GazeTrak: Exploring Acoustic-based Eye Tracking on a Glass Frame

Ke Li
Cornell University
Ithaca, USA
kl975@cornell.edu

Ruidong Zhang
Cornell University
Ithaca, USA
rz379@cornell.edu

Boao Chen
Cornell University
Ithaca, USA
bc526@cornell.edu

Siyuan Chen
Cornell University
Ithaca, USA
sc2489@cornell.edu

Sicheng Yin
University of Edinburgh
Edinburgh, United Kingdom
yinsicheng1999@outlook.com

Saif Mahmud
Cornell University
Ithaca, USA
sm2446@cornell.edu

Qikang Liang
Cornell University
Ithaca, USA
ql75@cornell.edu

François Guimbretière
Cornell University
Ithaca, USA
fvg3@cornell.edu

Cheng Zhang
Cornell University
Ithaca, USA
chengzhang@cornell.edu

ABSTRACT

In this paper, we present GazeTrak, the first acoustic-based eye tracking system on glasses. Our system only needs one speaker and four microphones attached to each side of the glasses. These acoustic sensors capture the formations of the eyeballs and the surrounding areas by emitting encoded inaudible sound towards eyeballs and receiving the reflected signals. These reflected signals are further processed to calculate the echo profiles, which are fed to a customized deep learning pipeline to continuously infer the gaze position. In a user study with 20 participants, GazeTrak achieves an accuracy of 3.6° within the same remounting session and 4.9° across different sessions with a refreshing rate of 83.3 Hz and a power signature of 287.9 mW. Furthermore, we report the performance of our gaze tracking system fully implemented on an MCU with a low-power CNN accelerator (MAX78002). In this configuration, the system runs at up to 83.3 Hz and has a total power signature of 95.4 mW with a 30 Hz FPS.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile devices**; • **Hardware** → *Power and energy*.

KEYWORDS

Eye Tracking, Acoustic Sensing, Smart Glasses, Low-power

ACM Reference Format:
Ke Li, Ruidong Zhang, Boao Chen, Siyuan Chen, Sicheng Yin, Saif Mahmud, Qikang Liang, François Guimbretière, and Cheng Zhang. 2024. GazeTrak: Exploring Acoustic-based Eye Tracking on a Glass Frame. In *The 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '24), September 30-October 4, 2024, Washington D.C., DC, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3636534.3649376

1 INTRODUCTION

Currently, state-of-the-art eye tracking technologies utilize cameras to capture gaze points. However, cameras-based eye tracking solutions are known to have a relatively high power signature, which may not work well for smart glasses with a relatively small battery capacity. For instance, Tobii Pro Glass 3 [2], which is considered as one of the best eye tracking glasses, can only last for 1.75 hours with an extended battery capacity of 3400 mAh. When using the battery of a Google Glass (570 mAh), this eye tracking system can only last 18 minutes. The limited tracking time has hindered its ability to collect gaze point data in everyday life, which can be highly informative for many applications, such as, monitoring users' mental or physical health conditions [49, 59], gaze-based input, and attention and interest analysis [15].

ACM MobiCom '24, November 18-22, 2024, Washington D.C., DC, USA

ACM ISBN 979-8-4007-0489-5/24/09…$15.00
https://doi.org/10.1145/3636534.3649376

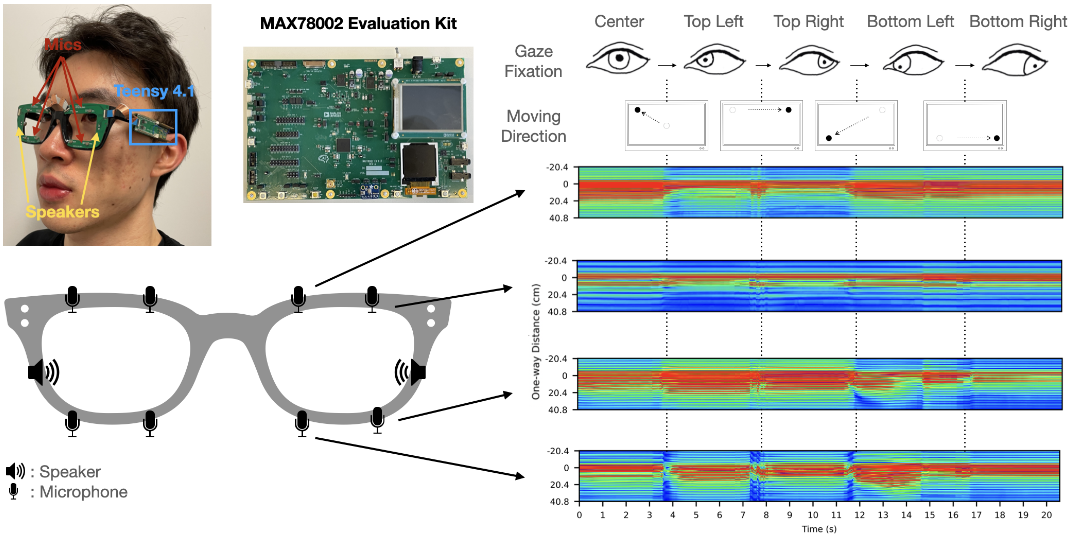


Figure 1: Echo Profiles of Different Microphones when Moving Gaze to Different Regions of The Screen.

To overcome this challenge, we introduce GazeTrak, which explores utilizing acoustic sensing (known for relatively low power, lightweight, and affordable) to continuously track gaze points on a glass frame. Its sensing principle is based on the fact that eyeballs are not perfectly spherical and rotating them would expose different shapes and stretch the skin around them with unique formations. This can provide highly valuable information for inferring gaze points. GazeTrak uses one speaker and four microphones on each side of the glass frame. The speaker emits frequency-modulated continuous-wave (FMCW) acoustic signals with the frequency above 18 kHz towards the eyeballs. The microphones capture the signals reflected by the eyeballs and their surrounding areas, which are used to process and calculate the echo profiles. These echo profiles are fed to a customized deep learning algorithm based on ResNet-18 to predict the gaze point.

We conducted two rounds of user studies to evaluate the performance of GazeTrak. During the studies, each participant was asked to look at and follow the instruction points on the screen. In the first round of the study, 12 participants evaluated our first hardware prototype, where the microphones and speakers were glued on a glass frame. The average cross-session tracking accuracy was 4.9°. It confirmed the optimal settings of the sensing system, which helped us design the final prototype. The final prototype (as shown in Fig. 1) features a more compact form factor, significantly lowers signal strength, and improves environmental sustainability as it can be attached to different glasses. To ensure consistent performance between the two prototypes, we conducted a second round of study with 10 participants, including some new participants, evaluating the final prototype. The final prototype achieved an average tracking accuracy of 4.9° for cross-session scenarios and 3.6° for in-session scenarios with a refreshing rate of 83.3 Hz. We made a demo video[1] to demonstrate the tracking performance and real-world applications of our system.

Although the current accuracy of our system was worse than commercial eye trackers such as Tobii Pro Glasses 3 [2] and Pupil Labs Glasses [28], it is still comparable to some webcam-based eye-tracking systems [21, 54]. Furthermore, due to the low-power feature of acoustic sensors, GazeTrak, including the data collection system, has a relatively low power signature of 287.9 mW. Compared to camera-based wearable eye tracking systems, our proposed system reduces the power consumption by over 95%. If using a battery with the capacity similar to Tobii Pro Glasses 3, our system can extend the usage time from 1.75 hours to 38.5 hours. It can even last 6.4 hours on the battery of normal smart glasses, such as Google Glass. The power signature of our system can be further improved using a recently introduced micro-controller with a low-power CNN accelerator (MAX78002). Hence, we implemented our gaze tracking pipeline fully on MAX78002. With the refresh rate set as 30 Hz, the power consumption of the whole system including the data preprocessing and model inference is measured as 95.4 mW.

In summary, the contributions of our paper are as follows:

- We designed and implemented the first acoustic-based continuous eye tracking system on glasses.
- A user study with 20 participants showed an average cross-session accuracy of 4.9° with a refreshing rate up to 83.3 Hz and a power signature of 287.9 mW.

[1] https://youtu.be/XvNLNkfQY7Q

- The performance of the system remained robust under different noisy environments and with different styles of glass frames.
- A real-time pipeline was implemented on MAX78002 to make inferences on the board with a power consumption of 95.4 mW at 30 Hz.

2 RELATED WORK

In this section, we introduce the prior webcam-based, non-wearable, and wearable eye tracking systems.

2.1 Webcam-based Eye Tracking Systems

Webcams have been widely used to implement eye tracking technologies because of its ubiquity on computers and its advantage of low-cost. Researchers have done many work to explore the potential of webcams for eye tracking [47, 48, 53, 63]. Currently, there are plenty of webcam-based eye tracking platforms that are available online, such as RealEye.io [54], GazeRecorder [19], and WebGazer.js [21].

These webcam-based eye tracking platforms provide affordable solutions for eye tracking with acceptable tracking performance for everyday users. However, the position of webcams are usually fixed and they have a relatively low resolution. Therefore, their performance can be more easily impacted by factors like lighting conditions, occlusions, camera orientations, etc.

2.2 Other Non-wearable Eye Tracking Technologies

In order to provide more accurate and reliable solutions to eye tracking and make them more applicable to assorted scenarios, researchers have put lots of efforts in implementing other non-wearable eye tracking technologies other than webcam-based systems, most of which are based on cameras with a higher resolution than webcams. Frontal camera-based eye tracking technologies based on computer vision techniques can take full advantage of the whole facial information of the user for eye tracking, which usually leads to high tracking accuracy. Different kinds of cameras have been used in tracking eye movements, such as RGB cameras [3], infrared (IR) cameras [45], and thermal cameras [60]. Beyond using just one camera, many technologies adopted multiple cameras in their eye tracking systems in order to improve the performance in different perspectives including providing larger tracking coverage [4], allowing for user motion [22] and tracking eye movements of multiple users [37]. Because of the reliable tracking performance and reasonable calibration time needed, frontal camera-based eye tracking technologies have been well commercialized, among which Tobii Pro Fusion [1] is one of the best desktop eye trackers because it only requires seconds of calibration process for new users and can provide a tracking accuracy as low as 0.3° in optimal conditions. As a result, this product has been used as reference in many research projects.

The aforementioned frontal camera-based technologies are mostly located at fixed positions and do not work well while users move to another position or are walking around. In order to allow some mobility for users while they are using the eye tracking technologies, many researchers investigated utilizing the cameras on mobile devices to track eye movements, such as mobile phones [25, 31, 68] or tablets [7, 25]. However, these eye tracking technologies based on mobile devices still require users to hold the mobile devices in front of their face all the time and cannot provide completely hands-free and motion-free experiences for users.

2.3 Wearable Eye Tracking Technologies

To overcome the challenges that non-wearable eye tracking technologies face as described in the last two subsections, many wearable eye tracking technologies based on cameras [12, 24, 29, 40, 41, 43, 52, 67], optical sensors [8, 35, 36, 57, 58], acoustic sensors [20], magnetic sensors [62], Electrooculography (EOG) sensors [9, 10], or inertial measurement units (IMU) [29] have been deployed on different kinds of wearables including glasses [8, 12, 20, 36, 40, 41, 52, 57, 58, 67], goggles [9, 10], hat [29], and head-mounted devices [24, 35, 43, 62]. Among all these wearable eye tracking technologies, camera-based ones usually outperform others in terms of tracking performance and do not require lots of calibration data from new users. Many wearable eye trackers using cameras, especially on glasses, have become commercial and can be used as a reliable way to track eye movements continuously, such as Tobii Pro Glasses 3 [2], Pupil Labs (Invisible, Core, VR/AR add-ons) [28], Dikablis Glasses 3 [16], and SMI Eye Tracking Glasses [23]. With these technologies, various novel gaze-based applications have been enabled, including detection of eye contacts [64], interaction with devices [26, 39, 46], and monitoring mental health [32, 59].

Despite of the promising tracking performance, current solutions to wearable eye tracking systems still have some limitations. First of all, many eye tracking systems above can only recognize discrete gestures [9, 10, 57, 58, 67], limiting their performance in applications that need continuous tracking of the eyes. Camera-based wearable eye trackers can provide high accuracy in continuous eye tracking, but cameras are usually power-hungry, which makes them relatively impractical while deployed in wearables that need to be worn in everyday settings. To address this issue, Mayberry et al. [40, 41] proposed low-power solutions to tracking gaze positions with cameras on glasses while maintaining promising accuracies. Despite of the impressive performance, changing lighting conditions can still be a problem for these

Table 1: GazeTrak and Other Continuous Eye Tracking Techniques. The power of GazeTrak (Teensy 4.1) does not include data preprocessing and deep learning inference. The reported accuracy is tested within the same session without users remounting the device. Both weight and cost include the recording unit. NS = Not Specified.

Reference	Form Factor	Sensors	Power	Accuracy	Refresh Rate	Weight	Cost
Cho et al. [12]	Glasses	Cameras	>7W	0.79°	NS	NS	NS
Ryan et al. [52]	Glasses	Cameras	>1.6W	2°	NS	NS	~$700
iShadow [40]	Glasses	Cameras	0.07W	3°	30 Hz	NS	NS
CIDER [41]	Glasses	Cameras	0.032W	0.6°	250 Hz	NS	NS
Pupil Labs Glasses [28]	Glasses	Cameras	8.6W	0.6°	30/60/120 Hz	202.75g	$2,849
Tobii Pro Glasses 3 [2]	Glasses	Cameras	10.7W	0.6°	50/100 Hz	388.5g	$16,055
SMI Glasses [23]	Glasses	Cameras	NS	0.5°	60/120 Hz	NS	$41,000
Li et al. [36]	Glasses	NIR LED & Photodiodes	395μW	<2°	120 Hz	<25g	NS
Li et al. [35]	Head-mounted	Photodiodes	791μW	6.3°	10 Hz	NS	NS
GazeRecorder [19]	Webcam	Camera	/	1.05°	30 Hz	/	$500/month
WebGazer.js [21]	Webcam	Camera	/	4.17°	NS	/	Free
RealEye.io [54]	Webcam	Camera	/	~5°	60 Hz	/	$600/month
GazeTrak (Teensy 4.1)	**Glasses**	**Acoustic Sensors**	**0.288W**	**3.6°**	**83.3 Hz**	**44.2g**	**~$75**
GazeTrak (MAX78002)	**Glasses**	**Acoustic Sensors**	**0.095W**	**4.2°**	**30 Hz**	/	/

camera-based systems, as the performance became worse in an outdoor setting [41]. Besides, commercial eye trackers are usually expensive and do not provide open-source software for users, preventing them from being easily accessed and adapted by general users.

Recently, Li et al. [36] proposed a low-cost and battery-free solution to continuous eye tracking using near infrared emitters and receivers on glasses. It achieves competitive performance but they stated in the paper that this system can be impacted by direct sunlight and glasses movement, i.e. the remounting of the glasses. Besides, this work tracks the position and size of the pupil so we cannot directly compare it to our system. After conversion, its tracking accuracy of gaze positions is smaller than 2° in angular error. Another system using similar technology from the same group [35] tracks gaze positions with an accuracy of 6.3°, worse than the tracking accuracy of our system at 4.9°. Golard et al. [20] conducted a modeling and empirical study to prove that ultrasound can provide a low-power, fast and light-insensitive alternative for camera-based eye tracking technologies. However, it was evaluated on a physical 3D model of a human eye and used time-of-flight estimated from acoustic signals and not clear how it can apply on a real user.

To the best of our knowledge, GazeTrak is the first wearable sensing technology based on active acoustic sensing that can track gaze points continuously. We summarized and compared GazeTrak with some aforementioned wearable and webcam-based eye tracking techniques that can continuously track gaze positions in Tab. 1. These techniques are those that are most related to our system. Please note that commercial eye tracking wearables [2, 23, 28] usually have camera(s) recording the video of the environment as well so we can only roughly compare them to our device in terms of power and weight.

3 PRINCIPLE AND ALGORITHMS

Active acoustic sensing is based on affordable sensors (speakers and microphones), the sizes of which are relatively small. Previous research work has proved that it is able to provide enough information to track subtle skin deformations such as facial expressions [18, 34]. In this section, we discuss how this approach can be adapted to eye tracking.

3.1 FMCW-based Active Acoustic Sensing

In order to capture the formation around eyeballs, we use FMCW-based acoustic sensing, which has been widely proven effective to estimate distance and movements from complex environments [42, 61].

3.1.1 Encoded FMCW Signals. While customizing the FMCW signals for our system, three main features are taken into account: 1) *Operating frequency range*: The device is expected to be worn by users for a long period of time in their everyday lives. As a result, the FMCW signals need to be transmitted in the inaudible frequency range. Besides, to ensure the encoded signals are minimally impacted by the noise in the environment, the operating frequency range we pick should also be uncommon in daily settings; 2) *Sampling rate*: To achieve a reasonable spatial and temporal resolution of tracking eye movements, the sampling rate of FMCW signals must be high enough; 3) *Gain*: As power signature increases with the signal gain, the signal gain should be properly determined to balance signal strength and power consumption.

Considering all the factors above, we set the operating frequency range of the FMCW signals that we emit in the

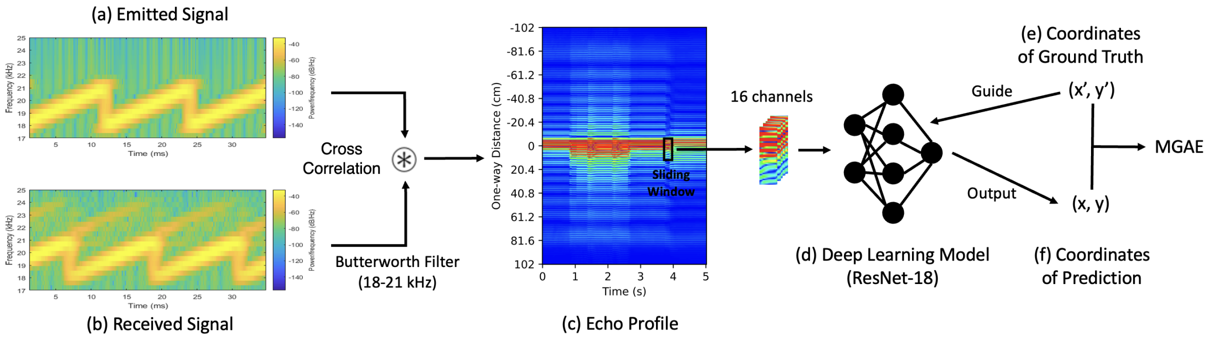


Figure 2: Overview of the GazeTrak System: Use the Speaker on the Right Side (18-21 kHz) for Illustration.

GazeTrak system above 18 kHz, because this range is near-inaudible and uncommon in the sounds generated by normal human activities. Because both eyes contain information while moving, we placed one speaker on each side of the glass frame. We set the speaker on the right side to operate at 18-21 kHz while the one on the left side operates at 21.5-24.5 kHz to make sure they do not interfere with each other. To guarantee that the system works reliably in these frequency ranges, we set the ADC sampling rate as 50 kHz with the frame length of FMCW signals as 600 samples. This gives the system a refresh rate of eye tracking at 83.3 Hz (50000 samples/s ÷ 600 samples). We believe a refresh rate of 83.3 Hz is sufficient to provide continuous gaze tracking since the frame rate of most videos are 30 Hz or 60 Hz. Lastly, the gain was experimentally adjusted to make sure that the signal does not saturate the microphones while the power consumption is relatively low.

3.1.2 Acoustic Patterns for Continuous Eye Tracking. After receiving the reflected FMCW signals, we first apply a Butterworth band-pass filter with a cut-off frequency range of 18-21 kHz or 21.5-24.5 kHz on the signal to remove the signals in the frequency range that we are not interested in. It also helps protect the privacy of users because we remove the audible range of the signals. Then we further process the filtered signal to obtain unique acoustic patterns. According to prior research work [30, 33, 34, 38, 55, 61, 65, 66], *Echo Profile* provides an accurate depiction of the status and movements of the reflecting objects in the environment. As a result, in this paper, we also use echo profiles as the acoustic patterns that our system monitors. As shown in Fig. 2 (a)-(c), echo profile is obtained by continuously calculating the cross-correlation between the received signals and transmitted signals. Fig. 1 demonstrates that different eye fixations and movements are correlated with different patterns in echo profiles. Based on these observations above, we believe that our GazeTrak system utilizing FMCW-based active acoustic sensing is able to track eye movements continuously with high accuracy.

3.2 Machine Learning Algorithms

3.2.1 Ground Truth Acquisition without using Eye Trackers. A professional eye-tracker (e.g., Tobii Pro Fusion) can provide highly accurate ground truth, but it is expensive. If our system needs a professional eye-tracker to train the system, it will make our eye-tracking system less accessible.

Therefore, we developed a new ground truth acquisition and calibration system that only needs a program running on a laptop. The program generates instruction points on the screen as the ground truth. During data collection, the users only need to look at and follow the movements of the instruction points. These ground truth data along with the echo profiles are fed into the machine learning model for training. This method is generally applicable on any device with a screen. For details about how the instruction points are generated, please refer to Sec. 5. To better compare our system with commercial eye trackers, we also use a Tobii Pro Fusion (120 Hz) [1] to record the eye movements to demonstrate the effectiveness of our training methods.

3.2.2 Deep Learning Model. We developed a customized deep-learning pipeline to learn the echo profiles calculated on the received signals. Because in the echo profiles (See Fig. 1), the temporal information has been converted to the spatial information on an image, we decided to use ResNet-18 as the encoder of our deep learning model because CNN networks are known to be good at extracting features from images. Then a fully-connected network is used as a decoder to predict gaze positions based on the features extracted from the images.

Because of the limited distance between the sensors on the glasses and the eyes, we are only interested in a certain range of the echo profiles (Fig. 2 (c)). As a result, we crop the echo profiles of each channel to get the center 70 pixels (23.8 cm) vertically. Then we randomly select 60 consecutive pixels

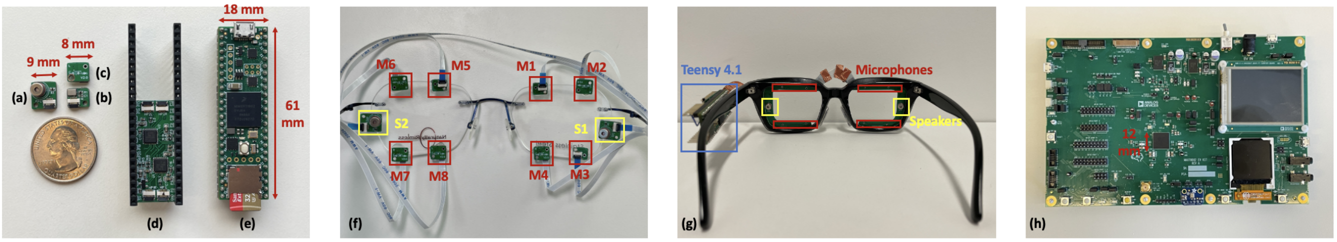


Figure 3: Hardware and Form Factor for GazeTrak: (a) Speaker board; (b) Microphone board (front view); (c) Microphone board (back view); (d) Customized PCB board for the audio chip NXP SGTL5000; (e) Teensy 4.1; (f) Glasses form factor with speakers and microphones attached (M1-8: microphones, S1-2: speakers); (g) Attachable and more compact prototype; (h) MAX78002 Evaluation Kit.

(20.4 cm) out of these 70 pixels for data augmentation purpose to make sure the system will not be severely impacted by the vertical shifting caused by remounting the device. To continuously track the gaze positions, we apply a sliding window of 0.3 seconds on the echo profiles. As a result, the dimension of the echo profile that we input into the deep learning model for one channel is 26 (0.3 s × 50000 Hz ÷ 600 samples + 1) × 60 (pixels). Because we use 2 speakers and 8 microphones in our system, which will be illustrated in Subsec. 4.2, we crop out the same dimension of echo profiles for all 2 × 8 = 16 channels, making the dimension of the input vector to the deep learning model as 26 × 60 × 16.

We use the instruction points as the labels (see Subsubsec. 3.2.1) and the mean squared error (MSE) as the loss function. We chose Adam optimizer and set the learning rate as 0.01. The model is trained for 30 epochs to get the estimation of the two gaze coordinates (x, y).

3.2.3 *Evaluation Metrics.* The prediction of our system is the coordinate (x, y) of our estimated gaze position on the screen in pixels. To evaluate the accuracy of GazeTrak, we adopted the accuracy defined in COGAIN eye tracker accuracy terms and definitions [11]. The evaluation metric we use in our system is the mean gaze angular error (MGAE) between the coordinate of our prediction (x, y) and that of the ground truth (x', y'). To calculate MGAE in degrees from the coordinates, we first need to get the angular error θ between the prediction and the ground truth of each data point. θ can be calculated using the law of cosines in a triangle as follows:

$$\theta = \arccos\left(\frac{d_{eg}^2 + d_{ep}^2 - d_{gp}^2}{2 \times d_{eg} \times d_{ep}}\right) \times 180 \div \pi \qquad (1)$$

where d_{eg}, d_{ep} and d_{gp} are the distance between user's eyes and ground truth, the distance between user's eyes and prediction, and the distance between ground truth and prediction respectively. MGAE is obtained by averaging θ over all the data points in the testing dataset.

4 DESIGN AND IMPLEMENTATION

4.1 Hardware Design

In order to implement the FMCW-based active acoustic sensing technique mentioned in the section above, we chose Teensy 4.1 [51] as the micro-controller to provide reliable FMCW signal generation and receiving in multiple channels. We designed a PCB board to support two SGTL5000 chips which are the same as the one on the Teensy audio shield [50]. With this customized PCB board plugged onto Teensy, it can support as many as 8 microphones and 2 speakers. We chose the speaker called OWR-05049T-38D [14] and the MEMS microphone called ICS-43434 [56] to support signal transmission and reception. We also built customized PCB boards for the speaker and the microphone to make them as small as possible. We used the Inter-IC Sound (I2S) buses on the Teensy 4.1 to transmit data between the Teensy 4.1 and the SGTL5000 chips, speakers and microphones. The collected data is stored in the SD card on Teensy 4.1. Fig. 3 (a)-(e) show these components.

4.2 Form Factor Design

We designed the first form factor using a commodity glass frame. We glued 1 speaker and 4 microphones to each inner side of a pair of light-weight glasses. The speakers and microphones are symmetrically placed on the glasses, as shown in Fig. 3 (f).

Based on the experience we learned during the iteration process, there are three key factors we took into consideration while designing the final form factor of GazeTrak: 1) *Type of glass frame*: We started designing the form factor with a large glass frame because we believe it has more room for us to place sensors. However, the larger the glass frame is, the easier it will be for the frame to touch the skin, blocking the signal transmission and reception. As a result, we finally picked a relatively small glass frame with a nose pad that can support the glass frame to a higher position. Besides, the light-weight glasses minimize the pressure attached on the user’s nose, making it more comfortable to wear; 2) *Sensor position*: The speakers and microphones on two sides are symmetric because we believe the movements of two eyes are usually synchronized. On each side, we place the speaker on the frame of the glasses next to the outer canthi because it is easier for the speakers to touch the skin if they are placed above the cheekbones or next to the eyebrows, considering their height. The microphones are scattered on the frame as far away from each other as possible to capture more information by receiving signals travelling in different paths. The sensors are attached as far away from the center of the lenses as possible in order to avoid blocking the view of the user; 3) *Stability*: We found that the stability of the device severely impacts the performance of our system especially when users need to remount the device frequently. The anti-slippery nose pad prevents the glasses from sliding down the user’s nose. Furthermore, we added two ear loops at the end of the legs of the glass frame. They greatly helps fix the glasses position from behind ears and improves the performance of the system. Finally, we made the form factor as shown in Fig. 3 (f).

4.3 Final Hardware Prototype

The prototype above is suitable for initial testing and comparison of different configurations. However, once the design of the prototype is finalized, we aim to create a more compact and less obtrusive form factor that is suitable for everyday use by users. To achieve this, we have designed two PCB boards, each containing one speaker and four microphones onboard, which can be attached to one side of the glasses. We have also deployed the Teensy 4.1 and the PCB board with SGTL5000 chips directly onto one leg of the glasses. To connect the micro-controller and the customized PCB boards, we have used flexible printed circuit (FPC) cables. The system has an interface that allows it to be powered by a Li-Po battery. The compact prototype is shown in Fig. 3 (g), and Fig. 1 shows a user wearing the prototype. We believe that this prototype can be easily adapted and attached to different types of glasses.

We have measured the weight of the prototype, and it carries a total weight of 44.2 grams, including the glasses, Teensy 4.1, PCB boards, and the Li-Po battery. Compared to camera-based eye tracking glasses, our GazeTrak device is much lighter. For example, Tobii Pro Glasses 3 [2] weigh 76.5 grams for the glasses and 312 grams for the recording unit. Our device has a significant advantage over camera-based eye tracking glasses in terms of weight.

5 USER STUDY PROCEDURE

The objective of our user study is to validate the performance of GazeTrak on continuously tracking gaze points. In order to reach this goal, we carefully designed the instruction video for participants’ gaze to follow. Basically, on the white screen, there would be one red dot moving around and we asked participants to stare at the point and follow it with their eyes. We divided the screen into 100 regions. For each data point, the instruction point appeared at a random position within one random region. The instruction point would move quickly to that random position and stay static at that position for a certain period of time because we mainly would like to test how GazeTrak performs to track the fixation of participants.

We recruited 20 participants (10 females and 10 males, 22 years old on average). Note that some participants participated in the study for multiple times to test different settings. The study was conducted in an experiment room on a university campus. During the study, the participants sit on a chair and put on the glasses form factor with our GazeTrak system. For each participant, we produced 12 sessions of instruction points. During the interval between sessions, participants were instructed to remove the device, place it on the table, and then put it back on. This step was taken to demonstrate that our system continued to function correctly even after the device was remounted. In each session, the instruction point moved to all the 100 pre-defined regions in a random order. The duration for which the instruction point remained at each position varied from 0.5 to 3.5 seconds, with an average of 2 seconds. As a result, the average length of each instruction session was 200 seconds. Before each session, there was a 15-second calibration process with the instruction point moving to the four corners of the screen and the center of the screen.

The full study took no more than 1.5 hours for each participant, during which we collected approximately 40 minutes of data (200 seconds × 12 sessions). Upon completing the study tasks, the participant was asked to complete a questionnaire to collect their demographic information and their feedback using this system.

6 EVALUATION RESULTS

In this section, we first evaluated the performance of GazeTrak with the initial prototype, comparing different ground truth acquisition methods, sensor configurations and amounts of training data. Then we tested our system under noisy environments and on glasses of various frame styles. Finally, we optimized the system on the final prototype and evaluated it with another study, with power consumption measured.

6.1 User-dependent Model

We first tested our system using the first prototype in Fig. 3 (f) with 12 participants. A user-dependent model was applied to train a separate model for each participant. Among 12 sessions we collected for each participant, we conducted

Table 2: Study Results for Different Mic Configurations.

Mic Configuration	M1+M5	M2+M6	M3+M7	M4+M8	Best 4 Mics (M2,4,6,8)	Best 6 Mics (M1,2,4,5,6,8)	All 8 Mics
MGAE	7.7°	7.2°	8.5°	6.9°	5.9°	5.5°	4.9°

a 6-fold cross validation to test the tracking performance of our system by using 10 sessions (33.3 minutes) of data for training and 2 sessions (6.7 minutes) of data for testing. Using the evaluation metrics defined in Subsubsec. 3.2.3, we calculated the mean gaze angular error (MGAE) in degree for all participants, and the result we obtained was 5.9°. To further improve the performance, we adopted the 15-second calibration data before each session to fine-tune the model, which resulted in an improved performance of 4.9°. It is worth mentioning that a similar calibration process is also required for commodity eye trackers (e.g., Tobii Pro). The distance between the participants' eyes and the screen center is measured to be around 60 cm so we have a field of view of 60° (the largest possible angular error) in this study. We made a demo video showing how our prediction looks like visually with this level of accuracy.

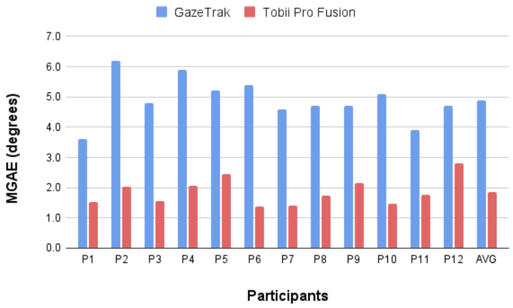


Figure 4: MGAE Distribution across Participants.

Next, we aimed to compare the impact on gaze tracking performance of using different ground truth acquisition methods: a commodity eye tracker (Tobii Pro Fusion) versus our method (using instruction points on the screen). We used the eye tracking data recorded by Tobii Pro Fusion as the ground truth to train the model, and the MGAE after fine-tuning was 4.9°. We conducted a repeated measures t-test between the results using Tobii data as the ground truth and those using instruction points as the ground truth for all 12 participants, and did not find a statistically significant difference ($p = 0.92 > 0.05$). This suggests that using instruction points on a screen monitor as the ground truth can be as effective as using Tobii data.

Apart from that, we also recorded the eye tracking accuracy of Tobii Pro Fusion itself which was reported after the calibration process of the Tobii platform. The results showed that Tobii Pro Fusion can track the gaze points with an average accuracy of 1.9° during the calibration process for all participants. We plotted the tracking performance of both GazeTrak and Tobii for all participants in Fig. 4.

6.2 Impact of Sensor Configurations

In this subsection, we evaluated the impact of the number and placement of microphones on tracking performance to determine the optimal sensor position for the best results. We assessed four different settings: 1) one microphone on each side (left and right); 2) two microphones on each side; 3) three microphones on each side and 4) all four microphones on each side. In the first setting, we compared the performance using data from four sets of microphone settings (M1+M5, M2+M6, M3+M7, and M4+M8 in Fig. 3 (f)), which is presented in Tab. 2. The findings demonstrate that the M4+M8 pair of microphones provides the best tracking performance among the four pairs tested. We conducted a one-way repeated measures ANOVA test on the results of the four settings and identified a statistically significant difference ($F(3, 44) = 6.74, p = 0.001 < 0.05$). These results indicate that microphone placement can affect gaze tracking performance, possibly due to differences in signal reflection before arriving at different microphones.

We further conducted experiments to evaluate performance using different combinations of microphones under settings 2 and 3. The results showed that the best performance was 5.9 degrees and 5.5 degrees, respectively. We also ran a one-way repeated measures ANOVA test among the results of these four settings using data from 12 participants. The results showed a statistically significant difference ($F(3, 44) = 51.61, p = 0.00001 < 0.05$). These findings suggest that our system requires four microphones on each side (eight microphones in total) to achieve the best performance.

6.3 Impact of Blinking

Blinking can introduce noise in our highly-sensitive acoustic sensing system as it can lead to relatively large movements around the eye. We conducted an evaluation to determine whether blinking affects the tracking performance of our system. For this evaluation, we selected data from three participants with the best, worst, and average tracking performance (P1, P2, P10). We removed the data where the participant blinks (about 10% of total data) based on the ground truth data obtained from Tobii Eye Tracker. We then used the processed data to retrain the user-dependent model for each participant. Our results showed that the performance did not improve after removing the blinking data. One possible reason for this result is that the blinking patterns are consistent and can be learned by the machine learning model. Therefore, our findings suggest that blinking does not significantly impact the performance of our system.

6.4 User-adaptive Model

To reduce the need of providing lots of training data for a new user, we employed a three-step process to train a user-adaptive model. Firstly, we trained a large base model using data from all participants except the one being tested. Secondly, we fine-tuned the model using the training data collected from the current participant. Notably, the user only needs to provide training data once during the initial system use. Finally, at the beginning of each session, we further fine-tuned the model using calibration data collected from the participant before testing or using the system. To determine the amount of data required to achieve competitive tracking performance, we reserved two sessions of data for testing and used varying amounts of training data from the participant to fine-tune the large model.

The results show that a new user only needs to provide six sessions of training data (approximately 20 minutes) to achieve good performance. Collecting more data does not necessarily result in better performance. Additionally, with only two or three sessions of data (approximately 6 minutes), the system can achieve a performance of 6.7° and 6.1°, respectively. If no user data is collected, the performance is 11.3°. This is likely because different people have unique head, face, and eye shapes. Therefore, to further reduce the amount of training data required from each new user, we may need to collect a significantly larger amount of training data from a more diverse set of participants.

6.5 Impact of Environment Noise

To ensure that our acoustic sensing system is resistant to different types of environmental noise, we conducted two experiments as described in this subsection.

6.5.1 Noise Injection.

In the first experiment, we recorded noises in different environments using the microphones on our glass frame. We then overlaid the noise onto the data collected in the user study to simulate different noisy environments. We recorded the noise in four different environments and measured the average noise levels using a sound level meter app called NIOSH provided by CDC [17]: 1) *street noise (70.8 dB(A))* recorded on the street near a crossroad; 2) *music noise (64.5 dB(A))* recorded while playing music on a computer; 3) *cafe noise (54.5 dB(A))* recorded in a cafe; 4) *driving noise (65.6 dB(A))* recorded while driving a vehicle. After overlaying each of these four noises, the tracking performance remained unchanged for every participant.

6.5.2 Real-world Noisy Environments.

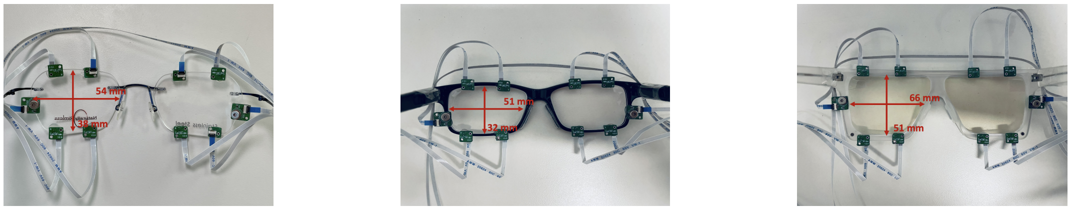
In the second experiment, we invited eight participants from the previous user study and recruited two new participants (P13 and P14) to test our device in different real-world noisy environments. Since this study required us to move to different environments, the study design differed slightly from the previous study described in Sec. 5.

In this study, we used an Apple MacBook Pro with a 13.3-inch display to play the instruction videos. We used the instruction points as the ground truth. The MacBook Pro was placed on a movable table, and participants were instructed to sit in front of the table to conduct the study. Additionally, according to Subsec. 6.4, 6 sessions of training data are sufficient to provide acceptable tracking performance. Therefore, for each participant, we collected a total of 8 sessions of data in a quiet experiment room, with 6 sessions for training and 2 sessions for testing. We then collected additional testing data under two different noisy environments. In the first environment, participants used our system while we played random music for 2 sessions. In the second environment, we collected 2 sessions of testing data at a campus cafe where staff and people were talking around during business hours. The noise levels under each environment were measured using the CDC NIOSH app: 1) *quiet room (33.8 dB(A))*; 2) *play music (64.0 dB(A))*; 3) *in the cafe (56.6 dB(A))*. This study design led to a total of 12 sessions of data collection for each participant, which is the same as the previous study.

We trained a personalized model for each participant using 6 sessions of data collected in the quiet room. Then, the 2 testing sessions collected in each scenario were used to test the performance of our system in different environments. The average gaze tracking performance of our system across 10 participants remained satisfactory at 3.8° and 4.8° under two noisy environments, playing music and in the cafe, while the performance in the quiet room was 4.6°. Overall, the average accuracy of gaze tracking did not change significantly with the presence of noise in the environment. We conducted a one-way repeated measures ANOVA test among the results of these three scenarios for all 10 participants and did not find a statistically significant difference ($F(2, 27) = 2.46, p = 0.11 > 0.05$). This again validates that our system is not easily affected by environmental noise.

6.6 Impact of Different Glass Frames

In our user study, we only tested our system on one glass frame (F1). However, we believe our GazeTrak system can be easily applied to glasses with different frame styles. In order to validate this assumption, we deployed our system on two



(a) Original Glasses (F1) **(b) Small Glasses (F2)** **(c) Large Glasses (F3)**

Figure 5: GazeTrak Deployed on Glasses with Various Frame Styles.

other pairs of glasses as shown in Fig. 5. The original glass frame in Fig. 5 (a) is frameless and relatively lightweight. In this study, we applied our system on two new glasses with different styles, size and weight. The first new glass frame (small glasses, F2) has a smaller size than F1 but a larger weight due to the frame around the lens (Fig. 5 (b)). The second new glass frame (large glasses, F3) with a frame around lens (Fig. 5 (c)) has a much larger size and weight compared to F1 and F2. To evaluate our system on these new glass frames, three participants from the original study (P1, P5 and P7) agreed to participate in this additional study. The study setups and procedures were exactly the same as the previous study described in Sec. 5.

We collected 12 sessions of data for each participant testing each glass frame. Since Subsec. 6.4 indicates that 6 sessions of training data is sufficient, we discarded the first 4 sessions for each glass frame and used the last 8 sessions to run a 4-fold cross validation in order to test the tracking performance of our system on different glasses. In this case, we can make sure that participants are familiar with the wearing of all the glass frames and eliminate the impact of some random factors. The evaluation result shows that the small glasses (F2) yielded a similar average performance to the original glasses (F1) (both at 5.3°), while the large glasses (F3) resulted in a relatively poorer average performance (at 6.1°), with a drop in performance of 15%. One possible reason for the performance difference is that the sensors on the larger glasses were much closer to the skin. Sometimes, the sensors may directly touch the skin, which could block the transmission and reception of signals, as we explained in Subsec. 4.2.

6.7 Evaluation on the Final Prototype

In the previous user studies, we evaluated GazeTrak with various configurations under different scenarios, using the initial prototype that we had developed. The results of these studies helped us confirm the prototype settings and develop an optimized system prototype, which features a more compact form factor as shown in Fig. 3 (g). In this subsection, our objective was to assess the performance and power consumption of this final prototype.

6.7.1 Gaze Tracking Accuracy. To evaluate the final prototype, we recruited 10 participants (four of whom participated in the previous study). The study design was similar to the previous study, except that we only used instruction points as the ground truth acquisition method. Each participant collected eight sessions of data (six sessions for training and two for testing). We reduced the signal strength from the speaker to 20% of the original setup, as we found that even with 2% of the original strength, the performance was similar in the pilot study. Hence, this final prototype has significantly lower signal strength and improved environmental sustainability. Additionally, we set the CPU speed of the Teensy 4.1 to 150 MHz in this study (standard speed: 600 MHz) to lower power consumption. With this setting, the system experienced a data loss rate of 0.002%, and the performance of our system was not affected by this loss, as shown in Tab. 3. Apart from the cross-session performance, we also conducted a test of the in-session tracking accuracy in which the training data and testing data were split from the same sessions without remounting the device to show the optimal performance of our system.

Table 3: Gaze Tracking Performance in MGAE with the Final Prototype.

Settings	P1	P2	P7	P10	P15	P16	P17	P18	P19	P20	Avg
Cross-session	4.4°	4.9°	5.9°	8.0°	3.6°	3.6°	3.9°	4.8°	4.7°	5.7°	**4.9°**
In-session	3.9°	3.6°	4.9°	5.3°	1.8°	2.6°	2.6°	3.6°	3.1°	4.9°	**3.6°**

As shown in Tab. 3, the mean gaze angular error (MGAE) is 4.9° for the cross-session evaluation, which is similar to the previous study. When evaluating the performance of GazeTrak within the same sessions, the accuracy improves to 3.6°. We did not add ear loops to this prototype because the legs of the glasses were wider than the ear loops we had. For most participants, the glasses fit well on their ears, but one participant (P10) reported that the glasses kept sliding down during the study, which may have affected their performance. Based on the questionnaires, no participant reported being able to hear the signal emitted from our system. We also measured the signal level from our system using the NIOSH app. We placed the phone running the app close enough to the speakers in our system and the app gave us an average signal level of 43.1 dB(A). This is below the maximum allowable daily noise recommended by CDC, which is 85 dB(A) over eight hours in the workspace [44].

6.7.2 Power Consumption. We measured the power consumption of our system with a current ranger [27]. The average current flowing through the system was measured as 88.3 mA @ 3.26 V, which gives us a power consumption of 287.9 mW. This value was tested with all 8 microphones and 2 speakers working, and with the data being written into the SD card. Our system can last up to 38.5 hours with a battery of similar capacity to Tobii Pro Glasses 3 (3400 mAh), while the working time of Tobii Pro Glasses 3 is only 1.75 hours. If applied to non-eye-tracking glasses, like Google Glass, our system can run for 6.4 hours. It is worth noting that these estimates do not include the power consumption of data preprocessing and deep learning inference running on a local server. We measured the power consumption of different components in our system (Tab. 4). Teensy 4.1 has a high base power consumption, while the sensors (speakers and microphones) consume much less power.

Table 4: Power Consumption of Different Components on Teensy 4.1. Power of data preprocessing and deep learning inference is NOT included.

Total	Speakers & Mics	SD card writing	Other operations
287.9 mW	16.4 mW	72.7 mW	198.8 mW

6.7.3 Usability. After the user study, we distributed a questionnaire to every participant to ask for feedback on our prototype. First, the participants evaluated the overall comfortableness and the weight of the prototype with a rating from 0 to 5. Across all 10 participants, the average scores they gave to these two aspects are 4.5 (std=0.7) and 4.2 (std=0.8), indicating that GazeTrak is overall comfortable to wear and easy to use. Furthermore, all 10 participants answered "No" to the question "Can you hear the sound emitted from our system?", verifying the inaudibility of the acoustic signals from the GazeTrak system.

7 INFERENCE ON MAX78002

In the previous evaluation, we recorded audio data with Teensy 4.1 first and run the signal processing and deep learning pipeline on a local server offline. To enable predictions of gaze positions in real-time on an MCU, we implemented the whole pipeline on a micro-controller with an ultra-low-power CNN accelerator (MAX78002 [13]).

7.1 ML Models

To achieve this goal, the deep learning models were trained and synthesized in advance, using the ai8x libraries [5]. We implemented two models with ai8x, which were ResNet-18 (used in the previous study) and MobileNet for comparison. Due to the hardware limit of MAX78002, we modified the models to be compatible with the chip. Specifically, for a Conv2d layer, the kernel size could only be set to 1x1 or 3x3 and the stride is fixed to [1, 1]. In addition, some convolution layers of ResNet-18 were substituted with depthwise separable convolution layers to avoid exceeding the limit of the number of parameters in the model. Furthermore, we quantized the input and the weights of the models with ai8x, which converted them all into 8-bit data format to save memory for storage and increase the speed of inference.

7.2 Data Preprocessing

Before the deep learning model, we also need to apply a band-pass filter on the received signals and perform cross-correlation between received signals and transmitted signals to obtain echo profiles as described in Subsubsec. 3.1.2. In the implementation on MAX78002, to reduce processing time, we removed the band-pass filter since all the computations are done on the MCU and transmitting private data is no longer a concern.

Then we experimented two different methods to realize the cross-correlation: (1) brute force to calculate echo profiles point by point; (2) the dot product function in the CMSIS-DSP library. Results of standard tests [6] revealed that it took the system 178.3 ms and 45.4 ms to compute one echo frame and make one inference with these two methods utilized respectively. Considering that one frame of our audio data comes every 12 ms in our system (600 samples ÷ 50000 samples/s), this processing time is too long to keep our system running in real-time with an FPS of 83.3 Hz. Finally, we explored method (3) a Conv2d layer (kernel size 1x1) with transmitted signals as the untrained weights and received signals placed along the channel axis of the input. This can increase the speed of echo profile calculation because it uses the CNN accelerator on MAX78002. We compressed the samples used for cross-correlation from 600x600 to 34x34 and the pixels of interest from 60 pixels (20.4 cm) to 30 pixels (10.2 cm) in this case to further decrease the processing time.

With this Conv2d layer added on top of the deep learning model, the model directly takes the raw audio data as input in instances with the size of 64 (34+30 samples) x 26 (frames) x 8 (microphones). This method allows the system to make one inference within 10.3 ms, which is enough for the real-time pipeline with a double-buffer method applied (DMA moves the current frame in one buffer while the CPU processes the previous frame in another buffer).

7.3 Accuracy and Refresh Rate

To validate these modifications and compression, we evaluated the in-session performance of different models with different settings using data collected with the final prototype in Subsec. 6.7 and showed the results in Tab. 5.

Table 5: Average In-session Performance across 10 Participants with Different Models and Settings.

Models	ML Libraries	Compressed?	Quantized?	MGAE
ResNet-18	pytorch	✗	✗	3.6°
	ai8x	✗	✗	4.0°
	ai8x	✓	✗	4.0°
	ai8x	✓	✓	4.2°
MobileNet	ai8x	✓	✗	4.2°
	ai8x	✓	✓	4.3°

As shown in the table, the same model trained with ai8x is slightly worse than that trained with PyTorch given the constraints of the convolution layers discussed above. Compressing the size of input data does not affect the accuracy. While MobileNet yields comparable accuracy to ResNet-18, both models suffer a slight performance drop after quantization since the precision of data is decreased.

Given the limitation of the I2S interfaces on MAX78002, to test our system in a more realistic condition, we still use Teensy 4.1 to control the speakers and microphones and transfer the received audio data to MAX78002 via the serial port. To accelerate the transmission speed, only the samples that are used for processing on MAX78002 are transferred. This generates a steady stream of audio data to MAX78002. In future, we will explore connecting microphones directly to MAX78002 using multi-channel audio protocols such as Time-division Multiplexing (TDM). Evaluation results showed that for ResNet-18 and MobileNet, MAX78002 spent 124.1 ms and 41.6 ms respectively loading the weights of the model. This is a one-time effort and can be done before running the real-time pipeline so it did not impact the refresh rate. Then it took 12 ms to load one instance and make an inference based on it in real-time for both ResNet-18 and MobileNet, giving a refresh rate of 83.3 Hz.

7.4 Power Consumption

We measured the power consumption of the MAX78002 evaluation kit while it made inferences. Tab. 6 demonstrates that MAX78002 consumes 96.9 mW and 86.0 mW respectively when making inferences with ResNet-18 and MobileNet at 83.3 Hz. The refresh rate can be reduced to 30 Hz to save power, which is enough for many applications. In this case, the power becomes 79.0 mW and 75.7 mW respectively.

If we can use MAX78002 to directly control speakers and microphones in future, we will be able to optimize the power efficiency and keep the overall power consumption of our real-time system around 95.4 mW, i.e., 79.0 mW (MAX78002 with ResNet-18 running at 30 HZ) + 16.4 mW (2 speakers and 8 microphones). One should keep in mind that this is just an estimate of the power of this real-time system and the power consumption of MAX78002 might increase if it does need to control the sensors but we do not expect it to be very high because the current power of MAX78002 already includes that of the CPU and the CNN accelerator running at full speed.

Table 6: Power Consumption of MAX78002 with Different Models Running.

Models	ResNet-18		MobileNet	
FPS (Hz)	83.3	30	83.3	30
Power (mW)	96.9	79.0	86.0	75.7

8 DISCUSSION

8.1 Evaluating Simpler Regression Models

We adopted two traditional regression models, which are linear regression (LR) and gradient boosted regression trees (GBRT), to predict gaze positions using the data collected in Subsec. 6.7 and the results showed that the average in-session tracking accuracy for these two models across 10 participants are 11.6° and 6.8° respectively. Compared to the results in Tab. 3, the traditional regression models output much worse accuracies than ResNet-18 (3.6°). We conducted an analysis of the impurity-based feature importance with GBRT, comparing the features in different channels of microphones in Tab. 7. It turns out that the channels receiving signals from 18-21 kHz (S1) are generally more important than channels receiving signals from 21.5-24.5 kHz (S2). Furthermore, the microphones that are closer to the inner corners of the eyes (M1, M4, M5 M8) are more important than those closer to the tails of the eyes (M2, M3, M6, M7).

Table 7: Feature Importance Analysis for Different Microphones using GBRT (Scaled to 0-100).

Microphones	M1	M2	M3	M4	M5	M6	M7	M8
Importance (S1/S2)	100/27	57/23	33/12	96/29	66/81	42/17	28/15	85/79

8.2 Impact of Real-world Factors

8.2.1 Head Movements. In the user study, we did not use a chin rest to fix the participants' head so they could turn their head freely. However, we believe that how head movements affect the system performance should be evaluated in more details in the future.

8.2.2 Near- and Far-sighted. In the user study, we collected participants' degrees of myopia in the questionnaires, which showed no connection to the gaze tracking performance.

8.2.3 User Speaking. One researcher evaluated our system when keeping silent and keeping talking to himself. The gaze tracking performance of the silent sessions and the talking sessions is the same at 3.9°.

8.3 Potential Applications

The goal of this paper is to demonstrate the feasibility of our new acoustic-based gaze tracking system on glasses. While our eye tracking accuracy of 4.9° is comparable to some webcam-based methods, it is lower than commercial eye trackers (1.9° in our study). Therefore, our system may not be immediately applicable to some applications requiring highly precise eye tracking. However, our system can still be used in many applications, such as interaction with interface elements like buttons in AR, that do not require very high accuracy eye trackers.

Our system can also be potentially used in tracking irregular eye movements, enabling healthcare applications for monitoring users' health conditions in everyday life. This requires monitoring the gaze movements throughout the day for analysis in everyday life, instead of just tracking their accurate gaze positions for a few hours in a controlled settings. The low-power and lightweight features of our GazeTrak system make it a good candidate solution to enabling a variety of applications that camera-based eye trackers cannot realize, by continuously understanding user gaze movements in the wild for extended periods. Furthermore, our system can alleviate the privacy concern from users as compared to camera-based methods.

8.4 Limitation and Future Work

8.4.1 Improving the Performance. There is room for further improvements of the performance of our system. For instance, we can apply calibration process on the output of the system to further enhance performance. We experimented with affine transformation and projective transformation to transform the output but they did not immediately improve the performance. According to the analysis of the error distribution of the eye tracking results, we believe this is because the error distribution is not linear in our system so we need to explore more non-linear transformation methods to improve the performance.

8.4.2 Calibration Process for Fine-tuning. Our system currently requires a 15-second calibration process before each session to fine-tune the model, which may be inconvenient for users. However, Subsec. 6.1 shows that the tracking accuracy without fine-tuning is still acceptable, at 5.9°, compared to the accuracy achieved with fine-tuning (4.9°).

8.4.3 Reducing Training Effort. Subsec. 6.4 suggests that GazeTrak achieves satisfactory performance on new users with approximately 20 minutes of training data using the user-adaptive model. This training effort can be further reduced by constructing a larger and more diverse dataset from much more participants to train the based model. Moreover, data augmentation methods, such as including simulation data to train the model, can be explored as well.

8.4.4 Towards a More Integrated System. In Sec. 7, we still used Teensy 4.1 to control the speakers and microphones, and transfer audio data to the MCU MAX78002. In future, we plan to further customize our own PCBs for MAX78002 to allow it to directly control speakers and microphones. We believe that the power consumption of our real-time system can be further reduced in this case because Teensy 4.1 with a high base power can be removed. Furthermore, we do not expect that the power consumption of MAX78002 will be significantly increased since the on-board CPU and CNN accelerator of MAX78002 were already operating at maximum speed in our current evaluation. With a solid system implementation, we plan to carry out an extensive evaluation of this more integrated system in future work to validate our speculation.

9 CONCLUSION

In this paper, we present the first acoustic-based eye tracking glasses capable of continuous gaze tracking. The study involving 20 participants confirms that our system can accurately track gaze points continuously, achieving an accuracy of 3.6° within the same session and 4.9° across different sessions. When compared to commercial camera-based eye tracking glasses such as Tobii Pro Glasses 3, our system reduces power consumption by 95%. A real-time pipeline is implemented on MAX78002 to make inferences with a power signature of 95.4 mW at 30 Hz.

ACKNOWLEDGMENTS

This work is supported by National Science Foundation (NSF) under Grant No. 2239569, NSF's Innovation Corps (I-Corps) under Grant No. 2346817, the IGNITE Innovation Acceleration Program, and the Ann S. Bowers College of Computing and Information Science at Cornell University. The authors would like to thank Prof. Susan Fussell and her lab for sharing the access to the commercial eye tracker Tobii Pro Fusion with us. We also appreciate the help of a Cornell student, Vipin Gunda, on the development of the pipeline on MAX78002.

REFERENCES

[1] Tobii AB. 2022. Tobii Pro Fusion. Retrieved Sept. 13, 2022 from https://www.tobiipro.com/product-listing/fusion/

[2] Tobii AB. 2022. Tobii Pro Glasses 3. Retrieved Sept. 13, 2022 from https://www.tobiipro.com/product-listing/tobii-pro-glasses-3/

[3] Artsiom Ablavatski, Andrey Vakunov, Ivan Grishchenko, Karthik Raveendran, and Matsvei Zhdanovich. 2020. Real-time Pupil Tracking from Monocular Video for Digital Puppetry. CoRR abs/2006.11341 (2020). arXiv:2006.11341 https://arxiv.org/abs/2006.11341

[4] Christer Ahlstrom and Tania Dukic. 2010. Comparison of Eye Tracking Systems with One and Three Cameras. In Proceedings of the International Conference on Methods and Techniques in Behavioral Research (MB). Article 3, 4 pages. https://doi.org/10.1145/1931344.1931347

[5] Analog Devices AI. 2022. MaximIntegratedAI. Retrieved Aug 23, 2023 from https://github.com/MaximIntegratedAI

[6] Analog Devices AI. 2023. MAX7800x Power Monitor and Energy Benchmarking Guide. Retrieved Aug 23, 2023 from https://github.com/MaximIntegratedAI/MaximAI_Documentation/blob/master/Guides/MAX7800x%20Power%20Monitor%20and%20Energy%20Benchmarking%20Guide.md

[7] Tanya Bafna, Per Bækgaard, and John Paulin Paulin Hansen. 2021. EyeTell: Tablet-based Calibration-free Eye-typing using Smooth-pursuit movements. In ACM Symposium on Eye Tracking Research and Applications. 1–6.

[8] Frank H Borsato and Carlos H Morimoto. 2016. Episcleral surface tracking: challenges and possibilities for using mice sensors for wearable eye tracking. In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications. 39–46.

[9] Andreas Bulling, Daniel Roggen, and Gerhard Tröster. 2008. It's in your eyes: Towards context-awareness and mobile HCI using wearable EOG goggles. In Proceedings of the 10th international conference on Ubiquitous computing. 84–93.

[10] Andreas Bulling, Daniel Roggen, and Gerhard Tröster. 2009. Wearable EOG Goggles: Eye-Based Interaction in Everyday Environments. In CHI Extended Abstracts on Human Factors in Computing Systems. 3259–3264. https://doi.org/10.1145/1520340.1520468

[11] Communication by Gaze Interaction Association. 2010. Woking copy of definitions and terminology for Eye Tracker accuracy and precision. Retrieved Sept. 13, 2022 from http://old.cogain.org/forums/eye-tracker-accuracy-and-precision-general-discussion/eye-tracker-accuracy-terms-and-definiti.html

[12] Chul Woo Cho, Ji Woo Lee, Kwang Yong Shin, Eui Chul Lee, Kang Ryoung Park, Heekyung Lee, and Jihun Cha. 2012. Gaze Detection by Wearable Eye-Tracking and NIR LED-Based Head-Tracking Device Based on SVR. Etri Journal 34, 4 (2012), 542–552.

[13] Analog Devices. 2022. MAX78002 Evaluation Kit. Retrieved Aug 23, 2023 from https://www.analog.com/media/en/technical-documentation/data-sheets/MAX78002EVKIT.pdf

[14] DigiKey. 2023. OWR-05049T-38D. Retrieved Mar 13, 2023 from https://www.digikey.com/en/products/detail/ole-wolff-electronics-inc/OWR-05049T-38D/13683703

[15] Sibo Dong, Justin Goldstein, and Grace Hui Yang. 2022. GazBy: Gaze-Based BERT Model to Incorporate Human Attention in Neural Information Retrieval. 182–192. https://doi.org/10.1145/3539813.3545129

[16] Ergoneers. 2022. Dikablis Glasses 3. Retrieved Sept. 13, 2022 from https://www.ergoneers.com/en/mobile-eye-tracker-dikablis-glasses-3/?gclid=CjwKCAiA9tyQBhAIEiwA6tdCrFd7F7xBwNa4XVP09wRHlBATh_jafRuH5ErUVdhJt5WVLK2_FdVs7RoCpK4QAvD_BwE

[17] Centers for Disease Control and Prevention (CDC). 2023. NIOSH Sound Level Meter App. Retrieved Mar 13, 2023 from https://www.cdc.gov/niosh/topics/noise/app.html

[18] Yang Gao, Yincheng Jin, Seokmin Choi, Jiyang Li, Junjie Pan, Lin Shu, Chi Zhou, and Zhanpeng Jin. 2022. SonicFace: Tracking Facial Expressions Using a Commodity Microphone Array. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5, 4, Article 156 (dec 2022), 33 pages. https://doi.org/10.1145/3494988

[19] GazeRecorder. 2022. GazeRecorder Webcam Eye Tracking. Retrieved Sept. 13, 2022 from https://gazerecorder.com/webcam-eye-tracking-accuracy/

[20] Andre Golard and Sachin S Talathi. 2021. Ultrasound for Gaze Estimation—A Modeling and Empirical Study. Sensors 21, 13 (2021), 4502.

[21] Brown HCI Group. 2021. WebGazer.js: Democratizing Webcam Eye Tracking on the Browser. Retrieved Sept. 13, 2022 from https://webgazer.cs.brown.edu/#publication

[22] Craig Hennessey and Jacob Fiset. 2012. Long range eye tracking: bringing eye tracking into the living room. In Proceedings of the Symposium on Eye Tracking Research and Applications. 249–252.

[23] iMotions. 2022. SMI Eye Tracking Glasses. Retrieved Sept. 13, 2022 from https://imotions.com/hardware/smi-eye-tracking-glasses/

[24] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication. 1151–1160.

[25] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye Tracking for Everyone. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[26] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A. Lee, and Mark Billinghurst. 2018. Pinpointing: Precise Head- and Eye-Based Target Selection for Augmented Reality. In Proceedings of the CHI Conference on Human Factors in Computing Systems. 1–14. https://doi.org/10.1145/3173574.3173655

[27] Low Power Lab. 2018. Introduction to Current Ranger. Retrieved Mar 13, 2023 from https://lowpowerlab.com/guide/currentranger/

[28] Pupil Labs. 2022. Pupil Invisible. Retrieved Sept. 13, 2022 from https://pupil-labs.com/products/

[29] Antonio Lanata, Gaetano Valenza, Alberto Greco, and Enzo Pasquale Scilingo. 2015. Robust head mounted wearable eye tracking system for dynamical calibration. Journal of Eye Movement Research 8, 5 (2015).

[30] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Deng, Ke Li, Mose Sakashita, Francois Guimbretiere, and Cheng Zhang. 2024. EchoWrist: Continuous Hand Pose Tracking and Hand-Object Interaction Recognition Using Low-Power Active Acoustic Sensing On a Wristband. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, 21 pages. https://doi.org/10.1145/3613904.3642910

[31] Yaxiong Lei. 2021. Eye Tracking Calibration on Mobile Devices. In ACM Symposium on Eye Tracking Research and Applications (ETRA Adjunct). Article 4, 4 pages. https://doi.org/10.1145/3450341.3457989

[32] Jue Li, Heng Li, Waleed Umer, Hongwei Wang, Xuejiao Xing, Shukai Zhao, and Jun Hou. 2020. Identification and classification of construction equipment operators' mental fatigue using wearable eye-tracking technology. Automation in Construction 109 (2020), 103000.

[33] Ke Li, Ruidong Zhang, Siyuan Chen, Boao Chen, Mose Sakashita, François Guimbretière, and Cheng Zhang. 2024. EyeEcho: Continuous and Low-power Facial Expression Tracking on Glasses. In Proceedings

of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, 24 pages. https://doi.org/10.1145/3613904.3642613

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

[35] Tianxing Li, Qiang Liu, and Xia Zhou. 2017. Ultra-Low Power Gaze Tracking for Virtual Reality. In Proceedings of the ACM Conference on Embedded Network Sensor Systems (SenSys). Article 25, 14 pages. https://doi.org/10.1145/3131672.3131682

[36] Tianxing Li and Xia Zhou. 2018. Battery-Free Eye Tracker on Glasses. In Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom). 67–82. https://doi.org/10.1145/3241539.3241578

[37] Bhanuka Mahanama. 2022. Multi-User Eye-Tracking. In Symposium on Eye Tracking Research and Applications (ETRA). Article 36, 3 pages. https://doi.org/10.1145/3517031.3532197

[38] Saif Mahmud, Ke Li, Guilin Hu, Hao Chen, Richard Jin, Ruidong Zhang, François Guimbretière, and Cheng Zhang. 2023. PoseSonic: 3D Upper Body Pose Estimation Through Egocentric Acoustic Sensing on Smartglasses. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 7, 3, Article 111 (sep 2023), 28 pages. https://doi.org/10.1145/3610895

[39] Mizuki Matsubara, Joachim Folz, Takumi Toyama, Marcus Liwicki, Andreas Dengel, and Koichi Kise. 2015. Extraction of read text using a wearable eye tracker for automatic video annotation. 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. 849–854.

[40] Addison Mayberry, Pan Hu, Benjamin Marlin, Christopher Salthouse, and Deepak Ganesan. 2014. IShadow: Design of a Wearable, Real-Time Mobile Gaze Tracker. In Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '14). 82–94. https://doi.org/10.1145/2594368.2594388

[41] Addison Mayberry, Yamin Tun, Pan Hu, Duncan Smith-Freedman, Deepak Ganesan, Benjamin M. Marlin, and Christopher Salthouse. 2015. CIDER: Enabling Robustness-Power Tradeoffs on a Computational Eyeglass. In Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (MobiCom '15). 400–412. https://doi.org/10.1145/2789168.2790096

[42] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. 2015. Contactless sleep apnea detection on smartphones. In Proceedings of the 13th annual international conference on mobile systems, applications, and services. 45–57.

[43] Basilio Noris, Jean-Baptiste Keller, and Aude Billard. 2011. A wearable gaze tracking system for children in unconstrained environments. Computer Vision and Image Understanding 115, 4 (2011), 476–486.

[44] U.S. Department of Health and Human Services. 1998. Criteria for a recommended standard: occupational noise exposure. DHHS (NIOSH) Publication No. 98–126 (1998). https://www.cdc.gov/niosh/docs/98-126/

[45] Takehiko Ohno, Naoki Mukawa, and Shinjiro Kawato. 2003. Just blink your eyes: A head-free gaze tracking system. In CHI'03 extended abstracts on Human factors in computing systems. 950–957.

[46] Lucas Paletta, Helmut Neuschmied, Michael Schwarz, Gerald Lodron, Martin Pszeida, Stefan Ladstätter, and Patrick Luley. 2014. Smartphone Eye Tracking Toolbox: Accurate Gaze Recovery on Mobile Displays. In Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA). 367–68. https://doi.org/10.1145/2578153.2628813

[47] Alexandra Papoutsaki. 2015. Scalable Webcam Eye Tracking by Learning from User Interactions. In Proceedings of the Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA). 219–222. https://doi.org/10.1145/2702613.2702627

[48] Alexandra Papoutsaki, James Laskey, and Jeff Huang. 2017. Searchgazer: Webcam eye tracking for remote studies of web search. In Proceedings of the 2017 conference on conference human information interaction and retrieval. 17–26.

[49] Lorenzo Piccardi, Basilio Noris, Olivier Barbey, Aude Billard, Giuseppina Schiavone, Flavio Keller, and Claes von Hofsten. 2007. WearCam: A head mounted wireless camera for monitoring gaze attention and for the diagnosis of developmental disorders in young children. In IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). 594–598. https://doi.org/10.1109/ROMAN.2007.4415154

[50] PJRC. 2023. Audio Adaptor Boards for Teensy 3.x and Teensy 4.x. Retrieved Mar 13, 2023 from https://www.pjrc.com/store/teensy3_audio.html

[51] PJRC. 2023. Teensy 4.1 Development Board. Retrieved Mar 13, 2023 from https://www.pjrc.com/store/teensy41.html

[52] Wayne J Ryan, Andrew T Duchowski, and Stan T Birchfield. 2008. Limbus/pupil switching for wearable eye tracking under variable lighting conditions. In Proceedings of the 2008 symposium on Eye tracking research & applications. 61–64.

[53] Shreshth Saxena, Elke Lange, and Lauren Fink. 2022. Towards Efficient Calibration for Webcam Eye-Tracking in Online Experiments. In Symposium on Eye Tracking Research and Applications (ETRA). Article 27, 7 pages. https://doi.org/10.1145/3517031.3529645

[54] RealEye sp. z o.o. 2022. RealEye Webcam Eye-Tracking. Retrieved Sept. 13, 2022 from https://www.realeye.io/

[55] Rujia Sun, Xiaohe Zhou, Benjamin Steeper, Ruidong Zhang, Sicheng Yin, Ke Li, Shengzhang Wu, Sam Tilsen, Francois Guimbretiere, and Cheng Zhang. 2023. EchoNose: Sensing Mouth, Breathing and Tongue Gestures inside Oral Cavity using a Non-contact Nose Interface. In Proceedings of the 2023 ACM International Symposium on Wearable Computers (Cancun, Quintana Roo, Mexico) (ISWC '23). Association for Computing Machinery, New York, NY, USA, 22–26. https://doi.org/10.1145/3594738.3611358

[56] TDK. 2023. ICS-43434. Retrieved Mar 13, 2023 from https://invensense.tdk.com/products/ics-43434/

[57] Cihan Topal, Atakan Dogan, and Omer Nezih Gerek. 2008. A wearable head-mounted sensor-based apparatus for eye tracking applications. In 2008 IEEE Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems. IEEE, 136–139.

[58] Cihan Topal, Ömer Nezih Gerek, and Atakan Doğan. 2008. A head-mounted sensor-based eye tracking device: eye touch system. In Proceedings of the 2008 symposium on Eye tracking research & applications. 87–90.

[59] MéLodie Vidal, Jayson Turner, Andreas Bulling, and Hans Gellersen. 2012. Wearable Eye Tracking for Mental Health Monitoring. Computer Communications 35, 11 (jun 2012), 1306–1311. https://doi.org/10.1016/j.comcom.2011.11.002

[60] Quan Wang, Laura Boccanfuso, Beibin Li, Amy Yeo-jin Ahn, Claire E. Foster, Margaret P. Orr, Brian Scassellati, and Frederick Shic. 2016. Thermographic Eye Tracking. In Proceedings of the Biennial ACM Symposium on Eye Tracking Research & Applications. 307–310. https://doi.org/10.1145/2857491.2857543

[61] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1, 4 (2018), 1–20.

[62] Eric Whitmire, Laura Trutoiu, Robert Cavin, David Perek, Brian Scally, James Phillips, and Shwetak Patel. 2016. EyeContact: scleral coil eye tracking for virtual reality. In Proceedings of the 2016 ACM International Symposium on Wearable Computers. 184–191.

[63] Katarzyna Wisiecka, Krzysztof Krejtz, Izabela Krejtz, Damian Sromek, Adam Cellary, Beata Lewandowska, and Andrew Duchowski. 2022. Comparison of Webcam and Remote Eye Tracking. In Symposium on Eye Tracking Research and Applications (ETRA). Article 32, 7 pages. https://doi.org/10.1145/3517031.3529615

[64] Zhefan Ye, Yin Li, Alireza Fathi, Yi Han, Agata Rozga, Gregory D Abowd, and James M Rehg. 2012. Detecting eye contact using wearable eye-tracking glasses. In Proceedings of the 2012 ACM conference on ubiquitous computing. 699–704.

[65] 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 (Cancun, Quintana Roo, Mexico) (ISWC '23). Association for Computing Machinery, New York, NY, USA, 60–65. https://doi.org/10.1145/3594738.3611365

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

[67] Yanxia Zhang, Andreas Bulling, and Hans Gellersen. 2011. Discrimination of gaze directions using low-level eye image features. In Proceedings of the 1st international workshop on pervasive eye tracking & mobile eye-based interaction. 9–14.

[68] Anjie Zhu, Qianjing Wei, Yilin Hu, Zhangwei Zhang, and Shiwei Cheng. 2018. MobiET: A New Approach to Eye Tracking for Mobile Device. In Proceedings of the ACM International Joint Conference and International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp). 862–869. https://doi.org/10.1145/3267305.3274174