# *Whispering Wearables*: Multimodal Approach to Silent Speech Recognition with Head-Worn Devices

Tanmay Srivastava*
Computer Science, Stony Brook University
United States
tsrivastava@cs.stonybrook.edu

R. Michael Winters
Microsoft Research Labs, Microsoft Corporation
United States
mikewinters@microsoft.com

Thomas M. Gable
Microsoft Corporation
United States
thomas.gable@microsoft.com

Yu-Te Wang
Academia Sinica
Taiwan
yutewang@microsoft.com

Teresa LaScala
Microsoft Research Labs, Microsoft Corporation
United States
teresalascala@microsoft.com

Ivan J. Tashev
Microsoft Research Labs, Microsoft Corporation
United States
ivantash@microsoft.com

## ABSTRACT

Silent speech recognition has emerged as a promising approach for enabling hands-free and discreet interaction with head-worn devices. In this paper, we present *QuietSync*, a multimodal system that combines inertial measurement unit (IMU) and contact electrode (ExG) signals to achieve accurate silent speech recognition using off-the-shelf devices. *QuietSync* utilizes an IMU attached to the lower part of the headphones near the ear and strategically places ExG electrodes on the headphones, glasses (nose and behind the ear), and face (for VR applications) to capture subtle movements and muscle activity associated with silent speech production. We conducted a user study with 9 participants and successfully recognized 12 commands with an accuracy of 94.2%. Our system leverages the complementary nature of IMU and ExG signals to enhance the robustness and reliability of silent speech recognition. The IMU captures subtle movements of the jaw and facial muscles, while the ExG electrodes detect low-amplitude surface muscle activity associated with speech production. We show that our system is not affected by the length and speech mannerisms of the commands, and can be fine-tuned for users of varied native languages with only 5 samples. Our findings demonstrate the feasibility of using off-the-shelf head-worn devices to enable silent speech recognition, opening up new possibilities for seamless and discreet interaction with devices such as VR/AR headsets and earables. To the best of our knowledge, *QuietSync* is the first system to enable silent speech interaction for multiple form factors.

## CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

## KEYWORDS

Silent speech recognition; Accessibility; EXG, and IMU sensing

**ACM Reference Format:**
Tanmay Srivastava, R. Michael Winters, Thomas M. Gable, Yu-Te Wang, Teresa LaScala, and Ivan J. Tashev. 2024. *Whispering Wearables*: Multimodal Approach to Silent Speech Recognition with Head-Worn Devices. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685720

## 1 INTRODUCTION

Silent speech interfaces (SSIs) are revolutionizing communication technologies [19] by enabling speech recognition through the capture of articulatory movements [15, 16, 31, 67] and neural signals [17], rather than vocal output. These interfaces enhance privacy and discretion, allowing users to command devices without being overheard, and provide communication alternatives for those with speech impairments. They are particularly useful in noisy settings and aiding accessible interactions where traditional voice recognition systems falter [25, 28] . In addition to the benefits specific to SSIs, speech as an input modality—whether vocalized or silent—offers several advantages over manual input methods such as typing or touch, especially in mobile contexts [60] . It enables hands-free operation, which is essential for accessibility and multitasking scenarios, and speeds up the interaction, allowing for more natural and efficient communication with devices.

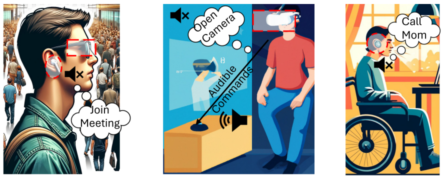
Previous works in the field of SSIs have explored various approaches to capture and interpret silent speech. Most of the approaches involve tracking one or multiple articulators like lips and tongue. Earlier studies have focused on visual methods [20, 56, 68, 75], wireless signals [3, 30], RFID [73] and electromyography (ExG) [27, 43, 48]. In recent times researchers have investigated the use of ultrasound [9, 10, 33] , acoustic signals [15, 26, 76] , and IMU [29, 67] to track tongue and, jaw, and lip movements.

Despite the potential of SSIs, their integration into wearable devices such as earbuds, headphones, VR headsets, and glasses involves complex challenges. Current systems often rely on custom prototypes, which are either intrusive or tailored to a single form factor, limiting their practicality across varied device types. This lack of versatility hinders the widespread adoption of SSIs in consumer electronics. Furthermore, while ExG electrodes are adept at

[*] This work was done when the author was interning at Microsoft Research, Redmond.

*ICMI '24, November 04–08, 2024, San Jose, Costa Rica*

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685720



(a) Private Interaction (b) Virtual Reality (c) Accessibility

**Figure 1: Use cases of *QuietSync*. (a) To enable discreet interaction in public spaces, (b) For interacting in Virtual Reality, and voice commands in "another reality", and (c) As a hands-free alternate accessible modality.**

detecting subtle muscle movements crucial for silent speech, they typically require gel applications and particular sensor placement to ensure sensitivity, which can be cumbersome and impractical for everyday wear. These drawbacks pose significant barriers to the seamless integration of SSIs into wearable devices.

This paper introduces *QuietSync*, a novel multi-modal system that addresses these challenges by combining IMUs and novel, foam-based, dry electrodes (ExG). QuietSync aims to overcome the limitations of previous works by providing a versatile and user-friendly solution for silent speech interaction. The system's key contributions lie in its ability to enable SSI across four different form factors—headphones, earphones, glasses, and VR/XR systems—making it the first of its kind. This breakthrough is achieved through the development of custom, dry, compressible, and novel ExG electrodes that can be seamlessly integrated into head-worn devices, improving usability and comfort for the user.

Moreover, *QuietSync* tackles the challenge of adaptability and personalization through sophisticated signal processing techniques and a lightweight machine-learning model that can be fine-tuned with as few as five samples per user. This approach ensures that the system can quickly adapt to individual users' speech patterns and mannerisms, enhancing its accuracy and reliability.

We demonstrate the effectiveness of *QuietSync* through a controlled study with nine users and twelve commands, achieving over 95% accuracy across different form factors, speech mannerisms, and native languages. This high level of performance showcases the system's robustness and potential for real-world applications. By successfully integrating IMU and ExG sensors into common head-worn devices, earphones, headphones, VR, and glasses, *QuietSync* paves the way for extending the accessibility and usability of mobile and wearable technology through discreet, non-vocal interaction. This advancement could significantly benefit privacy-sensitive applications, enhance accessibility for disabled users, and foster new modes of interaction in next-generation electronics. We show some of the potential applications of *QuietSync* in Figure 1 .

In the subsequent sections, we provide background on silent speech technologies, review related work, and detail the hardware and system design of *QuietSync*. Next, we describe the implementation and data collection methods, followed by an evaluation of the system's performance across different form factors. We conclude the paper with a discussion of our findings and their implications for the design of effective silent speech interfaces.

## 2 BACKGROUND

In this section, we will discuss human speech articulation and articulator sensing via IMU and ExG electrodes.

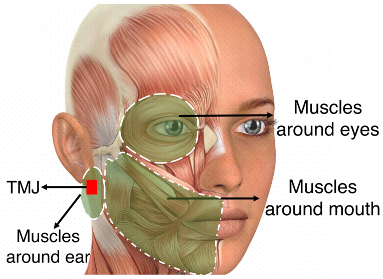


**Figure 2: Temporo-mandibular Joint (TMJ) and the muscle groups involved in speech production.**

*Note: Image Credit - Shutterstock*

### 2.1 Human Speech Articulation

Human speech articulation involves the movement of articulators, such as the lips, tongue, alveolar ridge, and hard and soft palate to produce speech sounds. The articulators are controlled by muscles that contract and relax to create the complex movements required for speech production [11]. While the jaw is not actively involved in speech articulation, it moves up and down to felicitate the movement of lips. Due to this, the jaw is also termed a secondary articulator for its weak involvement [51]. The temporomandibular joints (TMJ), located at the junction of the lower jaw and skull, allow the lower jaw to move up and down.

The muscles around the mouth (e.g., obicularis oris) control the shape and movements of the mouth and lips. The muscles around the eyes (e.g. obicularis oculi) contract and pull the skin of the forehead and cheek towards the nose, indirectly participating in speech production. These muscles are tightly connected and work together to produce speech sounds, affecting the contraction and relaxation of other facial muscles simultaneously. The movement of these muscles due to speech-related jaw motion produces subtle vibrations, called *facial vibrations* [44]. There is another class of vibrations present during speech articulation: bone-borne vibrations, that are generated at the vocal cords during audible speech articulation, propagating through bone and muscles around the face to the TMJ [47]. Since we are interested in developing SSI, we filter out bone-borne vibrations using a low pass filter [29]. Figure 2 shows the muscle groups involved in speech articulation and TMJ.

### 2.2 Articulator Sensing

The research community has recognized multiple sensing modalities, EMG, ExG, IMU, Utltrasound, Camera, WiFi, and recently mmWave, for sensing the movement of articulators associated with silent speech. In our work, we focus on IMU and ExG and explain our design choice in Section 4. IMU usually consists of a gyroscope measuring the rate of change of angle, an accelerometer measuring linear acceleration, and a magnetometer measuring the orientation and rotation of the head regarding the Earth's magnetic field. We use accelerometer and gyroscope in our system as magnetometer can be highly sensitive to electromagnetic noise [34]. Accelerometers can suffer from long-term drifts and gyroscopes with jerk noise and techniques like Kalman filter have been used in the past for noise suppression [37]. However, our data window is comparatively small (commands), therefore *QuietSync* is not very much affected by these noises.

We use custom electrodes (Section 4) to sense facial vibrations. These electrodes measure the electrical signals from muscles to the nervous system during contraction and relaxation. The electrode's conductive material, such as silver/silver chloride (Ag/AgCl), allows it to convert the ionic currents from the body into electrical currents that can be measured by the recording device. These minute, low-voltage signals are amplified for accuracy sensing. We use multi-step signal processing to remove effect of power line and body artifacts for robust sensing as described in Section 5 .

# 3 RELATED WORK

Silent Speech interfaces have been studied and researched thoroughly as an efficient alternative mode of interaction, with users widely accepting it [54]. SSIs broadly fall into two categories: contact-based methods and contactless methods.

## 3.1 Contactless Methods

Contactless methods use sensors that do not require physical contact with the user's face and use wireless signals to gauge the movement of articulators. Earliest contactless methods employed the use of camera-based systems to track facial movements of lips [6, 20, 32, 53, 56, 66, 68, 75]. Since these systems track a primary articulator and the availability of large video datasets they are accurate and robust. However, these systems are sensitive to lighting conditions, are not portable, require a clear line of sight to the user's face, and raise serious privacy concerns [6]. Along with camera, wireless signals, WiFi [71, 72] and Radio Frequency (RF) signals [14, 63], and ultrasound [9, 10, 33] have been used to track articulator movements. While these systems overcome the privacy concerns of camera-based systems, they are sensitive to environmental conditions, require calibration, and are susceptible to interference from other devices. Recently, the research community has employed acoustic sensing on mobile devices to capture articulator movements [15, 23, 24, 40, 70, 76, 79]. The idea is to transmit ultrasound signals from mobile devices and capture the channel response using microphones. The channel response is then used to infer articulator movements. These systems are portable, do not require a line of sight, and are less sensitive to environmental conditions. However, they still require holding the phone in the hand, which might not be feasible in all scenarios like driving or accessibility needs. *QuietSync* overcomes the limitations of these systems and provides a real-time hands-free, and portable system for silent speech recognition.

## 3.2 Contact-based Methods

Contact-based methods often require one or multiple sensors placed on the face, inside the mouth, or on the articulators to infer unvoiced speech. Electromyography (EMG) sensors are used to capture muscle activity associated with lips, jaw, and cheeks during speech production [27, 43]. These sensors are often not socially acceptable as skin electrodes are attached to the user's face around the cheek and lips [48] or require either the use of gel electrodes or specialized form-factor, hence can not be easily integrated with commercial wearable products. Sensors placed on the articulators [18, 22, 31, 38, 59, 61, 73], or even sensors retrofitted to masks [21] can capture the articulators' motion and hence infer unvoiced speech. However, some of these techniques are intrusive, wherein magnetic sensors are mounted on the tongue or inside the mouth, magnets glued to users' tongues, or tattoos placed around users' lips. These systems are not socially acceptable and require calibration. IMUs placed on TMJ have been used to capture jaw movements during speech production. However, these systems require custom form-factor [67] or can only recognize phonemes and not words, rendering their applications limited [29]. Also, researchers have explored retrofitting commercially available head-worn devices with sensors to capture articulator movements [16, 77, 78] or use earphones to capture silent speech associated ear canal deformation for speech recognition [26]. These systems are portable, socially acceptable, and can be integrated with commercial head-worn devices. However, these systems allow for only a single form factor, limiting their applications. In contrast, *QuietSync* is the first multimodal system that can be integrated with a variety of head-worn devices, such as headphones, glasses, and VR headsets, robust, socially acceptable, and portable for silent speech recognition.

## 3.3 Other SSI Systems

Along with speech-based SSI, non-speech-based SSI has been explored. These systems include gaze, tongue gestures, teeth gestures, and hand gestures. Gaze and dwell track the user's eyeball movement and dwell time to infer interaction [2, 74] and perform tasks like typing, selecting, and scrolling in VR/XR. Eye-based gestures have been explored extensively as well for device interaction [4, 42]. While these interactions provide a hands-free and discreet mode of interaction, they are not suitable for all scenarios, like driving, or when the user's hands are occupied and demand continuous attention. Mouth-based interactions include teeth [49, 57, 69] and tongue-based interactions [16, 39, 50] are discreet, require very little movement, and can be integrated with head-worn devices. However, these interactions have a steeper learning curve and are not as intuitive as spoken language. Hand [13, 52, 55] and ear [1, 7, 8] based gestures have been explored and used for device interaction in commercially available wearables. These interactions are intuitive, easy to learn, and reliable. However, these interactions require the user's hands to be free, can have a limited vocabulary, and are not discreet. We believe while *QuietSync* provides a hands-free, discreet, and intuitive mode of interaction, in the future, we can include other modes of interaction like tongue, teeth, and head motion gestures to provide a multimodal interaction system for head-worn devices.

# 4 HARDWARE DESIGN

In this section, we describe the design choices for our *QuietSync*'s prototype, including sensor placement and rationale. Our aim is to develop an unobtrusive, comfortable system that can be easily integrated with existing head-worn devices, requiring minimal or no modifications. We focus our design around popular commercially available devices, such as VR/XR systems (HP Reverb G2 Omnicept Edition, Apple Vision Pro), ExG headbands (Muse 2), smart glasses (Rayban glasses from Meta), and headphones (Microsoft Surface Headphones).

*QuietSync* integrates IMU and ExG sensors, which have been extensively used in developing silent speech recognition systems for

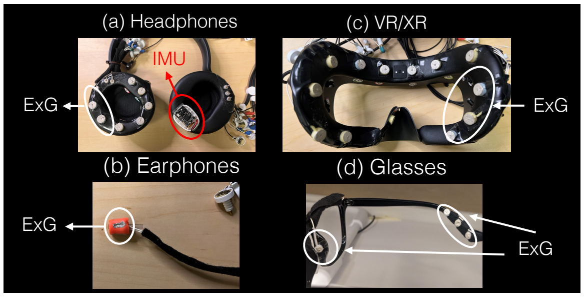


**Figure 3: Sensor placement for *QuietSync* on (a) headphones, (b) customized earphones, (c) VR/XR frame, and (d) customized glasses. White circles represent ExG electrodes, while red circles represent IMUs.**

upper body wearables. Previous studies have shown that IMUs can reliably capture jaw movements and facial muscle activity around the temporomandibular joint (TMJ) [29, 67], while ExG sensors can detect low-amplitude surface muscle activity associated with speech production [5, 27]. This multi-modal combination of sensors allows us to capture both large movements of the articulators near the lower ear and subtle movements and muscle activity of the facial muscles associated with silent speech production.

■**IMU Placement.** Most current ear-worn devices, such as Airpods Max, Bose NC 700, and Bose QC35, are equipped with IMU sensors. Although the IMUs in these devices might not be optimally positioned for capturing jaw movements, they can be repositioned within the earable to enable silent speech sensing. We place the IMU on the lower part of the headphones, near the ear, to capture jaw movements and facial muscle activity around the TMJ. Unlike previous works that placed the IMU directly on the TMJ, we observe that jaw motion signals can be detected from the lower end of the ear, making *QuietSync* integrable with commercially available headphones.

■**ExG Placement.** To determine ExG sensor placement, we studied human facial muscle anatomy and speech articulation presented in Background 2.1. While most current VR/XR or other head-worn devices are not equipped with ExG to capture facial vibrations, we believe that ExG can be augmented in future devices. To this end, we aim to select sensor locations that can be easily integrated. The ExG electrodes are placed on the headphones, glasses (nose and behind the ear), and face (for VR applications) to capture the subtle movements and muscle activity associated with silent speech production.

■**Custom Eletrodes.** To address the limitations of gel or sticker-based electrodes, we developed novel dry electrodes using 3D-printed molds, castable urethane [64] foam, and conductive medical [45] electrode ink. Our electrodes can be installed directly on devices, conforming to different body topographies due to the compressible properties of the urethane foam, thus enhancing comfort and ease of use. Our fabrication process involves designing a two-part mold with a cavity for the electrode and a lid with overflow holes. The cavity includes a 1.5 mm through-hole for wire placement and a channel for wire exit. After applying mold release and allowing it to dry, we thread a partially stripped wire through the hole, ensuring the stripped end extends out of the mold. We then cast the electrode with the chosen urethane foam. To complete the electrode, we clip the exposed wire end flush with the electrode body and coat the skin-facing surface with conductive ink, curing it according to the manufacturer's instructions. Finally, we solder a connector to the remaining wire end. Our choice of materials was based on factors such as durability, skin safety, and electrical conductivity. We chose SmoothOn FlexFoam-It™ 6 for its long pot life and ideal compression properties [65] , and Creative Materials 113-09 ink for its high conductivity, skin safety, and resistance to flexing and creasing [46] . The ink also allows us to create multiple discrete electrodes on a single surface by applying unconnected patches. To achieve successful results, we recommend several precautions, including using a mold release agent, incorporating overflow holes in the mold lid, taping the wire in place before casting, and properly curing the conductive ink. We acknowledge that limitations of our process include ink wear over time, which can be addressed by reapplying and curing the ink, and variation due to manual fabrication. In our future work, we plan to explore embedding electrodes directly into devices and investigating alternative materials for improved durability.

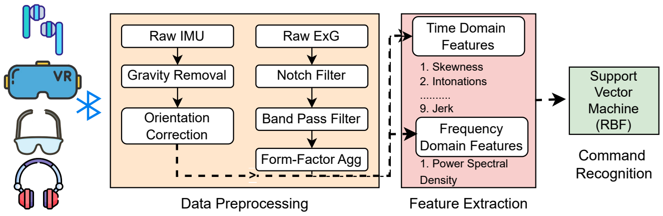


**Figure 4: System design of *QuietSync* for silent speech recognition.**

In summary, the IMU captures the large movements of the articulators around the lower ear, while the ExG electrodes detect the low-amplitude surface muscle activity associated with speech production. Figure 3 shows the placement of the sensors for (a) headphones, (b) earphones, (c) VR/XR, and (d) glasses for *QuietSync*.

## 5 SYSTEM DESIGN

This section elaborates on the design and working components of our system. *QuietSync* enables silent speech interaction with head-worn devices by fusing IMU and ExG signals. Our system design is shown in Figure 4. The system comprises three main components: data pre-processing, feature extraction, and classification. We aim to remove the effect of motion artifacts from ExG and the effect of wearing position from IMU, extract features from the time and frequency domain of the multi-modal system that encodes speech characteristics, and train a lightweight classifier that can be adapted to a new user with as low as 5 samples.

### 5.1 Data Pre-processing

We pre-process the raw IMU and ExG data to mitigate the effects of orientation, gravity, motion, and powerline noise. For the IMU's accelerometer data, we employ a high-pass filter with a 0.5 Hz cutoff frequency to remove the influence of gravity. Additionally, we perform offset removal by subtracting the mean of the first 0.1 seconds of data from the entire window for each axis. To eliminate the impact of orientation on the gyroscope data, we apply an inverse rotation matrix to align each data point to a consistent reference frame [41]. This ensures that the user's device wearing orientation does not affect the system's performance.

For the ExG data, we apply a notch filter at 60 Hz to suppress powerline noise, and then apply a 50 Hz low-pass filter. To mitigate motion artifacts, we employ the common rejection method, subtracting the data of reference electrodes [71] from the electrodes of interest. We also remove the effect of baseline drift by subtracting the mean of the first 0.1 seconds of data from the entire window. Furthermore, we aggregate the data from all electrodes to obtain a single-channel representation for each window and modality. This approach reduces the number of features and helps mitigate the impact of noise in the data.

### 5.2 Feature Extraction

Figure 5 shows time domain representations of all the commands for both modalities. For IMU we show the estimated orientation and for ExG we show summed values for all the electrodes. We extract both time and frequency domain features from the preprocessed IMU and ExG data to capture the subtle movements and muscle activity associated with silent speech production. We use 9 statistical features for the time domain, including mean, standard deviation, skewness, kurtosis, jerk, zero-crossing rate, area under the curve, energy, and range. For the frequency domain, we compute the power spectral density (PSD) using Welch's method with a 1-second window and 50% overlap. We extract the PSD features in the 0-50 Hz frequency range, as this captures the majority of the speech-related muscle activity. To make PSD features invariant to the length of the signal, we normalize them by dividing each frequency bin by the total power in the 0-50 Hz range. We concatenate the time and frequency domain features to create a feature vector for each window and modality. We take the first 8 coefficients of the PSD as features, as they capture the majority of the signal power. Our IMUs are sampled at 100 Hz and ExG at 1000 Hz.

### 5.3 Word classification

After we have preprocessed the data and extracted features, we train an SVM classifier to recognize the silent speech commands. For classification, we employed a Support Vector Machine (SVM) with a radial basis function (RBF) kernel, as it is well-suited for small datasets [12]. We trained both user-dependent and user-independent models to evaluate the system's performance across different scenarios. For the user-dependent classifier, we trained a separate SVM for each user, ensuring personalized silent speech recognition. We utilized 5-fold cross-validation to tune the hyperparameters, specifically, gamma($\gamma$) and cappa($\kappa$), optimizing for accuracy as the primary metric. This approach allowed us to find the best-performing model configuration for each user.

Additionally, we explored a user-independent model to assess the system's ability to generalize across users without the need for extensive individual training. To enhance the performance of the user-independent model, we fine-tuned it using just 5 samples of each word per user. This fine-tuning process helped adapt the model to the specific characteristics of each user's silent speech patterns while minimizing the training data requirements.

| Interaction Type | Commands |
|---|---|
| Device Interaction | Call Mom, Clear Notification, Clear Calendar, Lock Screen |
| Meeting Controls | Close Camera, Open Camera, Mute Microphone, Join Meeting, Leave Call |
| Media Controls | Decrease Volume, Increase Volume, Play Music |

**Table 1: We collect 12 commands from 3 categories.**

By focusing on SVMs and leveraging cross-validation for hyperparameter tuning, we achieve high accuracy in silent speech recognition, even with the limited dataset size.

## 6 IMPLEMENTATION AND DATA COLLECTION

This section describes the data collection process and implementation details of *QuietSync*. We first provide an overview of the data collection setup and software, followed by a detailed description of our user study.

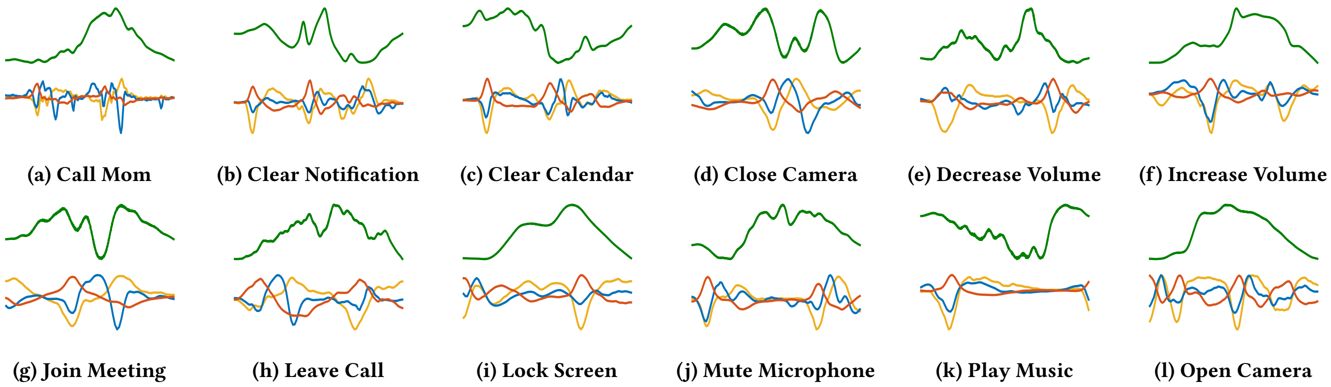
### 6.1 Data Collection Setup

For collecting IMU data, we used a sensor from Mbient [36], sampling the accelerometer and gyroscope at 100 Hz, with a full-scale reading of ± 4g and 250 degrees per second, respectively. To collect ExG data, we employed the BrainVision Recorder [58], labeling electrodes according to their position and sampling at 1000 Hz. In total, we collected data from 28 electrodes. In addition to IMU and ExG data, we recorded audio and video using a laptop camera and external microphone for data sanity checks. All data streams were synchronized using the Lab Streaming Layer (LSL) [35], which enables synchronization through per-sample timestamps and time synchronization for multi-modal interfaces. The data from all streams was stored in an extended data format (XDF) file using the LSL Lab Recorder. Figure 6 illustrates our data collection pipeline.

### 6.2 User Study

We conducted a user study with 9 participants of diverse age groups, ethnicities, and native languages, collecting data for 12 commonly used speech interaction commands, as shown in Table 1. Our data collection study was approved by the Institutional Review Board (IRB) of our institute prior to the study. All participants provided informed consent and were compensated for their time with a gift card. We recruited participants from our organization, and they were free to leave the study at any time without penalty. Table 2 shows the demographics of our participants, which included a mix of age groups, native languages, and facial structures (Asian, White, Black). We required participants to have no history of speech disorders or other medical conditions that could affect their speech production. To maintain hygiene and ensure better contact between the electrodes and the skin, we cleaned the prototypes with an alcohol solution before each data collection session and asked participants to wipe their faces with wet wipes.

We collected data from users in silent and audible manners. We displayed one command at a time on a monitor, using different color codes: green for audible and red for inaudible, randomizing



**(a) Call Mom** **(b) Clear Notification** **(c) Clear Calendar** **(d) Close Camera** **(e) Decrease Volume** **(f) Increase Volume**

**(g) Join Meeting** **(h) Leave Call** **(i) Lock Screen** **(j) Mute Microphone** **(k) Play Music** **(l) Open Camera**

**Figure 5: Time domain representation of the 12 commands for IMU and ExG. Green (Summed ExG values), Yellow (IMU orientation x), Blue (IMU orientation y), and Orange (IMU orientation z).**

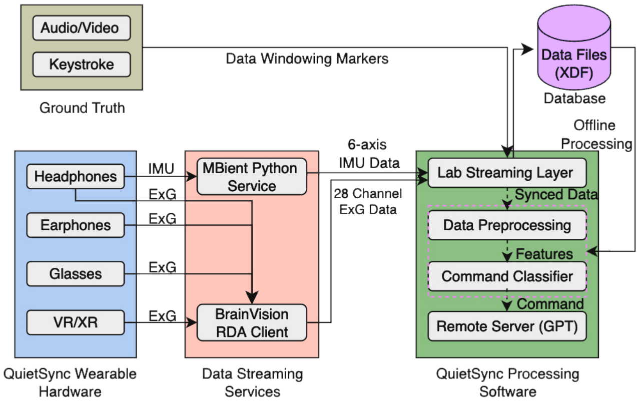


**Figure 6: Data flow of *QuietSync* for offline and online processing**

| | |
|---|---|
| **Age** | 22-31 |
| **Gender** | 5 male, 4 female |
| **Native language** | 3 Hindi, 2 English, 2 Chinese, 2 Italian |
| **Ethnicity** | 5 Asian, 3 White, 1 Black |

**Table 2: User demographics for our study**

the order of commands. Participants were instructed to press the spacebar while articulating each command, and we recorded the timestamps for key presses and releases, which were streamed in the LSL stream and later used to create windows for each command. The participants could pause the session and press the backspace key to display the previous command. We collected 40 samples for each command in the silent manner and 15 samples for each command in the audible manner. Also, we collected 5 samples per command in the silent manner for each user to train a user-independent model. The data was collected in two sessions. In total, we collected 1080 samples for silent speech and 405 samples for audible speech. Figure 7 shows the data collection setup.

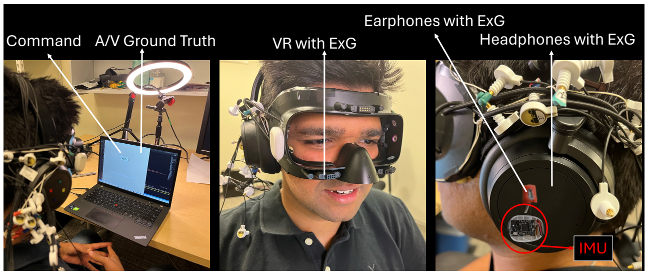


**Figure 7: Data collection setup for *QuietSync*. We collect data from IMU, ExG, audio, and video streams, synchronized using LSL.**

## 7 EVALUATION

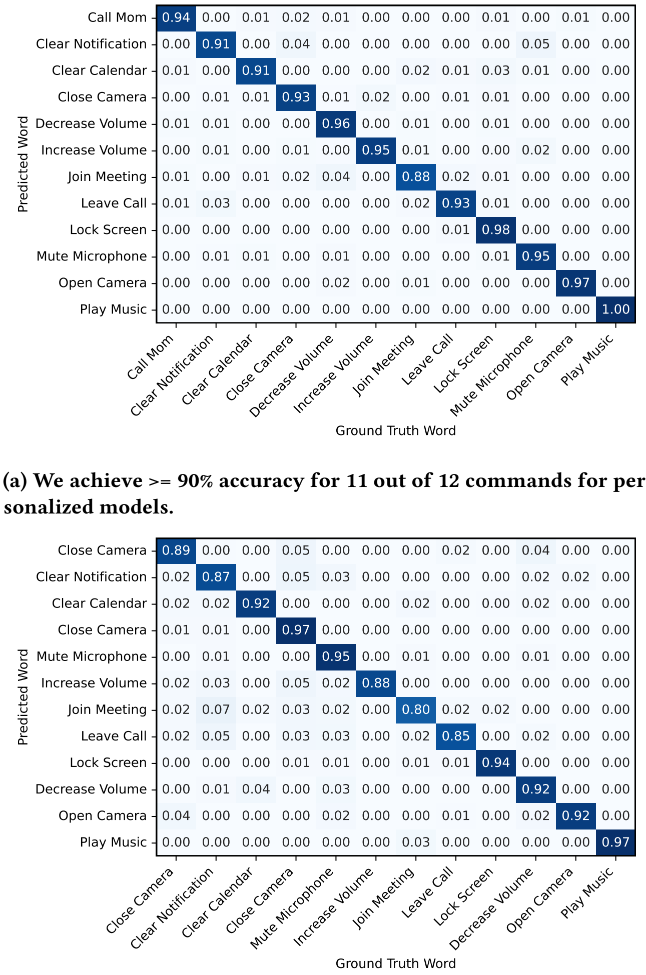
In this section, we will discuss the evaluation of our system, *QuietSync* for silent speech recognition. We will first present the results of our user study, followed by an analysis of the system's performance across different scenarios and settings. We will also discuss the impact of different sensing modalities and form factors on the system's accuracy and reliability. Finally, we will present *QuietSync* s real-time performance. We show that *QuietSync* chieves (1) more than 95% accuracy in user-dependent and > 93% accuracy in user-independent scenarios with only 5 samples (2) is agnostic to the native language of the user, speech mannerisms, and length of the command, and (3) can be integrated with different sensing modalities and form factors.

### 7.1 Overall Performance

We report the accuracy of user-dependent and user-independent models for all commands in Figure 8a and Figure 8b, respectively. We achieve more than 90% recognition accuracy for 11 out of 12 commands for user-dependent models achieve > 90% accuracy, with an average accuracy of 94.2%. To train the user-independent model, we use randomly selected 5 samples for each word and achieve > 90% accuracy in 6 out of 12 commands, with an average accuracy of 93.5%. We rest of the evaluations we report the results for user-dependent models, unless otherwise mentioned.

### 7.2 Performance Across Different Modalities

One of the most crucial aspects of *QuietSync* is its ability to integrate different sensing modalities and form factors. We evaluate the



**(a) We achieve >= 90% accuracy for 11 out of 12 commands for personalized models.**

**(b) Using only 5 samples for each word, we can train a user-independent model and achieve >85% accuracy for 11 commands.**

**Figure 8: Confusion matrix for commands recognition using IMU and ExG.**

performance of *QuietSync* across various modalities (Figure 9(a)), isolated form factors such as glasses, earphones, headphones, and VR/XR headsets (Figure 9(b)), and different form factor combinations, including headphones, glasses, and VR/XR headsets (Figure 9(c)). Remarkably, we achieve over 80% accuracy for all but two combinations, demonstrating the robustness and reliability of *QuietSync* across different settings. Our findings reveal that the combination of IMU and ExG sensors across all modalities provides the best performance, with an impressive accuracy of 94%. This underscores the value of multi-modal signals for silent speech recognition, as they capture both large movements of the articulators and subtle muscle activity associated with speech production. When used independently, ExG sensors achieve an accuracy of 84.3%, while IMU sensors alone reach 79.2%. Although this accuracy is lower than that reported in [67], it is important to note that, unlike their system, *QuietSync* can be integrated with custom headphones. In our analysis of isolated form factors, only glasses and earphones exhibit slightly lower accuracies of 76.5% and 77.3%, respectively. This can be attributed to the limited number of ExG sensors on glasses (only 2 on the nose) and earphones (3 inside the ear) that are in contact with facial muscles to capture facial vibrations. In contrast, headphones (with multiple ExG electrodes and IMU) and VR headsets (multiple ExG electrodes) demonstrate accuracies above 80%. One of the most significant findings of *QuietSync* is presented in Figure 1c, which highlights the performance of combining multiple form factors. We achieve over 85% accuracy when combining glasses, earphones, and headphones across various combinations. We believe these results are particularly meaningful, as most of these form factor combinations can be worn together in daily life, enabling hands-free and discreet interactions.

## 7.3 Impact of Different settings

We evaluate the impact of the native language of the user, speech mannerisms (silent v/s audible), and length of the command on the performance of *QuietSync*.

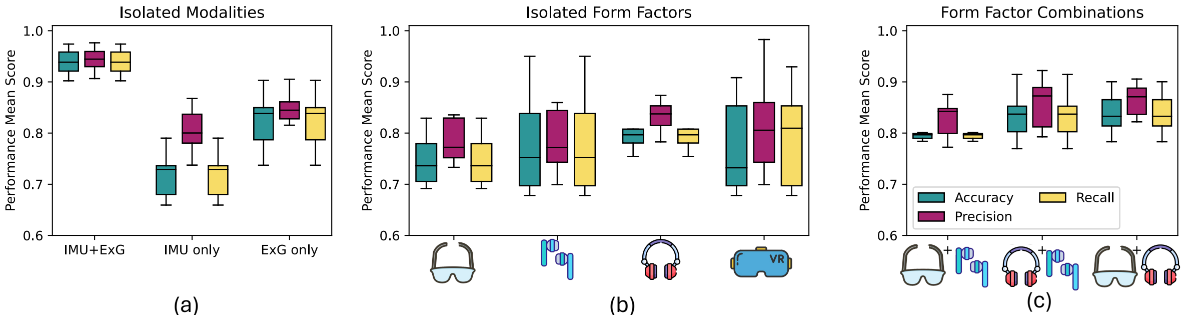
■ **Impact of native language.** Figure 10 shows the mean word recognition accuracy for users with different native languages. We achieve >=90% accuracy for all users, demonstrating the system's agnostic to the native language of the user. This is a significant advantage of *QuietSync* as it can be used by users from different linguistic backgrounds without the need for extensive training or customization. We attribute this ability of the system to multiple sensing locations for ExG. Previous research has shown that people can have varied facial structures based on ethnicity [62]. By placing ExG sensors in different locations, we introduce redundancy in the system, helping us in capturing facial vibrations for different users.

■ **Impact of command length.** We evaluate the system's performance for commands with different lengths. We achieve > 90% accuracy for all commands, demonstrating the system's agnosticity to the length of the command. This is a significant advantage of *QuietSync* as it can be used for wide applications and scenarios, such as short commands for device interaction (Call Mom (94%) ) or longer commands for device control (Clear Notification (91%)). We achieve > 90% accuracy for all commands, highlighting the system's robustness and reliability across different syllable lengths.

■ **Impact of speech mannerisms** We also evaluate the system's performance for silent and audible speech mannerisms by training a user-dependent classifier using silent testing on audible commands, and vice-versa. As shown in Table 3, we achieve > 90% accuracy for both silent and audible speech, highlighting the system's robustness and reliability across different speech styles. This is a significant advantage of *QuietSync* as it can be used in different scenarios and settings, such as noisy environments where the user can speak audibly but traditional speech recognition systems will not work or when the user wants to articulate silently for discreet interaction.

## 7.4 Real-time Performance

We implemented *QuietSync* for real-time evaluation using LSL streaming and Python, testing with five users. To demonstrate its interaction capabilities, we integrated it with GPT-4.turbo for a "Name, Place, Animal, Thing?" game. We mapped four commands to "yes", "no", "maybe", and "you got it". For each user, we saved a user-dependent model. We used an empirical threshold on IMU data for command detection, replacing keyboard keystroke timings.



(a) (b) (c)

**Figure 9: Performance of different sensing modalities and form factors. (a) shows the performance of different modalities, (b) shows the performance of combinations of different modalities for 4 form-factors, and (c) shows the performance of combinations of different form-factors**

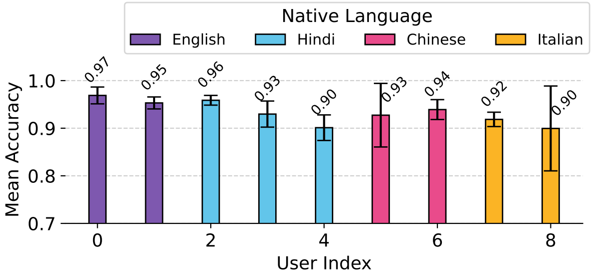


**Figure 10: Mean word recognition accuracy users with different native languages.**

| Classifier Configuration | Train Audible Test Silent | Train Silent Test Audible | Train Audible Test Audible | Train Silent Test Silent |
|---|---|---|---|---|
| **Accuracy%** | 91.5±2.4 | 92.3±1.7 | 90.8±5.9 | 94.2±3.8 |

**Table 3: Performance of *QuietSync* for silent and audible speech mannerisms.**

Data preprocessing and feature extraction ran in parallel on the windowed input stream. Our system achieved a mean latency of 1.35 seconds: 0.05 seconds for windowing, 0.9 seconds for preprocessing and feature extraction, and 0.3 seconds for classification. We believe optimization can reduce this latency further. Each user played the game 10 times. We defined task completion as GPT guessing the word or users correctly saying it 15 times consecutively. We achieved a 90% task completion rate across all users, demonstrating *QuietSync*'s real-time performance and reliability.

## 8 DISCUSSION

*QuietSync* is an early attempt at developing a multi-modal system for silent speech recognition. We have shown that combining IMU and ExG signals can significantly improve the accuracy and robustness of silent speech recognition.

**Limitations.** *QuietSync* to the best of our knowledge, is the first system to enable silent speech interaction with multiple form factors. However, in its current early stage, we have identified the areas of improvement. We tested with a small number of users. Although the sample size is limited, the field of SSI is still developing, and similar studies have provided valuable insights with comparable sample sizes. Our diverse sample includes a range of ages, genders, native languages, and ethnicities 2. Our future studies will expand the sample size to enhance the scope and impact of our findings. Also, our command set is limited and lacks an explicit 'other' category. We selected 12 commands of varying lengths and interaction categories based on a formative study within our organization. Our future iterations will expand the command set, include user-defined commands, and add an 'other' category to improve flexibility and robustness. Finally, our system was evaluated in a controlled environment. Real-world testing is crucial, as factors like body movement and motion artifacts may affect performance. Future work will focus on testing in diverse environments to assess robustness and identify challenges.

**Integrating *QuietSync* with existing systems** We implemented a gamified real-time demo of *QuietSync* to showcase its potential for integration with different services. In the future, we plan to integrate *QuietSync* with existing voice assistants like Cortana, Alexa, Google Assistant, and Siri to enable silent speech interaction with smart devices in day-to-day applications like MS Teams and hands-free typing for people with disabilities.

## 9 CONCLUSION

In this paper, we presented *QuietSync* a multimodal system that combines IMU and ExG signals to enable silent speech recognition with head-worn devices. We conducted a user study with 9 participants and successfully recognized 12 commands with an accuracy of 94.2%. Our system leverages the complementary nature of IMU and ExG signals to enhance the robustness and reliability of silent speech recognition. The IMU captures subtle movements of the jaw and facial muscles, while the ExG electrodes detect low-amplitude surface muscle activity associated with facial vibrations. *QuietSync* is agnostic to speech mannerisms, length of commands, and the native language of the users. Our findings demonstrate the feasibility of using off-the-shelf head-worn devices to enable silent speech recognition, opening up new possibilities for seamless and discreet interaction with devices such as VR/AR headsets and earables.

## ACKNOWLEDGMENTS

We would like to express gratitude for the valuable feedback from Hannes Gamper, Sebastian Braun, Ann Paradiso and David Johnston, as well as the wonderful time spent with peer interns Ard Kastrati, Azalea Gui, Eloi Moliner, Michele Mancusi, Ruihan Yang, and Yuanchao Li at the Audio and Acoustic Research Group, Microsoft Research - Redmond.

## REFERENCES

[1] Andrea F Abate, Michele Nappi, and Stefano Ricciardi. 2016. Smartphone enabled person authentication based on ear biometrics and arm gesture. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 003719–003724.

[2] 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

[3] Peter Birkholz, Simon Stone, Klaus Wolf, and Dirk Plettemeier. 2018. Non-Invasive Silent Phoneme Recognition Using Microwave Signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 12 (Dec. 2018), 2404–2411. https://doi.org/10.1109/TASLP.2018.2865609 Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[4] Andreas Bulling and Hans Gellersen. 2010. Toward mobile eye-based human-computer interaction. *IEEE Pervasive Computing* 9, 4 (2010), 8–12.

[5] Ho-Seung Cha, Won-Du Chang, and Chang-Hwan Im. 2022. Deep-learning-based real-time silent speech recognition using facial electromyogram recorded around eyes for hands-free interfacing in a virtual reality environment. *Virtual Reality* 26, 3 (2022), 1047–1057.

[6] Hang Chen, Jun Du, Yu Hu, Li-Rong Dai, Chin-Hui Lee, and Bao-Cai Yin. 2020. Lip-reading with Hierarchical Pyramidal Convolution and Self-Attention. arXiv:2012.14360 [cs.CV]

[7] Yu-Chun Chen, Chia-Ying Liao, Shuo-wen Hsu, Da-Yuan Huang, and Bing-Yu Chen. 2020. Exploring user defined gestures for ear-based interactions. *Proceedings of the ACM on Human-Computer Interaction* 4, ISS (2020), 1–20.

[8] Zicheng Chi, Yao Yao, Tiantian Xie, Xin Liu, Zhichuan Huang, Wei Wang, and Ting Zhu. 2018. EAR: Exploiting uncontrollable ambient RF signals in heterogeneous networks for gesture recognition. In *Proceedings of the 16th ACM conference on embedded networked sensor systems*. 237–249.

[9] Tamás Gábor Csapó, Csaba Zainkó, László Tóth, Gábor Gosztolya, and Alexandra Markó. 2020. Ultrasound-based articulatory-to-acoustic mapping with WaveGlow speech synthesis. *arXiv preprint arXiv:2008.03152* (2020).

[10] B. Denby, Y. Oussar, G. Dreyfus, and M. Stone. 2006. Prospects for a Silent Speech Interface using Ultrasound Imaging. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. I–I. https://doi.org/10.1109/ICASSP.2006.1660033

[11] Richard P Di Fabio. 1998. Physical therapy for patients with TMD: a descriptive study of treatment, disability, and health status. *Journal of orofacial pain* 12, 2 (1998).

[12] Kai-Bo Duan and S Sathiya Keerthi. 2005. Which is the best multiclass SVM method? An empirical study. In *International workshop on multiple classifier systems*. Springer, 278–285.

[13] Yikai Fang, Kongqiao Wang, Jian Cheng, and Hanqing Lu. 2007. A real-time hand gesture recognition method. In *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 995–998.

[14] David Ferreira, Samuel Silva, Francisco Curado, and António Teixeira. 2022. Exploring silent speech interfaces based on frequency-modulated continuous-wave radar. *Sensors* 22, 2 (2022), 649.

[15] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-Based Silent Speech Interface for Smartphone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 80 (Sept. 2020), 27 pages. https://doi.org/10.1145/3411830

[16] 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 *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 564–573. https://doi.org/10.1145/3577190.3614120

[17] P Ghane, G Hossain, and A Tovar. 2015. Robust understanding of EEG patterns in silent speech. In *2015 National Aerospace and Electronics Conference (NAECON)*. IEEE, 282–289.

[18] Jose A. Gonzalez, Lam A. Cheah, Angel M. Gomez, Phil D. Green, James M. Gilbert, Stephen R. Ell, Roger K. Moore, and Ed Holdsworth. 2017. Direct Speech Reconstruction From Articulatory Sensor Data by Machine Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 12 (2017), 2362–2374. https://doi.org/10.1109/TASLP.2017.2757263

[19] Jose A Gonzalez-Lopez, Alejandro Gomez-Alanis, Juan M Martín Doñas, José L Pérez-Córdoba, and Angel M Gomez. 2020. Silent speech interfaces for speech restoration: A review. *IEEE access* 8 (2020), 177995–178021.

[20] J. Han, L. Shao, D. Xu, and J. Shotton. 2013. Enhanced Computer Vision With Microsoft Kinect Sensor: A Review. *IEEE Transactions on Cybernetics* 43, 5 (2013), 1318–1334. https://doi.org/10.1109/TCYB.2013.2265378

[21] Hirotaka Hiraki and Jun Rekimoto. 2021. SilentMask: Mask-Type Silent Speech Interface with Measurement of Mouth Movement. In *Augmented Humans Conference 2021* (Rovaniemi, Finland) *(AHs'21)*. Association for Computing Machinery, New York, NY, USA, 86–90. https://doi.org/10.1145/3458709.3458985

[22] Robin Hofe, Stephen R. Ell, Michael J. Fagan, James M. Gilbert, Phil D. Green, Roger K. Moore, and Sergey I. Rybchenko. 2013. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Communication* 55, 1 (2013), 22–32. https://doi.org/10.1016/j.specom.2012.02.001

[23] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone. 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication* 52, 4 (2010), 288–300.

[24] Thomas Hueber, Gérard Chollet, Bruce Denby, and Maureen Stone. 2008. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. *Proc. of ISSP* (2008), 365–369.

[25] Madeline Jefferson. 2019. Usability of Automatic Speech Recognition Systems for Individuals with Speech Disorders: Past, Present, Future, and A Proposed Model. *undefined* (2019). https://www.semanticscholar.org/paper/Usability-of-Automatic-Speech-Recognition-Systems-A-Jefferson/73eefd141f43750b3ae0648e6ef099597e24c6c9

[26] 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 (2022), 1–28.

[27] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. Alterego: A personalized wearable silent speech interface. In *23rd International conference on intelligent user interfaces*. 43–53.

[28] Rachel Kaye, Christopher G Tang, and Catherine F Sinclair. 2017. The electrolarynx: voice restoration after total laryngectomy. *Medical Devices (Auckland, NZ)* 10 (2017), 133.

[29] 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*. 44–49.

[30] Rohan Khanna, Daegun Oh, and Youngwook Kim. 2019. Through-Wall Remote Human Voice Recognition Using Doppler Radar With Transfer Learning. *IEEE Sensors Journal* 19, 12 (June 2019), 4571–4576. https://doi.org/10.1109/JSEN.2019.2901271 Conference Name: IEEE Sensors Journal.

[31] 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*. 1–19.

[32] Naoki Kimura, Kentaro Hayashi, and Jun Rekimoto. 2020. TieLent: A Casual Neck-Mounted Mouth Capturing Device for Silent Speech Interaction. In *Proceedings of the International Conference on Advanced Visual Interfaces* (Salerno, Italy) *(AVI '20)*. Association for Computing Machinery, New York, NY, USA, Article 33, 8 pages. https://doi.org/10.1145/3399715.3399852

[33] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *Proceedings of CHI 2019* (Glasgow, Scotland Uk). Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300376

[34] Manon Kok and Thomas B Schön. 2016. Magnetometer calibration using inertial sensors. *IEEE Sensors Journal* 16, 14 (2016), 5679–5689.

[35] Christian Kothe, Seyed Yahya Shirazi, Tristan Stenner, David Medine, Chadwick Boulay, Matthew I Crivich, Tim Mullen, Arnaud Delorme, and Scott Makeig. 2024. The Lab Streaming Layer for Synchronized Multimodal Recording. *bioRxiv* (2024), 2024–02.

[36] Mbient Lab. 2020. Mbient IMU. https://mbientlab.com/metamotionr/

[37] Qiang Li, Ranyang Li, Kaifan Ji, and Wei Dai. 2015. Kalman filter and its application. In *2015 8th international conference on intelligent networks and intelligent systems (ICINIS)*. IEEE, 74–77.

[38] 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

[39] Zheng Li, Ryan Robucci, Nilanjan Banerjee, and Chintan Patel. 2015. Tongue-n-cheek: non-contact tongue gesture recognition. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*. 95–105.

[40] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A Noise-resistant Multi-modal Speech Recognition System Fusing mmWave and Audio Signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems* (Coimbra, Portugal) *(SenSys '21)*. Association for Computing Machinery, New York, NY, USA, 97–110. https://doi.org/10.1145/3485730.3485945

[41] Mark Looney. 2015. The basics of MEMS IMU/Gyroscope alignment. *Analog Dialogue* 49 (2015), 1–6.

[42] Päivi Majaranta and Andreas Bulling. 2014. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*. Springer, 39–65.

[43] Hiroyuki Manabe, Akira Hiraiwa, and Toshiaki Sugimura. 2003. Unvoiced speech recognition using EMG-mime speech recognition. In *CHI'03 extended abstracts on Human factors in computing systems*. 794–795.

[44] Tania Marur, Yakup Tuna, and Selman Demirci. 2014. Facial anatomy. *Clinics in dermatology* 32, 1 (2014), 14–23.

[45] Creative Materials. 2024. Conductive Ink. https://www.creativematerials.com/applications/medical-electrodes/

[46] Creative Materials. 2024. Electrically conductive medical electrode ink. https://server.creativematerials.com/datasheets/DS_113_09.pdf

[47] Maranda McBride, Phuong Tran, and Tomasz Letowski. 2008. Head mapping: Search for an optimum bone microphone placement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52. SAGE Publications Sage CA: Los Angeles, CA, 503–507.

[48] Geoffrey S Meltzner, James T Heaton, Yunbin Deng, Gianluca De Luca, Serge H Roy, and Joshua C Kline. 2018. Development of sEMG sensors and algorithms for silent speech recognition. *Journal of neural engineering* 15, 4 (2018), 046031.

[49] 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.* 269–282.

[50] Shuo Niu, Li Liu, and D Scott McCrickard. 2019. Tongue-able interfaces: Prototyping and evaluating camera based tongue gesture input system. *Smart Health* 11 (2019), 16–28.

[51] The University of Reading. 2021. The production of speech sounds. http://www.personal.rdg.ac.uk/~llsroach/phon2/artic-basics.htm

[52] Munir Oudah, Ali Al-Naji, and Javaan Chahl. 2020. Hand gesture recognition based on computer vision: a review of techniques. *journal of Imaging* 6, 8 (2020), 73.

[53] Laxmi Pandey and Ahmed Sabbir Arif. 2021. *LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445565

[54] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. *Acceptability of Speech and Silent Speech Input Methods in Private and Public.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445430

[55] Meenakshi Panwar. 2012. Hand gesture recognition based on shape parameters. In *2012 international conference on computing, communication and applications.* IEEE, 1–6.

[56] Yuru Pei, Tae-Kyun Kim, and Hongbin Zha. 2013. Unsupervised random forest manifold alignment for lipreading. In *Proceedings of the IEEE International Conference on Computer Vision.* 129–136.

[57] 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.

[58] BrainVision Recorder. 2024. BrainVision Recorder. https://www.brainproducts.com/downloads/recorder/

[59] Jun Rekimoto and Yu Nishimura. 2021. Derma: Silent Speech Interaction Using Transcutaneous Motion Sensing. In *Augmented Humans Conference 2021* (Rovaniemi, Finland) *(AHs'21)*. Association for Computing Machinery, New York, NY, USA, 91–100. https://doi.org/10.1145/3458709.3458941

[60] Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James Landay. 2016. Speech is 3x faster than typing for English and Mandarin text entry on mobile devices. *arXiv preprint arXiv:1608.07323* (2016).

[61] 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* (Seattle, Washington) *(ISWC '14)*. Association for Computing Machinery, New York, NY, USA, 47–54. https://doi.org/10.1145/2634317.2634322

[62] Muhammad Sajid, Tamoor Shafique, Imran Riaz, Muhammad Imran, Mirza Jabbar Aziz Baig, Shahbaz Baig, and Sohaib Manzoor. 2018. Facial asymmetry-based anthropometric differences between gender and ethnicity. *Symmetry* 10, 7 (2018), 232.

[63] Khairul Khaizi Mohd Shariff, Auni Nadiah Yusni, Mohd Adli Md Ali, Megat Syahirul Amin Megat Ali, Megat Zuhairy Megat Tajuddin, and MAA Younis. 2022. Cw radar based silent speech interface using CNN. In *2022 IEEE Symposium on Wireless Technology & Applications (ISWTA).* IEEE, 76–81.

[64] SMOOTH-ON. 2024. FlexFoam-iT!™ 6 Pillow Soft. https://www.smooth-on.com/products/flexfoam-it-6/

[65] SMOOTH-ON. 2024. FlexFoam-iT!™ 6 Pillow Soft Product Review. https://www.smooth-on.com/tb/files/FLEXFOAM-IT_SERIES.pdf

[66] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 6447–6456.

[67] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 140 (sep 2022), 26 pages. https://doi.org/10.1145/3550281

[68] 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.* 581–593.

[69] 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.* 161–169.

[70] J. Tan, C. Nguyen, and X. Wang. 2017. SilentTalk: Lip reading through ultrasonic sensing on mobile phones. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications.* 1–9. https://doi.org/10.1109/INFOCOM.2017.8057099

[71] Michal Teplan et al. 2002. Fundamentals of EEG measurement. *Measurement science review* 2, 2 (2002), 1–11.

[72] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni. 2016. We Can Hear You with Wi-Fi! *IEEE Transactions on Mobile Computing* 15, 11 (2016), 2907–2920. https://doi.org/10.1109/TMC.2016.2517630

[73] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I. Hong, Carmel Majidi, and Swarun Kumar. 2019. RFID Tattoo: A Wireless Platform for Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 155 (Dec. 2019), 24 pages. https://doi.org/10.1145/3369812

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

[75] Wai Chee Yau, Sridhar Poosapadi Arjunan, and Dinesh Kant Kumar. 2008. Classification of voiceless speech using facial muscle activity and vision based techniques. In *TENCON 2008-2008 IEEE Region 10 Conference.* IEEE, 1–6.

[76] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling Word and Sentence-level Lip Interaction for Smart Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–28.

[77] Ruidong Zhang, Hao Chen, Devansh Agarwal, Richard Jin, Ke Li, François Guimbretière, and Cheng Zhang. 2023. HPSpeech: Silent Speech Interface for Commodity Headphones. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers.* 60–65.

[78] 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.* 1–18.

[79] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing acoustic-based imaging for issuing contact-free silent speech commands. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–26.