO [24] as the tongue and jaw motion are similar.

4 IMPLEMENTATION

4.1 Data Collection Software

The data from the Muse 2 and Reverb G2 OE devices was synchronized using the Lab Streaming Layer (LSL) [31], a system for time synchronization commonly used for multimodal brain-computer interfaces. LSL allows both real-time streaming as well as recording streamed data to an extended data file (XDF) using its own Lab Recorder software. For the Muse 2, we used BlueMuse [32], an open-source tool for streaming LSL data from Muse. For the Reverb G2 OE, we created a custom data streaming tool in the Unity game engine built on HP's Omnicept software and the C# endings for LSL. Outside the Omnicept software, the Reverb G2 also provides the

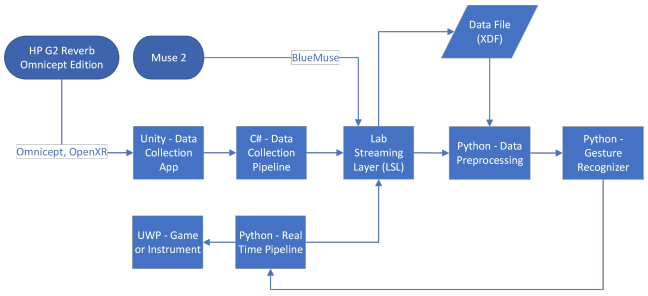

Figure 2: Data flowchart for both offline and online recognition with TongueTap.

position tracking data used in VR applications via OpenXR, which we added to our streaming tool for another measure of motion tracking.

During data collection, the user can press the "A" button on a Windows Mixed Reality controller to start a gesture and release it to stop the gesture, continuing to the next one. As gestures often take variable duration to complete, this allows more accurate boundaries to the gesture while also measuring the duration. If the user believes they made a mistake, they can instead press the "B" button to delete the previous gesture and redo it. The "Press", "Release" and "Delete" signals from these controller activities are also synchronized over LSL. All data is either stored in an XDF using the LSL Lab Recorder for offline recognition or streamed directly to a Python script processing moving windows from the data stream. The full data flowchart is shown in 2.

4.2 Gesture Recognition Approach

Our pre-processing pipeline used a 128Hz low-pass filter using SciPy and Independent Component Analysis (ICA) on the EEG signals while applying Principal Component Analysis (PCA) to the other sensors, each sensor separately from the others. ICA and PCA components were equal to the number of channels or axes for each sensor, for example, five components for EEG and six for IMU. The accelerometer values from the IMUs had gravity subtracted onboard the devices, so no additional pre-processing was performed for them. Then, we extracted 400ms windows from each gesture using MNE, beginning 100ms before the button press and ending 300ms after. Our gesture recognition models were not capable of handling invalid or raw time series data, so we removed chunks of the time series where any sensor was invalid, flattened the data into a single vector for every gesture and concatenated the sensors. We note that a model meant for time series may not require flattening and have better accuracy, although the varying frequencies of the different sensors make applying such a model to the data challenging.

For gesture recognition, we designed a hierarchical model as shown in Figure 3. Our final model used a Support Vector Machine (SVM) in Scikit-Learn using a radial basis function (RBF) kernel with hyperparameters C=100 and gamma=1 to do binary classification and determine whether a moving window of data contained a gesture or if it was a non-gesture. If the model decided it was a gesture, the final classification was done by a multi-class Random Forest Classifier with hyperparameters: 40 max. depth, 2 min. samples per leaf, 800 estimators.

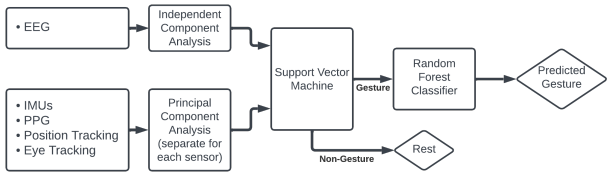

Figure 3: Architecture of gesture recognition model. Each sensor is processed separately until the SVM binary classification stage.

Age	20 to 34, average 25.6
Gender	9 Men, 5 Women, 2 Non-binary or gender diverse
Ethnicity	9 Asian, 4 White, 2 Middle Eastern, 1 Black
Occupation	12 Students, 2 Interns, 1 Researcher, 1 Teacher

Table 3: Participant demographics for both study locations.

Prior to reaching the hierarchical model, we experimented with Support Vector Machines, Random Forest Classifiers, Multi-Layer Perceptrons and Logistic Regression for the classifier. For dimensionality reduction, we tried PCA, ICA as well as Linear Discriminant Analysis (LDA). We found that the Random Forest Classifier always outperformed when doing multi-class classification yet the Support Vector Machine outperformed in binary classification, leading to the hierarchical approach for more optimally handling rest sequences. The dimensionality reduction differed for the sensing modalities, where ICA was more effective for EEG while other sensors were more successful using PCA. For tuning, as well as testing the accuracy of these models, we used 5-fold cross-validation while keeping a distinct testing set from 20% of the data. By doing so, we prevented overfitting on the testing set while tuning the models.

We attempted traditional machine learning methods instead of deep learning approaches as we were aiming for a classifier that could be executed in less than 100ms reliably. However, given the size of our dataset, deep learning methods could be plausible in recognizing tongue gestures. In our case, we didn't find it particularly necessary as we were already able to achieve a high enough accuracy in multi-class classification without leveraging deep neural networks.

5 DATA COLLECTION

The goal of our study was to create a large dataset of tongue gestures for evaluating tongue gesture recognition with sensors in off-the-shelf devices. Our study procedure was reviewed and approved by the Ethics Review Board at Microsoft prior to recruitment.

5.1 Participants

Participants were recruited at two locations (Redmond, WA, USA and Atlanta, GA, USA) through fliers around campus with a QR code, a mailing list for participants of past studies, and channels on Microsoft Teams. Participants were required to be 18-69 years in age, fluent in English and have normal vision, motor and cognitive abilities to be able to follow instructions and use the VR headset

safely. After the study, participants were compensated $50 in the form of a gift card of their choice. The demographics for the 16 participants are shown in Table 3. Participants also had a diverse range of hair length, style and texture including braided and curly hair, ensuring signals could be obtained even for users for whom BCIs traditionally fail to work.

5.2 Tasks and Procedure

When participants arrived at the study, we described the procedures and obtained informed consent. After participants were introduced to the study, we fitted the Muse 2 and then the Reverb G2 onto the participant's head and verified that they were able to see the Unity experimental interface on their display. We confirmed EEG contact quality by ensuring all electrodes had a standard deviation below 20 microvolts and waited 1 minute for all the sensors and calculated measures to stabilize before starting data collection.

The participants were then asked to do a practice round where they performed each gesture 5 times. The practice round served to help verify the signal quality, familiarize the participants with the press-and-release approach to recording and ensure that participants were doing the gestures correctly. As the gesture descriptions weren't very clear and difficult to demonstrate, this step served an important role in normalizing gesture movements across participants.

Afterward, participants started the main study for collecting the full dataset. The study consisted of 60 self-paced trials, separated into six batches of our 10 gestures. Participants performed the study fully in VR using the visual display in the Reverb G2, shown in Figure 4. At the start of a trial, participants were prompted which gesture they were to perform. During a trial, participants performed that gesture repeatedly, marking the start and end point of each gesture using the "A" button on the Windows Mixed Reality controller (i.e. button-down, button-up). Participants repeated the gesture 50 times in each trial while a visual counter incremented with each button press. Once they reached 50, the trial would end. This created a total of 3000 training examples per participant. In between batches, participants received a 10 second mandatory rest period to recoup attention. They were allowed to make other movements during rests, and we handled this "non-gesture" data as a null sequence where normal mouth and head motions could occur. Due to the long duration of the study, participants could also take an optional break of up to two minutes after every 15 trials.

At the end of the study, participants filled out a basic demographic survey and gave qualitative feedback on their experience with the interface. Additionally, the eight participants in the Atlanta site completed a NASA-TLX questionnaire for each of the 10 gestures. This was not completed at the other site due to being a later addition to the study protocol. The study took approximately 1.5 hours in total.

6 RESULTS

After data was collected from all participants and the models were optimized as described in Section 4.2, we performed a series of offline experiments for gesture classification. For the below experiments, unless otherwise specified, we used an 80/20 train-test split to build a user-dependent model with the eight gestures and rest

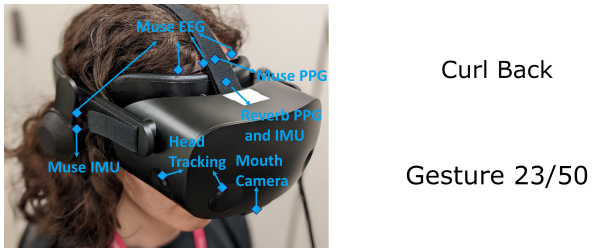

Figure 4: A participant wearing our experimental interface while performing data collection in VR, with sensor positions shown over it. Eye tracking is within gasket and omitted. Participants marked gestures using a VR controller, and received visual feedback indicating their position in the study.

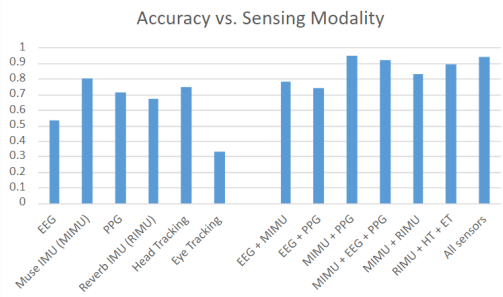

Figure 5: Sensing modality experiment. Classification across 8 gestures, mean of 16 participants.

condition using the hierarchical random forest and support vector machine model.

6.1 Classification and Sensing Modality

The result of most interest to us from the study was which sensors were most effective at classifying tongue gestures. While some of our sensors already contained multiple modalities, such as the IMUs including an accelerometer and gyroscope, we treat each stream as its own modality for the purpose of this comparison as they can be packaged together. Initially, we compared each sensing modality independently, but we observed that multimodal combinations were able to achieve a higher accuracy than a single modality alone. In particular, the most effective method was to combine the IMU on the Muse EEG headset with the PPG. The results for each modality and multimodal configuration is shown in Figure 5.

To our surprise, EEG was not particularly effective, although this may have been due to the location of the sensors being too close to the eyes, which produce a much stronger artifact. The IMU on the Muse turned out to be our most effective sensor, achieving 80% alone. Multimodal combinations including the Muse IMU were even more efficient, with a combination with the PPG sensor achieving 94% accuracy. While we have not observed this in prior literature as the PPG has never been used, we suspect this may be due to a greater blood flow to the entire face during tongue movement. We also found promising results when using the head tracking of the VR headset, although the head tracking may be less effective in a more ecologically valid configuration.

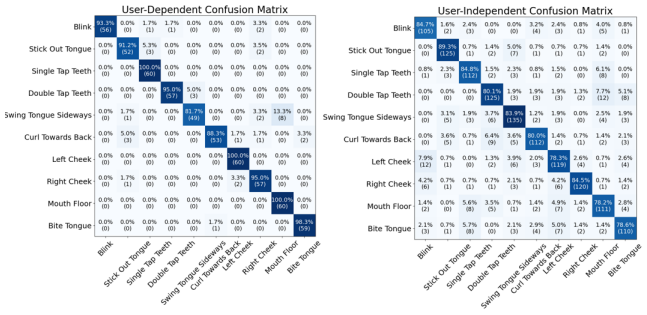

Figure 6: Confusion matrices for user-dependent and user-independent classification with all gestures and controls.

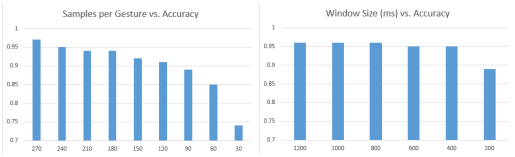

Figure 7: Samples per gesture and window size experiments. Classification across 8 gestures, mean of 16 participants.

6.2 Gesture Classification

For classifying between the gestures, we created confusion matrices for both user-dependent and user-independent classification In this case, we decided to include the "Blink" and "Stick Out" control gestures as we were curious if gestures would be confused with other movements of the face. For user-independent classification, we used a leave-one-user-out cross-validation for testing instead of a 80/20 test split averaged across users. We chose this approach as we sought to include no data from the participant being tested in the training dataset. As shown in Figure 6, the "Shake" gesture where the tongue is swung sideways was the gesture with the most error in the user-dependent model, being confused for the "Mouth Floor" gesture. The user-independent model had the classification error far more distributed, although the overall accuracy decreased to 80%.

In addition to classifying between gestures, we evaluated the amount of data per participant and window size necessary to get reliable accuracy. As shown in Figure 7, we found that recognition accuracy decayed rapidly after a dataset size of 180 samples per gesture and window size of 400 milliseconds. While this may be due to the hyperparameters chosen, the inability to reduce the dataset size suggests data augmentation methods may be necessary for achieving generalizable tongue gestures without collecting even larger datasets.

6.3 Gesture Usability

Quantitative metrics on the usability of gestures was collected using a NASA-TLX questionnaire, reported in Figure 8. Some of the gestures, such as curling the tongue back were challenging to perform, with participants pointing out they felt more tired after trials for the gesture. However, we found that the single and double tap, as

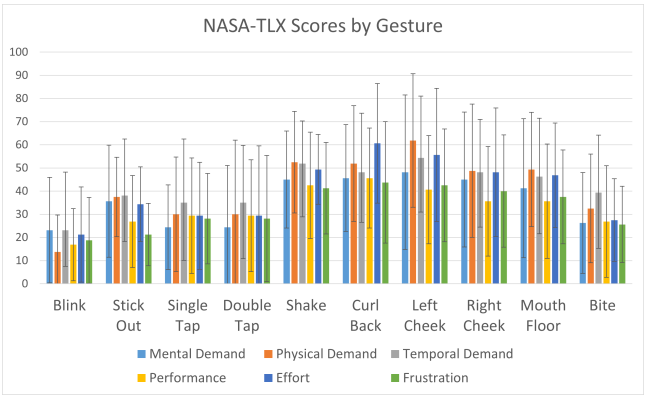

Figure 8: NASA-TLX questionnaire responses comparing the tongue gestures.

well as biting the tongue were comparable to blinking in cognitive load. Aligning with Chen et al.'s results [10], participants showed a preference for tongue gestures that were shorter in duration.

For the informal qualitative feedback, participants noted that tongue movements in the front of the mouth, such as "Bite" and "Single Tap" were more convenient, which aligns with the NASA-TLX results. Many participants struggled with interpreting what gestures meant; almost all of them asked for clarification on how "Mouth Floor" should be performed. P3 pointed out that they didn't know when to stop "Shake", which made them confused. P9 mentioned that while they felt they touched the cheek for "Left Cheek" and "Right Cheek", they weren't sure when was the right time to stop. Such an issue could be solved by real-time feedback when using the interface.

7 DISCUSSION

7.1 Integrating Tongue Gestures to Devices

Based on the sensors with the best accuracy, we can observe that the IMU behind the ear is a low-cost method of detecting tongue gestures with a position allowing it to be combined with past mouth sensing approaches such as Nguyen et al.'s ear EMG [47] and Jin et al.'s in-ear acoustics [23]. As a result, an IMU or a combination of these approaches can be used in earables or smart headphones and head-worn displays with relatively few modifications to existing hardware. Discreet, hands-free tongue gestures could replace touch-based gestures on these devices or be an alternative configuration for them. Potentially, the user could receive additional haptic feedback after performing gestures by adding additional components such as ultrasound transducers [55] or capacitive electrodes [24]. We chose not to include any custom hardware in this study as it would be against our goal of convenient replicability, although our omission of existing custom approaches makes comparison harder.

Another step critical for making tongue gestures viable for products is a reliable, user-independent classification model. While the user-independent model can already achieve above 80% accuracy, this wouldn't be sufficient for using such a classifier repeatedly to control an application. We also expect that current user-independent accuracy would decay when taken outside the lab conditions. A two-IMU approach as shown by Srivastava et al.

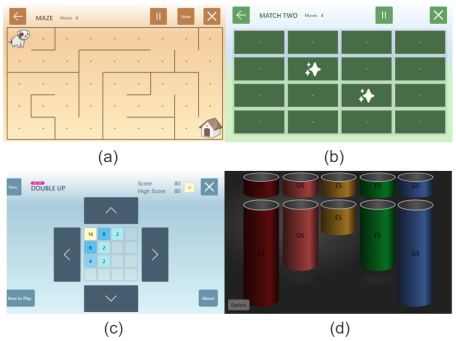

Figure 9: Four hands-free applications used for real-time tongue interactions. (a) A maze game. (b) A matching game. (c) A "double up" game similar to 2048. (d) A tubular bell musical instrument interface.

may help improve the generalization of the IMU signals [56]. An ear IMU could be combined with VR position tracking in a similar manner. An especially vital task would be obtaining a better model of non-gesture motion.

7.2 Real-time Implementation

In addition to the offline results, we implemented a real-time recognizer for tongue gestures. The offline model was pickled in Python, then unpickled for the real-time algorithm, which then captured a window of data repeatedly from the LSL stream using a ring buffer. As extra conditions for a positive gesture classification, there had to be enough data to fill the ring buffer and 400 milliseconds must have passed since the previous positive gesture classification. This helped prevent repeated activations of the gestures. When the real-time recognizer detected a gesture, it would trigger a keyboard press or mouse click, which was used to facilitate interactions with Universal Windows Platform apps. Given the success of IMU and PPG, we decided to test the real-time interface using only the Muse 2, without a VR headset. We combined the real-time tongue gesture recognizer with a Tobii desktop eye tracker, successfully controlling multiple hands-free games and an interface for controlling a music instrument [19] as shown in Figure 9. Real-time recognition typically required participants to be sitting still due to the limitations of the current data, but the interface successfully detected tongue gestures with a latency of 400 milliseconds.

7.3 Ecological Validity

While our study included multiple locations, they were both controlled lab environments where the participant was only moving to execute the gestures. For the gestures to translate to more realistic environments, a more ecologically valid study design with multiple sessions and mobility between environments is necessary. The real-time model developed in this paper will help in testing the usability of tongue gesture interactions in realistic settings.

In the future, we plan to conduct studies where participants may be performing other tasks such as walking, eating and talking to better characterize gestural and non-gestural movements. We would also like to test our tongue interface in settings emulating typical tongue gesture use cases where the sensor fit quality may decay and environmental noise would vary due to the mobility of head-worn devices. Our chunked block-based study design could be replaced by an event-based study design to better capture and evaluate gestures in daily interactions.

8 FUTURE WORK

8.1 Sensing Alternatives

The most useful sensors in our experiment were the IMU and PPG, but there are still many sensors that could be used in head-worn displays to detect tongue gestures. Acoustic approaches have been effective in sensing mouth movements in earables [23, 35] and could easily be applied in current head-worn devices. Moreover, the motion at the back of the ear captured by the Muse 2 IMU could potentially be detected in other modalities such as stretch sensors.

The sensors we used could also have different positions to make them more effective. Our EEG results fell short of our expectations based on prior work by Kæseler et al. [25, 34], which we suspect is due to the poor facial positioning of the Muse 2 EEG sensors, where there was much greater noise due to eye movements as demonstrated by the Blink control gesture. EEG and EMG sensors in the ears, around the nose piece and on the head strap may prove more useful for sensing tongue movements in head-worn displays. For earables, in-ear EEG would likely be a better option.

8.2 Tongue Interaction for Augmented Reality

We believe the most promising application for tongue interactions is in controlling AR interfaces. In contrast with VR, AR headsets are often used while interacting with other people in shared spaces, making discreet control of the interface more important. While typically not near the ear, most AR headsets contain an IMU already, making it easy to integrate our system into these devices. Tongue taps could be a suitable alternative to hand gestures like the "air tap" present in the HoloLens. AR headsets like the HoloLens 2 and Magic Leap 2 also include eye tracking as a potential sensor for facilitating interactions [57], which could be combined with tongue gestures to enable point & click interactions [19]. We plan to study this multi-organ interaction further by experimenting with its use in AR headsets and comparing it to other gaze-based interactions.

9 CONCLUSION

In this paper, we presented TongueTap, a tongue gesture interface that does not require any additional sensors beyond those available in commercial head-worn devices. We found that IMUs, PPG and motion tracking capabilities in head-worn devices can perform eight-gesture classification at greater than 70% accuracy and tried combinations of sensors that may enable tongue gesture interactions in head-worn displays with minimum cost and hardware modifications. We determined that some tongue gestures like tapping on the front teeth and biting on the tongue had cognitive demands comparable to blinking. We also found that tongue gestures can be executed and recognized at 400ms, less than gaze and dwell rates for most users. Through the sensing capabilities, gestures and interactions demonstrated by TongueTap, we put forth tongue gestures as a method for discreet, hands-free control of head-worn devices without requiring any additional hardware.

REFERENCES

[1] Kei Asano, Naoki Kimura, and Jun Rekimoto. 2023. HMDspeller: Fast and Hands-Free Text Entry System for Head Mount Displays Using Silent Spelling Recognition. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 457, 4 pages. https://doi.org/10.1145/3544549.3583910

[2] Olivier Augereau, Gabriel Brocheton, and Pedro Paulo Do Prado Neto. 2022. An Open Platform for Research about Cognitive Load in Virtual Reality. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 54–55. https://doi.org/10.1109/VRW55335.2022.00020

[3] G. Beach, C.J. Cohen, J. Braun, and G. Moody. 1998. Eye tracker system for use with head mounted displays. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218)*, Vol. 5. 4348–4352 vol.5. https://doi.org/10.1109/ICSMC.1998.727531 ISSN: 1062-922X.

[4] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sept. 2017), 1–20. https://doi.org/10.1145/3130902

[5] Abdelkareem Bedri, Himanshu Sahni, Pavleen Thukral, Thad Starner, David Byrd, Peter Presti, Gabriel Reyes, Maysam Ghovanloo, and Zehua Guo. 2015. Toward Silent-Speech Control of Consumer Wearables. *Computer* 48, 10 (Oct. 2015), 54–62. https://doi.org/10.1109/MC.2015.310 Conference Name: Computer.

[6] M. G. Bleichner, J. M. Jansma, E. Salari, Z. V. Freudenburg, M. Raemaekers, and N. F. Ramsey. 2015. Classification of mouth movements using 7 T fMRI. *Journal of Neural Engineering* 12, 6 (Nov. 2015), 066026. https://doi.org/10.1088/1741-2560/12/6/066026 Publisher: IOP Publishing.

[7] Andreas Bulling, Daniel Roggen, and Gerhard Tröster. 2008. It's in your eyes: towards context-awareness and mobile HCI using wearable EOG goggles. In *Proceedings of the 10th international conference on Ubiquitous computing (UbiComp '08)*. Association for Computing Machinery, New York, NY, USA, 84–93. https://doi.org/10.1145/1409635.1409647

[8] T.P. Caudell and D.W. Mizell. 1992. Augmented reality: an application of heads-up display technology to manual manufacturing processes. In *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, Vol. ii. 659–669 vol.2. https://doi.org/10.1109/HICSS.1992.183317

[9] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. 2020. C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-mounted Miniature Cameras. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*. Association for Computing Machinery, New York, NY, USA, 112–125. https://doi.org/10.1145/3379337.3415879

[10] 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

[11] Jingyuan Cheng, Ayano Okoso, Kai Kunze, Niels Henze, Albrecht Schmidt, Paul Lukowicz, and Koichi Kise. 2014. On the tip of my tongue: a non-invasive pressure-based tongue interface. In *Proceedings of the 5th Augmented Human International Conference (AH '14)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/2582051.2582063

[12] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg. 2010. Silent speech interfaces. *Speech Communication* 52, 4 (April 2010), 270–287. https://doi.org/10.1016/j.specom.2009.08.002

[13] Murtaza Dhuliawala, Juyoung Lee, Junichi Shimizu, Andreas Bulling, Kai Kunze, Thad Starner, and Woontack Woo. 2016. Smooth eye movement interaction using EOG glasses. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 307–311. https://doi.org/10.1145/2993148.2993181

[14] 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

[15] S. S. Fisher, M. McGreevy, J. Humphries, and W. Robinett. 1987. Virtual environment display system. In *Proceedings of the 1986 workshop on Interactive 3D graphics (I3D '86)*. Association for Computing Machinery, New York, NY, USA, 77–87. https://doi.org/10.1145/319120.319127

[16] Pablo Gallego Cascón, Denys J.C. Matthies, Sachith Muthukumarana, and Suranga Nanayakkara. 2019. ChewIt. An Intraoral Interface for Discreet Interactions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300556

[17] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (Sept. 2020), 80:1–80:27. https://doi.org/10.1145/3411830

[18] 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

[19] 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* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 456, 4 pages. https://doi.org/10.1145/3544549.3583930

[20] Mayank Goel, Chen Zhao, Ruth Vinisha, and Shwetak N. Patel. 2015. Tongue-in-Cheek: Using Wireless Signals to Enable Non-Intrusive and Flexible Facial Gestures Detection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 255–258. https://doi.org/10.1145/2702123.2702591

[21] W.E.L. Grimson, G.J. Ettinger, S.J. White, T. Lozano-Perez, W.M. Wells, and R. Kikinis. 1996. An automatic registration method for frameless stereotaxy, image guided surgery, and enhanced reality visualization. *IEEE Transactions on Medical Imaging* 15, 2 (April 1996), 129–140. https://doi.org/10/dpc987 193 citations (Crossref) [2021-11-05] Conference Name: IEEE Transactions on Medical Imaging.

[22] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[23] 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

[24] Arata Jingu, Yudai Tanaka, and Pedro Lopes. 2023. LipIO: Enabling Lips as Both Input and Output Surface. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 695, 14 pages. https://doi.org/10.1145/3544548.3580775

[25] Rasmus L. Kaeseler, Tim Warburg Johansson, Lotte N. S. Andreasen Struijk, and Mads Jochumsen. 2022. Feature and Classification Analysis for Detection and Classification of Tongue Movements From Single-Trial Pre-Movement EEG. *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society* 30 (2022), 678–687. https://doi.org/10.1109/TNSRE.2022.3157959

[26] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. Association for Computing Machinery, New York, NY, USA, 43–53. https://doi.org/10.1145/3172944.3172977

[27] Prerna Khanna, Tanmay Srivastava, Shijia Pan, Shubham Jain, and Phuc Nguyen. 2021. JawSense: Recognizing Unvoiced Sound using a Low-cost Ear-worn System. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications (HotMobile '21)*. Association for Computing Machinery, New York, NY, USA, 44–49. https://doi.org/10.1145/3446382.3448363

[28] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Zixiong Su, Alex Olwal, Jun Rekimoto, and Thad Starner. 2022. SilentSpeller: Towards mobile, hands-free, silent speech text entry using electropalatography. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3502015

[29] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–11. https://doi.org/10.1145/3290605.3300376

[30] Bonkon Koo, Hwan-Gon Lee, Yunjun Nam, and Seungjin Choi. 2015. Immersive BCI with SSVEP in VR head-mounted display. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 1103–1106. https://doi.org/10.1109/EMBC.2015.7318558 ISSN: 1558-4615.

[31] Christian Kothe et al. 2014. Lab streaming layer (LSL). https://github.com/sccn/labstreaminglayer

[32] Jason Kowaleski. 2022. BlueMuse. https://github.com/kowalej/BlueMuse original-date: 2017-08-19T05:24:36Z.

[33] Olave E. Krigolson, Chad C. Williams, Angela Norton, Cameron D. Hassall, and Francisco L. Colino. 2017. Choosing MUSE: Validation of a Low-Cost, Portable EEG System for ERP Research. *Frontiers in Neuroscience* 11 (2017). https://www.frontiersin.org/articles/10.3389/fnins.2017.00109

[34] 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.

[35] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (July 2022), 62:1–62:24. https://doi.org/10.1145/3534621

[36] Richard Li and Gabriel Reyes. 2018. Buccal: low-cost cheek sensing for inferring continuous jaw motion in mobile virtual reality. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers (ISWC '18)*. Association for Computing Machinery, New York, NY, USA, 180–183. https://doi.org/10.1145/3267242.3267265

[37] Richard Li, Jason Wu, and Thad Starner. 2019. TongueBoard: An Oral Interface for Subtle Input. In *Proceedings of the 10th Augmented Human International Conference 2019* (Reims, France) *(AH2019)*. Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3311823.3311831

[38] Li Liu, Shuo Niu, Jingjing Ren, and Jingyuan Zhang. 2012. Tongible: a non-contact tongue-based interaction technique. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility (ASSETS '12)*. Association for Computing Machinery, New York, NY, USA, 233–234. https://doi.org/10.1145/2384916.2384969

[39] Päivi Majaranta, Ulla-Kaija Ahola, and Oleg Špakov. 2009. Fast gaze typing with an adjustable dwell time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, New York, NY, USA, 357–360. https://doi.org/10.1145/1518701.1518758

[40] Ignas Martišius and Robertas Damaševičius. 2016. A Prototype SSVEP Based Real Time BCI Gaming System. *Computational Intelligence and Neuroscience* 2016 (2016), 3861425. https://doi.org/10.1155/2016/3861425

[41] Katsutoshi Masai, Kai Kunze, Daisuke Sakamoto, Yuta Sugiura, and Maki Sugimoto. 2020. Face Commands - User-Defined Facial Gestures for Smart Glasses. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 374–386. https://doi.org/10.1109/ISMAR50242.2020.00064 ISSN: 1554-7868.

[42] Denys J. C. Matthies, Bernhard A. Strecker, and Bodo Urban. 2017. EarFieldSensing: A Novel In-Ear Electric Field Sensing to Enrich Wearable Gesture Input through Facial Expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1911–1922. https://doi.org/10.1145/3025453.3025692

[43] Mostafa Mohammadi, Hendrik Knoche, Bo Bentsen, Michael Gaihede, and Lotte N. S. Andreasen Struijk. 2020. A Pilot Study on a Novel Gesture-Based Tongue Interface for Robot and Computer Control. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. 906–913. https://doi.org/10.1109/BIBE50027.2020.00154 ISSN: 2471-7819.

[44] Takuro Nakao, Yun Suen Pai, Megumi Isogai, Hideaki Kimata, and Kai Kunze. 2018. Make-a-face: a hands-free, non-intrusive device for tongue/mouth/cheek input using EMG. In *ACM SIGGRAPH 2018 Posters (SIGGRAPH '18)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3230744.3230784

[45] Yunjun Nam, Bonkon Koo, Andrzej Cichocki, and Seungjin Choi. 2014. GOM-Face: GKP, EOG, and EMG-Based Multimodal Interface With Application to Humanoid Robot Control. *IEEE Transactions on Biomedical Engineering* 61, 2 (Feb. 2014), 453–462. https://doi.org/10.1109/TBME.2013.2280900 Conference Name: IEEE Transactions on Biomedical Engineering.

[46] 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.

[47] Phuc Nguyen, Nam Bui, Anh Nguyen, Hoang Truong, Abhijit Suresh, Matt Whitlock, Duy Pham, Thang Dinh, and Tam Vu. 2018. TYTH-Typing On Your Teeth: Tongue-Teeth Localization for Human-Computer Interface. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '18)*. Association for Computing Machinery, New York, NY, USA, 269–282. https://doi.org/10.1145/3210240.3210322

[48] Shuo Niu, Li Liu, and D. Scott McCrickard. 2019. Tongue-able interfaces: Prototyping and evaluating camera based tongue gesture input system. *Smart Health* 11 (Jan. 2019), 16–28. https://doi.org/10.1016/j.smhl.2018.03.001

[49] Laxmi Pandey and Ahmed Sabbir Arif. 2021. LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3411764.3445565

[50] Rupert Reif and Willibald A. Günthner. 2009. Pick-by-vision: augmented reality supported order picking. *The Visual Computer* 25, 5 (May 2009), 461–467. https://doi.org/10.1007/s00371-009-0348-y

[51] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (Sept. 2022), 135:1–135:57. https://doi.org/10.1145/3550314

[52] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. 2014. The tongue and ear interface: a wearable system for silent speech recognition. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers (ISWC '14)*. Association for Computing Machinery, New York, NY, USA, 47–54. https://doi.org/10.1145/2634317.2634322

[53] 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

[54] Makoto Sasaki, Kohei Onishi, Dimitar Stefanov, Katsuhiro Kamata, Atsushi Nakayama, Masahiro Yoshikawa, and Goro Obinata. 2016. Tongue interface based on surface EMG signals of suprahyoid muscles. *ROBOMECH Journal* 3, 1 (April 2016), 9. https://doi.org/10.1186/s40648-016-0048-0

[55] Vivian Shen, Craig Shultz, and Chris Harrison. 2022. Mouth Haptics in VR using a Headset Ultrasound Phased Array. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3491102.3501960

[56] 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

[57] Sophie Stellmach. 2022. Eye tracking overview - Mixed Reality. https://docs.microsoft.com/en-us/windows/mixed-reality/design/eye-tracking

[58] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 581–593. https://doi.org/10.1145/3242587.3242599

[59] Wei Sun, Franklin Mingzhe Li, Benjamin Steeper, Songlin Xu, Feng Tian, and Cheng Zhang. 2021. TeethTap: Recognizing Discrete Teeth Gestures Using Motion and Acoustic Sensing on an Earpiece. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 161–169. https://doi.org/10.1145/3397481.3450645

[60] 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

[61] Tomás Vega Gálvez, Shardul Sapkota, Alexandru Dancu, and Pattie Maes. 2019. Byte.It: Discreet Teeth Gestures for Mobile Device Interaction. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312925

[62] Colin Ware and Harutune H. Mikaelian. 1986. An evaluation of an eye tracker as a device for computer input2. In *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface (CHI '87)*. Association for Computing Machinery, New York, NY, USA, 183–188. https://doi.org/10.1145/29933.275627

[63] Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. 2002. Brain-computer interfaces for communication and control. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 113, 6 (June 2002), 767–791. https://doi.org/10.1016/s1388-2457(02)00057-3

[64] Qiao Zhang, Shyamnath Gollakota, Ben Taskar, and Raj P.N. Rao. 2014. Non-intrusive tongue machine interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2555–2558. https://doi.org/10.1145/2556288.2556981

[65] Ruidong Zhang, Mingyang Chen, Benjamin Steeper, Yaxuan Li, Zihan Yan, Yizhuo Chen, Songyun Tao, Tuochao Chen, Hyunchul Lim, and Cheng Zhang. 2022. SpeeChin: A Smart Necklace for Silent Speech Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (Dec. 2022), 192:1–192:23. https://doi.org/10.1145/3494987

[66] Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. 2023. EchoSpeech: Continuous Silent Speech Recognition on Minimally-Obtrusive Eyewear Powered by Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 852, 18 pages. https://doi.org/10.1145/3544548.3580801