TeethTap: Recognizing Discrete Teeth Gestures using Motion and Acoustic Sensing on an Earpiece

Teeth gestures become an alternative input modality for different situations and accessibility purposes. In this paper, we present TeethTap, a novel eyes-free and hands-free input technique, which can recognize up to 13 discrete teeth tapping gestures. TeethTap adopts a wearable 3D printed earpiece with an IMU sensor and a contact microphone behind both ears, which works in tandem to detect jaw movement and sound data, respectively. TeethTap uses a support vector machine to classify gestures from noise by fusing acoustic and motion data, and implements K-Nearest-Neighbor (KNN) with a Dynamic Time Warping (DTW) distance measurement using motion data for gesture classification. A user study with 11 participants demonstrated that TeethTap could recognize 13 gestures with a real-time classification accuracy of 90.9% in a laboratory environment. We further uncovered the accuracy differences on different teeth gestures when having sensors on single vs. both sides. Moreover, we explored the activation gesture under real-world environments, including eating, speaking, walking and jumping. Based on our findings, we further discussed potential applications and practical challenges of integrating TeethTap into future devices.

Additional Key Words and Phrases: Teeth Gestures; Eyes-free Input; Hands-free Input; Motion Sensing; Acoustic Sensing; Earpiece

ACM Reference Format:

. 2021. TeethTap: Recognizing Discrete Teeth Gestures using Motion and Acoustic Sensing on an Earpiece. In *XX*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The vast-majority of input techniques for mobile devices demands the use of hands as an input source, which may constraints user experiences. For example, it would be inconvenient for a user to interact with a smartwatch to reject a phone call while both hands are occupied (e.g. carrying objects [23]). Therefore, providing hands-free interactions may improve the wearable interactive experiences under different situational uses.

To understand different hands-free interaction options, prior works explored eye-tracking systems [8, 9], tongue input [27], teeth input [4, 21], facial expression [20], and voice recognition [11]. Eye-tracking systems [9] usually require a mounted camera attached to glasses or a stationary device to become hands-free. Beyond having a camera facing users, more research leveraged head-worn sensors to track input gestures from the face. However, many of these works either require complex on-body sensor contacts on the face [14, 26, 36], abnormal sensor locations [21] or placing sensors in the mouth [12, 28]. To simplify the sensing requirements and hardware complexity, further research explored hands-free interaction techniques through sensors that are easily mounted on existing wearable devices, such as glasses or earbuds. For example, CanalSense leveraged barometers in the earbuds to classify face-related movements [3]. However, past works mostly explored face-related gestures through either barometers [3] or bone-conduction microphones [4]. Furthermore, many existing devices have already embedded inertial measurement unit (IMU) sensors, such as eSense [16], into earpieces. However, little research has explored the feasibility of leveraging motion sensing (e.g., IMU) combined with acoustic sensing (e.g., microphone) to recognize face-related gestures.

Manuscript submitted to ACM

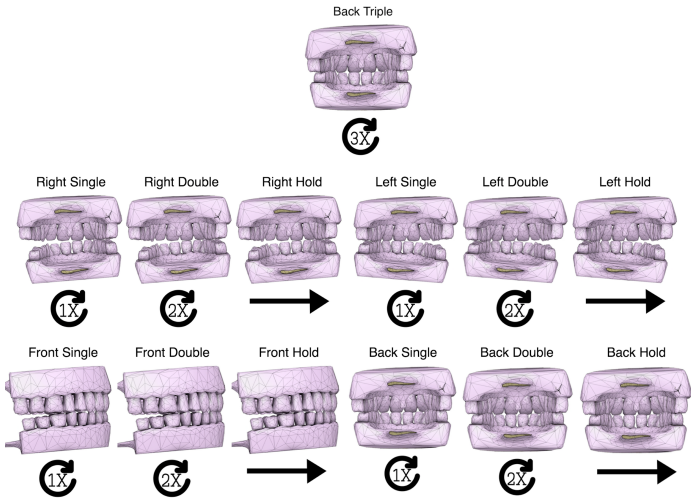

Fig. 1. Out gesture set with 13 teeth gestures.

In this paper, we present TeethTap, a minimally invasive eyes-free, hands-free input technology that can recognize up to 13 discrete teeth gestures (Fig. 1), which cover both places of contact (i.e., left side, right side, front and back) and methods of contact (i.e., single bite, double bite, or hold). To recognize these 13 teeth gestures, we built a lightweight earpiece, which secures a microphone and IMU sensor behind each ear. The earpiece was made of 18 small components which were 3D printed and then fitted together, and was adjustable to various ear sizes and head widths. To understand the feasibility of leveraging TeethTap to recognize teeth gestures, we conducted a user study with 11 participants in five sessions (i.e., one practice session, one training session, two testing sessions, and one remounting session). We then analyzed the accuracy to recognize gestures using a DTW-based K-Nearest-Neighbor(KNN) algorithm, which has been widely used to classify IMU-based data in previous literature [35].

Overall, TeethTap achieved 90.9% accuracy on average to classify 13 discrete teeth input gestures in the testing sessions. We further compared the differences in the accuracy of having sensors on both sides vs. one single side. We found that it is sufficient to only use a single side sensor to recognize 'manner' gestures, such as single-tap, double-tap, and hold. We also uncovered the accuracy differences in remounting the sensors by participants themselves and participants' subjective feedback. We further discussed the existing challenges of using TeethTap in-the-wild, the potential applications (e.g., volume control with "Hold" gestures), integrating TeethTap to other devices, and how to avoid remounting problems of TeethTap. We believe our findings shed light on future research that leverages motion and acoustic sensing on earpieces to recognize teeth gestures. Our contributions are summarized as follows:

- We explored the feasibility of leveraging motion sensing captured around the ear, and fusing motion and acoustic signals to filter noises, to recognize 13 discrete teeth gestures with an average accuracy of 90.9%.
- We uncovered the effect on different gestures of having motion sensors on both sides vs. one side and discussed the influences on recognition accuracy from remounting the devices.
- We proposed a set of design implications to apply the combination of motion and acoustic sensing on earpieces (e.g., in-the-wild scenario, design form factors, integrating to other head-worn devices).

2 RELATED WORK

Hands-free interaction techniques benefit people under different situational uses. Prior research exploring hands-free wearable input devices focuses on tracking eye movement [5, 9, 31, 37], head movement [10], jaw movement [33] and lip movement [1]. For example, Rantanen et al. [26] leveraged head-mounted capacitive and electromyography (EMG) sensors to detect different facial gestures. Similarly, Interferi [14] allowed users to wear a face sensing mask that used acoustic interferometry to track face-related gestures. However, these approaches often require aggressive instrumentation, such as cameras, magnets or headsets, to accurately distinguish between user input gestures.

To explore other hands-free interaction techniques that require minimal hardware instrumentation and complexity, prior works explored different approaches to recognize teeth gestures [4] and tongue gestures [24, 27, 32]. Researchers first explored the approaches by adding sensors inside the human's mouth, such as embedded optical sensors into orthodontic dental retainers to detect tongue gestures [28] and intraoral sensing bit to detect different tongue and teeth gestures [12]. Moreover, Li et al. [19] used sensor-embedded teeth to recognize four mouth-related activities: coughing, chewing, drinking and speaking. However, these approaches might be obtrusive to some people who do not have dental retainers or do not want to hold a sensor bit in the mouth.

To avoid placing sensors inside the mouth, past researchers further explored other approaches like placing bone-conduction microphones (e.g., [4]) on the skin to track teeth gestures or tongue gestures. For example, TeethClick [21] placed a single throat microphone that touched the cheek and picked up vibration signals from the jawbone to recognize single vs. double teeth clicks. To further make the hardware instrumentation less obtrusive, Bitey [4] recognized tooth click sounds from up to five different pairs of teeth gestures with bone-conduction microphones worn above the ears. However, Bitey tested user-specific gesture sets tailored to each participant, and the study relied solely on acoustic data, which has several limitations related to noise from speaking, chewing or outside agents.

Another approach to track tooth-clicks is to use motion sensing (e.g., IMU). Simpson et al. [29] introduced Tooth-Click Detector, which used a three-axis accelerometer on an earbud to pick up strong vibrations from tooth-clicks to control computer cursors. Zhao et al. [38] further employed the Tooth-Click Detector [29] as well as an eye-gaze tracker to type on an on-screen keyboard. Researchers had also used tooth-touch sound as an alternative mouse device for accessibility [17, 30]. However, these approaches were binary—only being able to detect whether or not there was a tooth click. Recently, many existing earpieces have already embedded IMU sensors, such as eSense [16], and have been applied for activity recognition [15, 18]. However, it is unknown whether it is feasible to detect different teeth gestures through earpieces with IMU sensors.

Previous works also explored the combination of using IMU sensors and microphones to detect eating behaviors [6, 7]. We understand that acoustic sensors are more error-prone to noise, and motion sensors are more likely to have false positives while the user is in motion. Therefore, it is important to explore how both motion and acoustic sensors can be used in tandem on earpieces to detect different teeth gestures and reduce false positives. In our work, TeethTap fuses acoustic sensing with motion sensing to accurately classify a large set of 13 universally applied teeth gestures. By combining data from two separate sensing modalities into one device, our system is able to better realize noise and recognize teeth tapping movements. Furthermore, our instrumentation is minimally-intrusive, securing both sensors discreetly behind each ear. Strategic sensor placement combined with a robust classification system makes TeethTap a viable future accessory to the ear.

3 GESTURE DESIGN

Our approach in designing teeth gestures was inspired in part by two linguistic vowel sound features: the degree of aperture (jaw openness) and tongue frontness (or backness) [25]. The degree of aperture functions as z-axis, and is relevant for gesture release detection. We applied the idea of tongue frontness to the jaw, functioning as a y-axis. Lastly, we added a final axis (x-axis) for side-side movement. The four extremes of our x-y plane can therefore be described as front, back, left, and right. This design maximized the spread of each point of contact to best avoid confusion when classifying one gesture from another.

In linguistics, there are two primary categories of articulation: the place of articulation and the manner of articulation[12]. As described above, our gestures have four places the teeth can make contact: front, back, left, and right. For each place of contact, TeethTap employs three possible "manners" of contact: single tap, double tap, and hold. "Single tap" is a quick tap and release. Naturally, "double tap" is composed of two quick single taps followed by a release. "Hold" is a tap with a delayed release. The time passed from the start of the hold gesture when the teeth first make contact, and the release of the gesture is registered as a continuous variable representing analog input. All non-hold gestures represent discrete digital inputs. We also added a gesture into our gesture set. This gesture is composed of three quick single back (regular) bites in sequence. We designed this gesture to be natural to produce yet easily recognizable for the purpose of testing under various real-world conditions such as walking, jumping, eating and speaking.

4 METHODOLOGY

4.1 Sensing Principle

By positioning our IMU sensors just behind the bottom of the ear where the jawline begins, we are able to collect gyroscope movement across three axes whenever the jaw shifts upwards, downwards or sideways. Fig. 2 illustrates this principle on the IMU's y-axis under the left ear. As the jaw extrudes leftward, it presses against the bottom part of the left IMU, causing the gyroscope to rotate upwards. The resulting rotation causes its y-axis value to increase. On each earpiece, we also placed a microphone to collect and analyze acoustic data from different teeth gestures.

Fig. 3(a) shows the left ear's IMU data during a left single gesture, clearly depicting this positive y-axis peak. Conversely, the right ear IMU depicts a negative y-axis peak, as the right jaw retracts, rotating the right IMU in the opposite direction (Fig. 3(b)). Similarly, Fig. 3(d) further shows a similar peak, this time illustrating the right ear's IMU data for a single right gesture. Again, the negative peak in Fig. 3(c) is caused by the left side of the jaw retracting and rotating the IMU downwards.

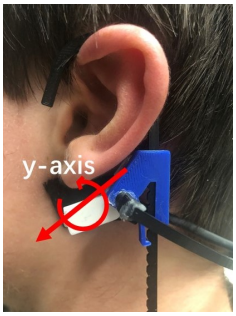

Fig. 2. Y-axis in relation to jaw movement

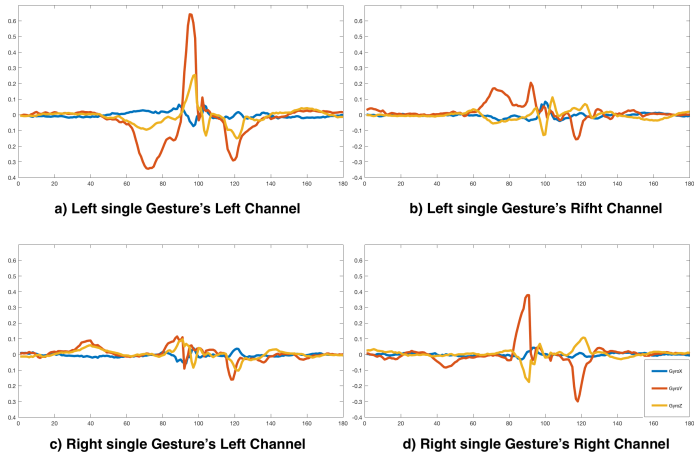

Fig. 3. Raw Gyroscope data for Left Single and Right Single Gestures

Fig. 4 illustrates the gyroscope data from four gestures: back triple, back double, back single, and back hold (top to bottom, respectively). The first three high-amplitude peaks in Figure 4(a) represent the back-triple gesture. The fourth smaller peak at the end of the window represents the gesture release. Release energy is captured when the mouth opens after performing a gesture. Back double (Fig. 4(b)) and back single (Fig. 4(c)) also end with a release peak. Notice that back hold (Fig. 4(d)), has no release peak because hold gestures delay the release, categorizing it instead as a separate sub-gesture.

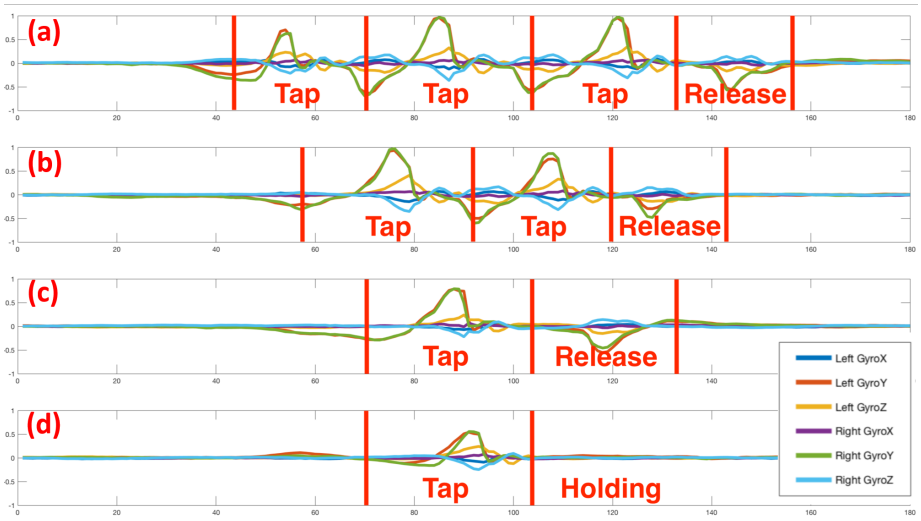

Fig. 4. Raw Gyroscope data for back triple, back double, back single, back hold Gestures

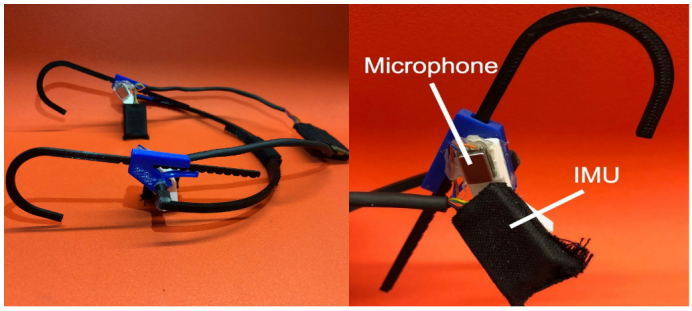

Fig. 5. 3D printed earpiece housing two microphones and two IMUs

4.2 Hardware Design

TeethTap's hardware is composed of a 3D printed earpiece housing two contact microphones and two IMUs. Our 3D printed earpiece is made from 18 small individual components assembled together to form a single unit (Fig. 5). The design is adjustable around the ears and behind the head to accommodate for various ear sizes and head widths. The natural flex of thinly printed PLA filament presses the IMU sensors against the jawline just under the ear and secures the microphones to the temporal bone behind the ear. We used two contact microphones (BU-30179-000) [22] and two inertial measurement units (IMU) (MPU-9250) [13] to capture sound and motion on the skin behind each ear. The contact microphones are connected to a customized PCB board, which amplifies and filters the acoustic signals. The filtered data from acoustic sensors and the gyroscope data from IMUs are sent to a micro-controller (HUZZAH32) [2] using its on-board 12-bit analog to digital converter (ADC) and its inter-integrated circuit (I^2C) communication, respectively. The microphone data is sampled at 8000 Hz, and the IMU data is sampled at 120 Hz. Lastly, the HUZZAH32 sends the data to a computer for processing using WiFi.

4.3 Data Processing Pipeline

To collect sensor data from the HUZZAH32 board, we created a Python program on the receiving computer. We also used the same program to analyzes the data for gesture recognition in two stages: gesture segmentation and gesture classification. Figure 6 illustrates TeethTap's data processing pipeline.

4.4 Gesture Segmentation

Our algorithm first segmented a two-second sliding window from the continuous data stream generated from the microphones and IMUs. As data flowed in and out of the queue, our sliding window shifted 20 times a second with an overlap of 95 percent. Every window, we checked if the microphone data exceeded a predetermined energy threshold,

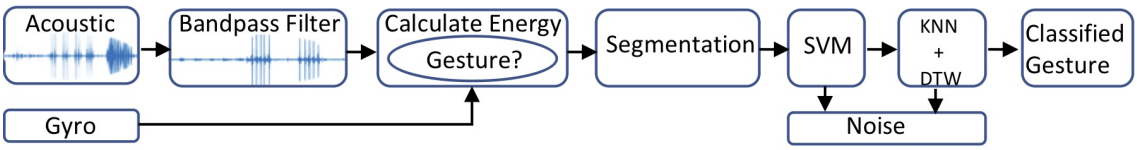

Fig. 6. Data Processing Pipeline

which indicated a gesture was possibly performed. Once our system detected a sufficient spike in audio data, we then grabbed that window's corresponding two-second gyroscope data window. Next, we checked if the gyroscope's y-axis absolute maximum value exceeded a predetermined energy threshold to understand whether a gesture was performed. At this stage, we waited until the gesture was centred within the two-second sliding window in preparation for segmentation. Because most participants finished each gesture in roughly 1.5 seconds, further segmentation was needed. To segment the data, we smoothed the absolute value of the peak(s) to find the gesture's center-point and added a 90 data point buffer on each side to form a finalized event region of 1.5 seconds (i.e., 180 data points).

4.5 Noise Detection with Acoustic Sensing

Although TeethTap's contact microphone was hardly affected by outside noise, self-generated noise such as eating, talking or walking might interfere with the system. To address this issue, we implemented an SVM model classifier with a linear kernel to train both acoustic features and IMU features in the frequency domain. To collect acoustic data for noise and gestures, we asked one researcher and two pilot participants (one female) to each perform each teeth gesture five times, and we collected noise information by asking them to talk, walk, eat food, and remain static in random order. Overall, we collected 650 gesture segments and 650 noise segments.

TeethTap extracted features from the IMU data and the microphone data for SVM classification. Seven of the eight IMU-related features were calculated across each of the three axes for both gyroscopes (six axes total). These included the number of peaks, peak values, root mean square (RMS), zero-crossing rate, standard deviation, minimum value, and maximum value. The eighth IMU-derived feature was calculated by finding the correlation between each of the left gyroscope axes with each of the right gyroscope axes. We also collected two acoustic features from the microphone data: the 30 lower bins of the Fast Fourier Transform (FFT) and 26 Mel-frequency cepstral coefficients (MFCC) [35]. This was made for a total of 64 features used to train our SVM model. We then applied the model to classify noise segments vs. gesture segments from acoustic data in TeethTap (Fig. 6).

4.6 Gesture Classification Algorithm

After segmenting the data and filtering out the noise, we classified the gestures by using K-Nearest-Neighbor (k=1) with a distance measurement of multi-dimensional Dynamic Time Warping (DTW) [34]. DTW is known for finding temporal patterns (similarities) between time-series datasets (especially with small training sets). Our first ran DTW (Dynamic Time Warping) on the data gathered during gesture segmentation with each gesture instance from training, one at a time. DTW's distance function would then output a value from every iteration. The gesture with the smallest distance value was determined as the predicted gesture.

5 USER STUDY

5.1 Participants

To evaluate the real-time gesture recognition performance and the usability of TeethTap, we conducted a user study with 11 participants with an average age of 24.3 (five female). Each participant received a $10 gift card or cash for participating in our study. The study for each participant lasted around one hour and was conducted in a laboratory environment. The study was approved by the institutional review board (IRB).

5.2 Procedure

At the beginning of the user study, we played a video that demonstrated how to perform 13 TeethTap gestures using animations of teeth constructed with AutoCAD (Fig. 1), followed by a live demonstration of the system by the researcher. Next, we helped the participant put on the device and explained the user interface (UI) of the system. The participant was asked to sit in front of a table with a monitor that displayed the testing UI. We then conducted the study in five different sessions: one practice session, one training session, two testing sessions, and one remounting session.

In each of the five sessions, participants were asked to perform each of the 13 gestures five times in a random order, which was indicated in the monitor. The first session was the practice session, which was designed to help the participant familiarize themselves with the gestures and testing system. The second session was the training session. The data collected in the training session was used to train an ML model, which was then used to provide real-time classification in later sessions. In the two testing sessions, we provided real-time classification results to the participant. The model was trained using the data collected in the training session. If the gesture was recognized as the same gesture appearing on the screen, we changed the background to green. Otherwise, the background was turned to red, and the recognized gesture's name and the picture were displayed. Whenever the system detected a holding gesture, the UI displayed a clock and asked the participants to hold for a randomly generated time interval (two to four seconds) until release. If no release gesture was detected within five seconds after the timer ended, the system timed out, counting the attempt as a recognition failure and proceeding to the next gesture in sequence.

To further understand the effect of taking TeethTap off and put it back on the same training data, we conducted a remounting session. Participants were asked to take off our prototype and put it back before this session started. Afterward, participants followed the same instructions as the testing session. In total, we collected 2860 (11*13*5*4) gesture instances in the training, testing and remounting sessions. At any point in the study, if the participant misconducted a gesture (performed a gesture than was different from the requested gesture), we asked the participant to report this to the researcher, and we removed these instances from the training and testing data. In total, 89 out of 2860 instances were removed. The real-time classification results and sensor data were saved for later analysis. After

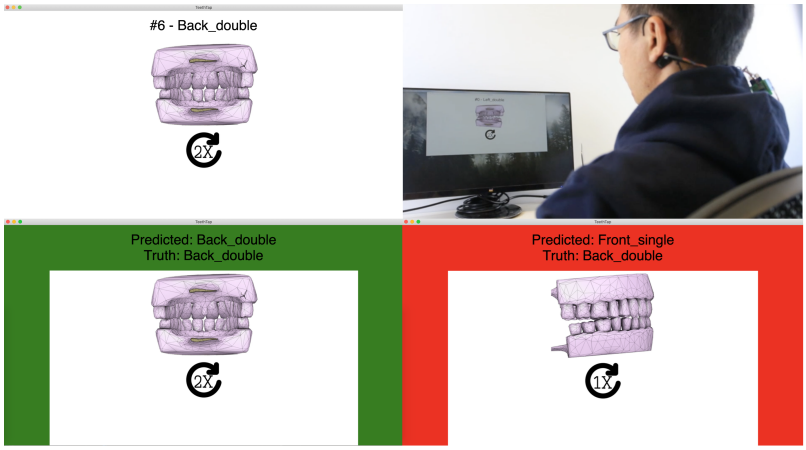

Fig. 7. The GUI of the user study

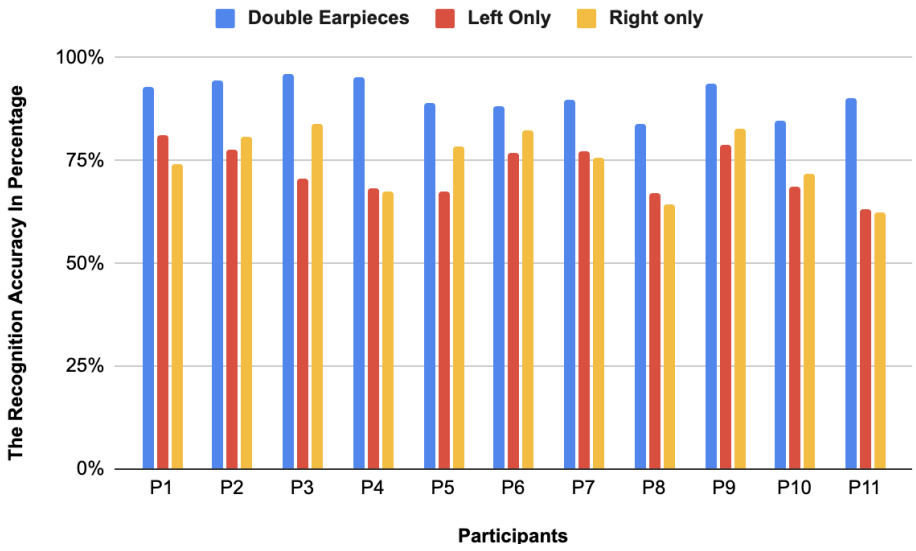

Fig. 8. The recognition accuracy for 11 participants in the testing sessions

our participants finished the five sessions, we asked their subjective feedback on our system, potential applications, and improvements.

5.3 Results

5.3.1 Results from Both Earpieces. In the two testing sessions of recognizing 13 different gestures, we found that our participants reached an average accuracy of 90.9% (SD = 4.1%). Within the 1382 total teeth gestures from 11 participants, TeethTap successfully recognized 1256 gestures. Fig. 8 demonstrates each participant's individual teeth gestures recognition accuracy. We found that P3 and P8 had the highest and lowest accuracy of 96.2% and 83.9%, respectively. There were only five holding gesture instances where the system failed to detect the release gesture. As shown by the confusion matrix presented in Fig. 9, the back-triple gesture had the highest accuracy, which reached over 99.1%. Among all different gestures, the left-hold-gesture had the lowest accuracy of 81.9%. By analyzing the false positive from the confusion matrix, we found that the right-hold-gesture is most likely to be falsely recognized as back-hold-gesture (9.1%). Overall, confusion was more prominent among similar gestures, such as single and holding gestures (holding gesture is a single tap gesture with a delayed release).

5.3.2 Comparison between Single and Double Earpieces. To understand whether having both earpieces are necessary and which gestures are less prone to errors by only having sensors on one side, we further analyzed the accuracy with Left-only earpiece or right-only earpiece. We used the segmentation data saved in the training session and two testing sessions and followed the same data processing pipeline as both earpieces. We found that the average accuracy dropped 18.4% to 72.4% for only using the left earpiece, and it dropped 16.0% to 74.8% for the right one (Fig. 8).

In our gesture set, there are three "manners" of contact: double-tap, single-tap, and hold. To understand whether having a single earpiece affects the accuracy, we relabeled the data by only dividing it into three groups: double-tap, single-tap, and hold. From the results (Fig. 10 a), we first found that the average accuracy across all eleven participants in the testing sessions reached 96.1% to recognize these three different gesture groups by using both earpieces. By only

having single side earpiece, we found the average accuracy only decreased by 1.6% for the left earpiece and 2.9% for the right earpiece, respectively. Therefore, we can conclude that using a single side earpiece could reach relatively similar accuracy as double-side earpieces to classify among single-tap gestures, double-tap gestures, and hold gestures.

To understand the effect of earpiece positions on different teeth-contact areas, we relabeled the data as front-teeth-tap, back-teeth-tap, right-teeth-tap, and left-teeth-tap. We found that the average accuracy in classifying the four kinds of gesture groups with both-side earpieces stayed around 90.9% (Fig. 10 b). However, the accuracy decreased dramatically to 74.9% with the left only earpiece and 75.1% with right only earpiece, respectively. From the results, we found that the accuracy with both-side earpieces are vital to recognize teeth gestures are different positions.

To further explore the accuracy correlation between the position of earpiece placement and the teeth-tap position, we first conducted a comparative analysis of the accuracy of left-teeth-tap gestures with the left only earpiece and right only earpiece. For the three left-teeth-tap gestures (i.e., 'left-single-tap,' 'left-double-tap,' and 'left-hold'), the average accuracy with left only earpiece (93.2%) outperformed 3.3% than the right earpiece (89.9%). On the other side, we also analyzed the same results for right-teeth-tap gestures (i.e., 'right-single-tap,' 'right-double-tap,' and 'right-hold'). We found that the average accuracy with a left only earpiece (88.4%) was 5.8% less than the right one (94.5%). Therefore, the accuracy is higher for a single earpiece the recognize the teeth gestures that reside on the same side as the earpiece.

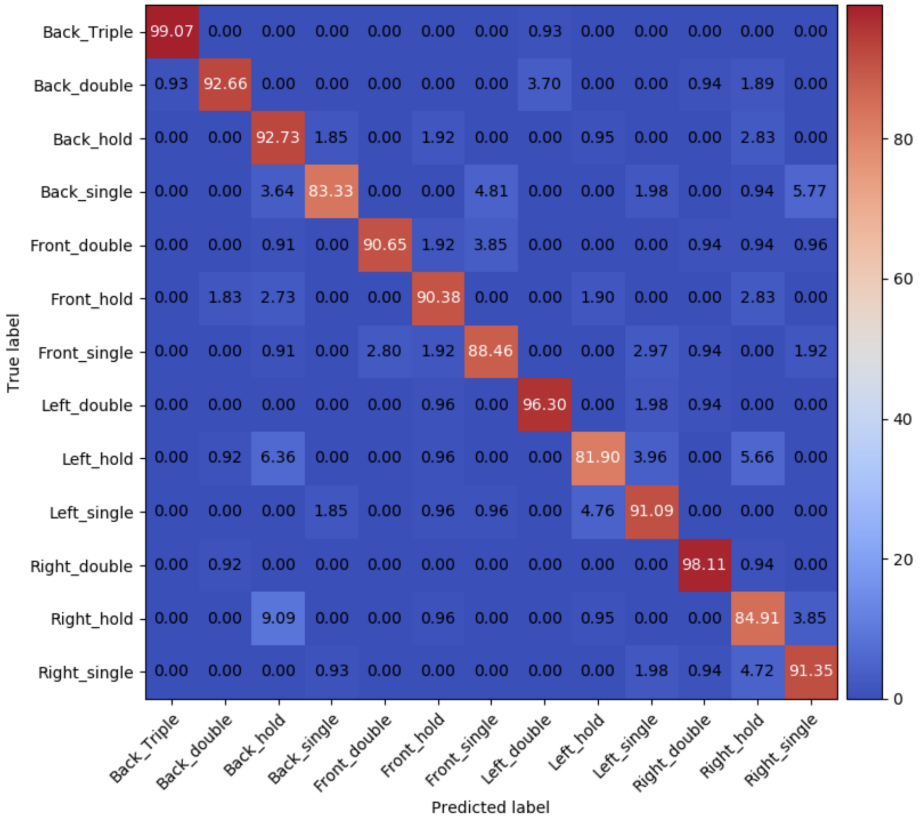

Fig. 9. The confusion matrix for recognition accuracy of the testing sessions

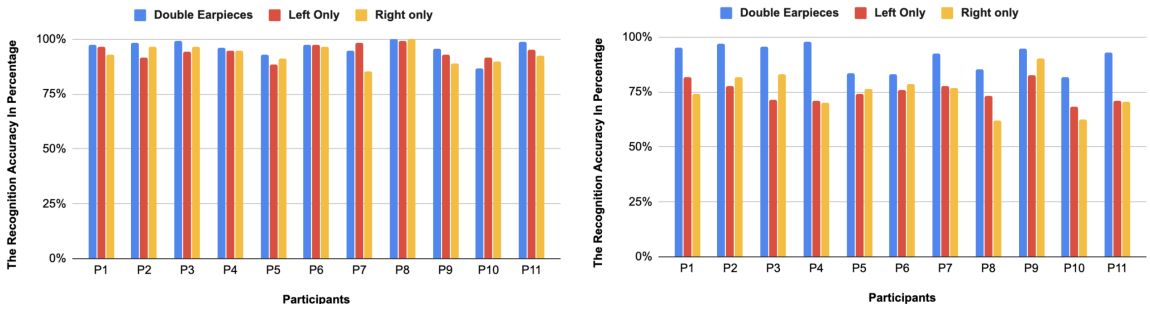

Fig. 10. a) The accuracy of different channels for 'manner' gestures b) The accuracy of different channels for four teeth-tap positions

5.3.3 Remounting Effects and Subjective Feedback. In our study, we conducted a remounting session to understand whether taking the earpieces off and putting them back would affect the accuracy. Overall, we found the average accuracy of recognizing 13 gestures across 11 participants reached 85.3%, which dropped 5.5% from the testing sessions. Therefore, we agree that having the participant remount the sensor by themselves may affect the recognition performance. By analyzing the accuracy changes across different participants, we found that P5 (-12.1%), P3 (-11.5%), and P9 (-10.9%) had the accuracy dropped over 10% in the remounting session. After finishing the study, P5 mentioned his experiences of the remounting session and concerns on making sure the system stays at the relatively same position every time:

> "...To be honest, I forgot where the previous position was after I took it off and trying to put it back. Therefore, I would recommend the researchers to design the artifact that fit on a fixed position on my ears, such as using my ears' shape and force to keep the sensor at the same position, just like the sporting earphones, they always fix at the same place when I use them..."

We further analyzed the results to uncover how does remounting affect the performance of different gesture sets. For 'manner' gestures, we found that the performance only dropped 3% from 96.1% to 93.1% after the participants remounted our prototype. Therefore, we revealed that 'manner' gestures are less influenced by remounting the devices.

6 DISCUSSION

From our study and findings, we uncovered the feasibility of leveraging IMU sensors on earpieces to track teeth gestures with an accuracy of 90.9% on average. We also introduced what gesture sets were more error-prone to a single earpiece and whether the subjective feedback and design implications for remounting the device. In this section, we further discuss activation gestures for in-the-wild scenarios, and the opportunities and challenges of deploying TeethTap in real-world future applications.

6.1 In-the-wild Scenario

From the findings, we found that TeethTap successfully recognized 99.1% for the back-triple gesture. We then conducted a short evaluation to explore whether the back-triple gesture could function as an activation gesture, such as "Hey Siri," to reduce the concerns from false positives. For the "in-the-wild" evaluation, we evaluated how well the activation gesture works while conducting different daily activities. For the same participant group, they were instructed to conduct the following activities in sequence: talking with the researcher, writing on a paper while talking, walking or running around the lab, and eating or drinking. At ten randomized intervals during this process, the researcher asked

the participant to perform an activation gesture. Throughout this process, TeethTap was running in real-time on a laptop to detect activation gestures (binary classification). If the performed activation gesture was not detected, we counted the attempt as a false-negative error. If the participant did not perform a gesture and the system detected an activation gesture, we counted this as a false-positive error. The recognition model was built using the five activation gesture instances collected in the previous training sessions.

Eleven participants tested the activation gesture while performing various activities over a total span of 71 minutes and 33 seconds. Among all Eleven participants, zero false-positive errors were triggered. However, we detected 23 false-negative errors from the 133 gestures. One thing worth mentioning here is that the training data for the activation gesture (back-triple gesture) was collected in the training session while the participants were sitting still in a chair. The added motion introduced from the prescribed activities likely influenced recognition performance. However, we intentionally designed the system to avoid false-positive errors while being more tolerant of false-negative errors, since false-positive errors arguably interfere more with performing daily activities. Therefore, future research could leverage a similar approach to generate an activation gesture to prevent false positives.

6.2 Applications and Gesture Sets

TeethTap offers up to 13 discrete teeth input gestures with an average accuracy rate of 90.9%. Our participants in our study showed strong interests in embedding teeth gestures to control their smart devices. P4 specifically mentioned the benefit of teeth gestures on privacy and 'faster response':

> "...I think the key benefit of having teeth gestures as another input modality is you can make instant responses without even take out the smart devices. Such as playing or pausing music, taking a phone call, I could simply use teeth taps to interact with my smart devices, especially I want this to be applied to my Airpods. Another benefit is that nobody else knows what I did, this will be very useful if I want to reject a call in a meeting..."

However, to interact with most applications, we may not need to recognize all 13 input gestures at the same time. In other words, a subset of the 13 gestures may be enough for many applications, enabling an even higher accuracy rate. In this section, we discuss potential applications of TeethTap and map possible gesture subsets to each application.

6.2.1 Navigating through audio or video content. The task of navigating through audio or video content could call for the following five gestures: back single (pause/play), left single (previous track), right single (next track), left double (rewind), and right double (fast-forward). The accuracy of recognizing these five gestures is 92.3% using training and testing data from the user study.

6.2.2 Volume control. The holding gesture is designed to provide continuous input, such as changing the volume. A user could simply hold down a gesture to raise or lower volume and release the gesture when the volume has reached the desired level. In this application, only two gestures would be needed: left hold (turn down the volume) and right hold (turn up the volume). The accuracy of recognizing these two gestures is 93.2% using training and testing data from the user study.

6.2.3 Operating phone call or videochat. Phone calls often come at socially inappropriate times. TeethTap could provide discreet gestures that are eyes-free and hands-free to operate a call with two gestures: back double (accept the call) and back single (reject call/hand up). The accuracy of recognizing these two gestures is 98.6% using training and testing

data from the user study. Even in the remounting session, we found that TeethTap could still successfully recognize these two gestures at an accuracy of 94.6%.

6.3 Integrating TeethTap to existing head-worn devices

The current form factor is an independent earpiece, as shown in Figure 5. However, we envision TeethTap could be easily adopted into the form factor of existing earphones, headphones, VR headsets or Glass frame technologies. The key integration step is to attach IMUs and contact microphones behind the ear. The form factor could be an extended piece attaching to a Glass frame. Sensors could also be embedded in headphones, following the curvature of the headphone ear-pad around the back of the ear. We believe that integrating such a change would require only hardware alterations, with no changes in the algorithm being necessary.

6.4 Improving the performance of session-independent models

The goal of TeethTap is to provide a user-dependent, but session-independent technology to recognize discrete teeth gestures. In other words, the user needs to provide a few training samples (e.g. five instances per gesture) when they use TeethTap for the first time. However, they should not have to recollect training data every time they wear the device. The testing results from the fifth session showed that after taking off the device and putting it back on again, the recognition accuracy of TeethTap decreased to around 85.3%. Apparently, there is room for improvement in the performance.

There are several potential solutions that can help improve TeethTap's performance as a session-independent input technology. Firstly, we can improve the design of the form factor, by improving its precision in applying consistent pressure to the same areas of the body every time it is put on. After all, form factor displacement between sessions is the primary reason for this performance decrease across sessions. Secondly, we can further process sensor data (e.g. normalization) to account for deviations in form factor positioning. Lastly, we could also consider utilizing acoustic sensor data in the gesture classification process (not just the segmentation process), as wearing a position would likely have less of an effect on the acoustic sensor. For instance, we can calculate energy differences between the left and right acoustic sensors to reliably dictate whether an incoming gesture comes from the left side or right side of the mouth.

7 LIMITATION AND FUTURE WORK

TeethTap demonstrates the proof-of-concept for detecting a rich set of discrete teeth gestures using an earpiece. However, we do find several limitations and future work from our user evaluation and prototype design. In the current user study, we did not evaluate TeethTap's performance in recognizing 13 gestures when the user is in motion (e.g. walking, running), which can be limiting in the context of daily life. In the future, we plan to further optimize our system to be functional while the user is moving. There are a few solutions we plan to explore to achieve this feat: 1) we plan to build two separate models—one for static posture and the other for motion—allowing our system to toggle between modes depending on context; 2) we plan on collecting a larger set of training samples and using more advanced machine learning techniques. Furthermore, our participants were from 21 to 34, which lead to being unknown about how well do aging population perform in our study by using our system. In future work, we will further conduct a study with older adults and also discover how well does TeethTap help people with motor impairments who have problems using their smart devices to provide input commands. Although we claimed the existing limitations of our current work, we do believe the current approach has proved the feasibility of leveraging motion tracking on earpieces and combined with noise-filtering from acoustic sensing to recognize different teeth gestures.

8 CONCLUSION

In this paper, we present TeethTap, a wearable technology that can recognize up to 13 discrete teeth gestures. It uses an earpiece which attaches an IMU sensor and a contact microphone behind both the left and right ears. A KNN-based (with the distance measurement of DTW) algorithm is developed for gesture recognition. A user study with 11 participants shows that it can recognize 13 gestures with an accuracy of 90.9%. We also uncovered the importance of having both-side earpiece available when recognizing position-based gestures comparing with the left-only or right-only earpiece. We further showed the sufficiency of only using a single earpiece to leverage motion sensing to recognize 'manner' based gestures. In the discussion, we introduced the approach of reducing false positives through an in-the-wild evaluation with an activation gesture. We also discussed the opportunity and challenges of widely deploying TeethTap on real-world devices in the future. We believe that by fusing motion and acoustic sensing into a minimalist earpiece, TeethTap offers a promising set of novel interactions for future applications.

REFERENCES

[1] Dalka , Piotr , and Czyzewski Andrzej. 2010. Human-computer interface based on visual lip movement and gesture recognition. *International Journal of Computer Science Applications* 7 (01 2010).

[2] Adafruit. [n.d.]. Adafruit HUZZAH32 – ESP32 Feather Board ID: 3405 - $19.95 : Adafruit Industries, Unique & fun DIY electronics and kits. https://www.adafruit.com/product/3405. (Accessed on 10/03/2020).

[3] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. Canalsense: Face-related movement recognition system based on sensing air pressure in ear canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 679–689.

[4] Daniel Ashbrook, Carlos Tejada, Dhwanit Mehta, Anthony Jiminez, Goudam Muralitharam, Sangeeta Gajendra, and Ross Tallents. 2016. Bitey: An exploration of tooth click gestures for hands-free user interface control. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 158–169.

[5] Michael Barz, Andreas Bulling, and Florian Daiber. 2015. Computational Modelling and Prediction of Gaze Estimation Error for Head-mounted Eye Trackers.

[6] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 37 (Sept. 2017), 20 pages. https://doi.org/10.1145/3130902

[7] Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald A. Peterson, Kofi Odame, Kelly Caine, Ryan J. Halter, Jacob Sorber, and David Kotz. 2018. Auracle: Detecting Eating Episodes with an Ear-mounted Sensor. *IMWUT* 2 (2018), 92:1–92:27.

[8] Murtaza Dhuliawala, Juyoung Lee, Junichi Shimizu, Andreas Bulling, Kai Kunze, Thad Starner, and Woontack Woo. 2016. Smooth eye movement interaction using EOG glasses. In *ICMI*.

[9] Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans-Werner Gellersen. 2015. Orbits: Gaze Interaction for Smart Watches using Smooth Pursuit Eye Movements. In *UIST*.

[10] Augusto Esteves, David Verweij, Liza Suraiya, Md. Rasel Islam, Youryang Lee, and Ian Oakley. 2017. SmoothMoves: Smooth Pursuits Head Movements for Augmented Reality. In *UIST*.

[11] S. Furui. 2000. Speech recognition technology in the ubiquitous/wearable computing environment. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, Vol. 6. 3735–3738 vol.6. https://doi.org/10.1109/ICASSP.2000.860214

[12] Pablo Gallego Cascón, Denys JC 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*. 1–13.

[13] TDK InvenSense. 2014. MPU-9250, Nine-Axis (Gyro+ Accelerometer+ Compass) MEMS MotionTracking™ Device.

[14] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019. Interferi: Gesture Sensing Using On-Body Acoustic Interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[15] Shin Katayama, Akhil Mathur, Marc Van den Broeck, Tadashi Okoshi, Jin Nakazawa, and Fahim Kawsar. 2019. Situation-Aware Emotion Regulation of Conversational Agents with Kinetic Earables. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 725–731.

[16] Fahim Kawsar, Chulhong Min, Akhil Mathur, Alessandro Montanari, Utku Günay Acer, and Marc Van den Broeck. 2018. eSense: Open Earable Platform for Human Sensing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 371–372.

[17] Koichi Kuzume. 2011. Tooth-touch Sound and Expiration Signal Detection and Its Application in a Mouse Interface Device for Disabled Persons - Realization of a Mouse Interface Device Driven by Biomedical Signals. In *PECCS*.

[18] Seungchul Lee, Chulhong Min, Alessandro Montanari, Akhil Mathur, Youngjae Chang, Junehwa Song, and Fahim Kawsar. 2019. Automatic Smile and Frown Recognition with Kinetic Earables. In *Proceedings of the 10th Augmented Human International Conference 2019*. 1–4.

[19] Cheng-Yuan Li, Yen-Chang Chen, Wei-Ju Chen, Polly Huang, and Hao-hua Chu. 2013. Sensor-embedded Teeth for Oral Activity Recognition. In *Proceedings of the 2013 International Symposium on Wearable Computers* (Zurich, Switzerland) *(ISWC '13)*. ACM, New York, NY, USA, 41–44. https://doi.org/10.1145/2493988.2494352

[20] Katsutoshi Masai, Yuta Sugiura, Katsuhiro Suzuki, Sho Shimamura, Kai Kunze, Masa Ogata, Masahiko Inami, and Maki Sugimoto. 2015. AffectiveWear: towards recognizing affect in real life. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 357–360.

[21] Tamer Mohamed and Lin Zhong. 2006. *Teethclick: Input with teeth clacks*. Technical Report. Technical Report. Rice University.

[22] Mouser. [n.d.]. BU-30179-000 Knowles | Mouser. https://www.mouser.com/ProductDetail/Knowles/BU-30179-000?qs=wo4x%252BUeoG8Wi4aOonYRo4g==. (Accessed on 10/03/2020).

[23] Alexander Ng, Stephen A Brewster, and John Williamson. 2013. The impact of encumbrance on mobile interactions. In *IFIP Conference on Human-Computer Interaction*. Springer, 92–109.

[24] 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*. ACM, 269–282.

[25] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[26] Ville Rantanen, Hanna Venesvirta, Oleg Spakov, Jarmo Verho, Akos Vetek, Veikko Surakka, and Jukka Lekkala. 2013. Capacitive measurement of facial activity intensity. *IEEE Sensors Journal* 13, 11 (2013), 4329–4338.

[27] 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* (Victoria, BC, Canada) *(UIST '09)*. ACM, New York, NY, USA, 177–180. https://doi.org/10.1145/1622176.1622209

[28] 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*. 177–180.

[29] T. Simpson *, C. Broughton, M. J. A. Gauthier, and A. Prochazka. 2008. Tooth-Click Control of a Hands-Free Computer Interface. *IEEE Transactions on Biomedical Engineering* 55, 8 (Aug 2008), 2050–2056. https://doi.org/10.1109/TBME.2008.921161

[30] Tyler Simpson, Michel Gauthier, and Arthur Prochazka. 2010. Evaluation of tooth-click triggering and speech recognition in assistive technology for computer access. *Neurorehabilitation and neural repair* 24 2 (2010), 188–94.

[31] Yusuke Sugano and Andreas Bulling. 2015. Self-Calibrating Head-Mounted Eye Trackers Using Egocentric Visual Saliency. In *UIST*.

[32] Kazuhiro Taniguchi, Hisashi Kondo, Mami Kurosawa, and Atsushi Nishikawa. 2018. Earable TEMPO: a novel, hands-free input device that uses the movement of the tongue measured with a wearable ear sensor. *Sensors* 18, 3 (2018), 733.

[33] Kazuhiro Taniguchi and Atsushi Nishikawa. 2018. Mouthwitch: A Novel Head Mount Type Hands-Free Input Device that Uses the Movement of the Temple to Control a Camera. *Sensors* 18, 7 (2018), 2273.

[34] Gineke A ten Holt, Marcel JT Reinders, and EA Hendriks. 2007. Multi-dimensional dynamic time warping for gesture recognition. In *Thirteenth annual conference of the Advanced School for Computing and Imaging*, Vol. 300.

[35] Cheng Zhang, Anandghan Waghmare, Pranav Kundra, Yiming Pu, Scott M. Gilliland, Thomas Plötz, Thad Starner, Omer T. Inan, and Gregory D. Abowd. 2017. FingerSound: Recognizing unistroke thumb gestures using a ring. *IMWUT* 1 (2017), 120:1–120:19.

[36] Qiao Zhang, Shyamnath Gollakota, Ben Taskar, and Raj PN Rao. 2014. Non-intrusive tongue machine interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2555–2558.

[37] Xiaoyi Zhang, Harish Kulkarni, and Meredith Ringel Morris. 2017. Smartphone-based gaze gesture communication for people with motor disabilities. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2878–2889.

[38] 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* (Santa Barbara, California) *(ETRA '12)*. ACM, New York, NY, USA, 341–344. https://doi.org/10.1145/2168556.2168632