CLC____________ Number___

UDC____________ Available for reference □Yes□No

SUSTech Southern University of Science and Technology

# Undergraduate Thesis

**Thesis Title：** Acoustic-Based Eye Tracker On Glasses

**Student Name：** Wentao Xie

**Student ID：** 11510010

**Department：** Computer Science and Enginerring

**Program：** Computer Science and Technology

**Thesis Advisor：** Jin Zhang

# COMMITMENT OF HONESTY

1. I solemnly promise that the paper presented comes from my independent research work under my supervisor's supervision. All statistics and images are real and reliable.
2. Except for the annotated reference, the paper contents no other published work or achievement by person or group. All people making important contributions to the study of the paper have been indicated clearly in the paper.
3. I promise that I did not plagiarize other people's research achievement or forge related data in the process of designing topic and research content.
4. If there is a violation of any intellectual property right, I will take legal responsibility myself.

Signature:

Date:

# Acoustic-Based Eye Tracker on Glasses

谢文涛

（计算机科学与工程系　指导老师：张进）

**[ABSTRACT]:** With the development of wearable devices, there is a large number of eye tracking products in the market. However, traditional eye tracking systems are both energy-intensive and facing the risk of privacy leakage. This paper presents an acoustic-based on-glasses eye tracker that solves the above problems. The proposed acoustic-based eye tracker tracks the eye gaze position using the eyeball structural properties. With several microphones and speakers around the eye, the speakers send ultrasonic signals with different frequencies and the microphones receive the signals reflected by the eyeball. This paper deeply analyzes the sound propagation model in the system and proposes that the eye shape feature can be derived from the received ultrasonic signals via a regular demodulation algorithm. The eye shape feature is leveraged to infer the eye gaze position using Support Vector Machine Regression (SVR). We use common and off-the-shelf (COTS) hardware components to build a prototype of our system and we integrate it into a common pair of glasses. Experiments with 4 participants show that our system achieves 0.92 *cm* mean distance error at 480 *Hz* output rate. We also analyze the potential causes of the prediction errors and point out the future directions of this work.

**[Keywords]:** Acoustic Signal; Eye Tracking; Human-centered Computing

**[摘要]：**随着可穿戴设备的发展以及普及，市面上出现了各种各样的眼动仪。然而，传统的眼动仪的能耗非常高，并且使用传统的眼动仪会有隐私泄露的风险。本文介绍了一种基于超声波信号的可穿戴眼动仪解决了上述问题。它利用了眼球结构特性来跟踪人眼的注视位置。在此系统中，眼睛周围安装有数个麦克风以及扬声器，扬声器发出不同频率的超声波信号，麦克风接收眼球反射的信号。本文深入分析了此系统中声音传播的模型，并提出可以通过一种常见的解调算法从接收到的超声信号中提取眼形特征。我们发现，基于这种眼形特征，利用支持向量机的回归模型，此系统可以推断出用户当前的视线位置。基于以上研究，我们利用市场中常见的电子元件搭建了一个基于此系统的原型，并将此原型集成到一副常见的眼镜中。我们设计了数个实验并且招募了四名志愿者参与到上述实验中去。实验结果表明，我们的系统在 480 赫兹输出频率下实现了 0.92 厘米的平均距离误差。在本文的结尾处，我们还分析了误差产生的潜在原因，并指出了这项工作的未来的发展方向。

**[关键词]：**声波信号；眼动追踪；人本计算

# CONTENTS

# 1. INTRODUCTION

Recent years have witnessed the surge of human-centered computing (HCC) technologies. Unlike traditional computing technologies, human in HCC systems not only act as the service receiver but also get themselves involved in the computing process. The objective of human-centered computing technologies mainly lies in assisting people with their daily activities [20], monitoring people's health condition [22] and providing people with a better human-computer interaction experience [3]. The human eye, which is one of the most important sense organs, provides people with massive amounts of information about the world and it also indicates both physical and mental health conditions. Thus, eye tracking technologies should be studied in HCC systems.

According to recent research, eye movement is a marker of some health issues (e.g., mental disorders and cognitive disorders) [1]. Thus, continuous eye tracking systems have the potential ability to assist in the diagnosis of these diseases. Furthermore, because the eye movement indicates where the user's current attention is on, a hands-free and attention-based human-computer interaction method can be achieved by tracking the user's eye movement. With the development of wearable devices, smart glasses (including VR/AR devices) are entering people's horizons. Except that smart glasses can perform basic tasks as a smartphone does (e.g., displaying text, image, and videos), smart glasses can also provide people with a new human-computer interaction experience using eyes. However, the existing smart glasses generally use cameras to capture the user's eye movement which consumes lots of energy and faces the risk of privacy leakage.

In recent years, researchers have proposed that acoustic signals can be used to perform short-distance tracking especially for mobile devices [1]. In these scenarios, the acoustic signals are used because they consume less energy than the traditional RF-based and image-based methods. Besides, similar to RF-based methods, acoustic-based tracking systems have no risk of privacy leakage. With those good features of acoustic-based sensing technologies and the rapid development of smart glasses technologies, an acoustic-based on-glasses eye tracking system is desirable.

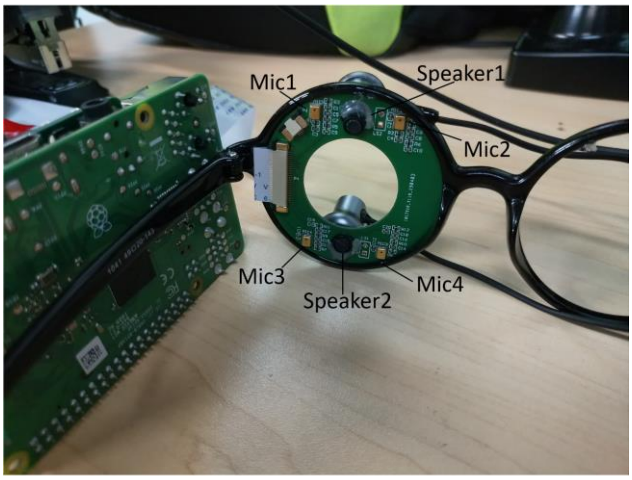


**Figure 1. System prototype.**

In this paper, we propose an acoustic-based eye-tracking system on glasses. As is shown in Figure 1, we mount two speakers on the upper and lower frame of the glasses to send ultrasonic signals and we mount four microphones at the four corners of the frame to receive the acoustic signals. Our system achieves eye tracking through two phases: the training phase and the testing phase. In the training phase, we let the subjects wear the glasses and instruct them to stare at a moving ball on a computer screen. Then, we collect the ultrasonic signals from the four microphones and we use quadrature demodulation to extract the eye shape feature. After the eye shape feature from the four microphones is derived, we label the data with the actual coordinates of the moving ball and use the data as well as the labels to train a Support Vector Machine Regression model (SVR). In the testing phase, similar to the training phase, we let the subject wear our glasses and stare at an arbitrarily moving ball on the screen while the system computes the eye shape feature from the four microphones. Then, the system passes the eye shape features to the SVR model to predict the current pupil position.

We summarize our main contribution as follow:

- We proposed an acoustic-based eye-tracking system with high accuracy and low energy consumption. To the best of our knowledge, this is the first work to explore the feasibility of tracking eye movement using acoustic signals.

- We implement the system on a regular pair of glasses using COTS (commercial off-the-shelf) hardware. The total cost of our prototype is within 250 RMB.

The following of this paper is organized as follow: Section 2 summarizes the related work of this paper. Section 3 shows the basic idea of our system as well as the challenges we need to overcome. Section 4 elaborates on the design details of our system. Section 5 presents the hardware design of the prototype of our system. Section 6 evaluates the system performance and Section 7 concludes this paper.

# 2. RELATED WORK

In this section, we briefly summarize the related work of this paper which can be roughly classified into the following two categories: acoustic sensing systems and eye tracking systems.

## 2.1 Acoustic Sensing Systems

Recent years acoustic sensing technologies have been lucubrated by researchers mainly because microphones and speakers are cheap, energy saving and have already been widely deployed (e.g., on smartphones). The following summarizes the most studied acoustic sensing systems. LLAP [1] uses the phase shift of a sinusoidal signal to estimate the distance change between the target object and the smartphone to perform tracking. FingerIO [2] and CovertBand [3] let the speaker sends OFDM symbols and perform cross-correlation on the received signal with the original OFDM symbols, then extracting the object motion by analyzing the correlation profile. Strata [4] basically uses the 26-bit GSM signal to estimate the channel impulse response (CIR), and the system derives the target motion by analyzing the phase change of each channel tap. In [4], the authors compare the performance of Strata with LLAP-based (low latency acoustic phase) methods and CC-based (cross-correlation) methods. They claim through their experiments that Strata achieves higher accuracy than the other two methods. EchoTrack [5] uses the FMCW signal as the transmitted signal. EchoTrack passes the received signal through a matched filter to localize the target object. [7] uses FMCW as well. They develop a technique called C-FMCW to detect the user's respiration rate when sleeping. Unlike traditional FMCW-based techniques, C-FMCW estimates the round-trip propagation time of the acoustic signals by discovering the maximally correlated offset between the transmitted signal and the received signal, then estimating the breathing rate through the variation of the propagation time. AAMouse [8] uses the Doppler shift of a sinusoidal signal to estimate the velocity, then tracking the motion of a mobile device. CAT [9] develops a distributed FMCW and uses Doppler shift to track the mobile device's motion. Vernier [10] develops a differentiated window-based phase change estimation method to estimate the device motion.

## 2.2 Eye Tracking Systems

Most existing eye tracking systems are camera-based. To achieve high accuracy and robust eye tracking, the existing techniques use near-infrared (NIR) LED to illuminate eyes and highlight the pupils [11]. There are also works that examine the use of camera under natural light [12]. With the development of wearable devices, researchers have developed eye tracking systems in the wearable context. iGaze [13] deploys a camera on glasses to track eye movement. Since it is an image-based system, its power consumption is high. iShadow [14] and CIDER [15] are camera-based on-glasses eye-tracking systems as well, they reduce the system power consumption by sampling the image's pixels while not losing any useful information. Apart from the camera-based systems, there is some work proposed that eye tracking can be achieved in a non-camera way. [16] is an eye-tracking system on VR devices. It leverages the pupil's light absorption effect of the screen light of a VR device to estimate the pupil position. [17] is another eye-tracking system on general glasses. Similar to [16], they make use of the pupil's light absorption effect to predict the pupil position. Rather than measuring the light absorption of the screen light passively, [17] put an array of NIR LEDs on glasses and actively illuminate the eyeball. Thus, their system can be used in general scenarios.

# 3. METHODOLOGY AND CHALLENGES

In this work, we propose a wearable eye tracking system using acoustic signals. As is shown in Figure 2, the human eyeball is not a perfect sphere which consists of a sphere-like vitreous body and an ellipsoid-like lens and anterior chamber. In this way, the surface of the human eye exposed to the air is not a smooth spherical surface. When the eye is moving, its surface will have different shapes.

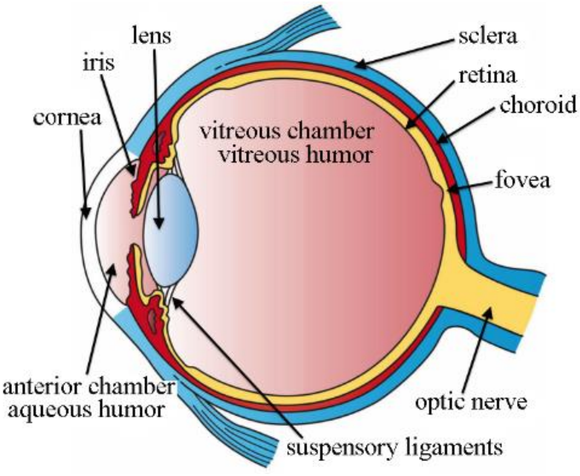


**Figure 2. The eyeball structure[1].**

We exploit this effect for eye tracking without cameras, *i.e.*, we leverage eye shape to infer eye gaze position. As is shown in Figure 1, we use speakers and microphones to build a simple sonar system on glasses to capture the eye surface changes when the eye is moving. Specifically, we mount two speakers on the top and lower edge of the frame of glasses, and we mount four microphones at the four corners of the frame. We let the speakers transmit ultrasonic CW (continuous wave) signals. Since the speakers are mounted toward the eyeball, the transmitted acoustic signals thereby will be reflected by the eyeball and finally received by the four microphones. Intuitively, when the eyeball is moving, the transmitted signals will propagate through varied reflection

[1] Image from http://open.umich.edu/education/med/resources/second-look-series/materials

paths because the reflection surface is changed (*i.e.*, the eyeball surface is changed). Thus, the received signal changes accordingly. The system measures the changes of eye shape by capturing the variation of the received signals. The system infers the eye gaze positions using Support Vector Machine Regression (SVR). The system diagram is shown in Figure 3.

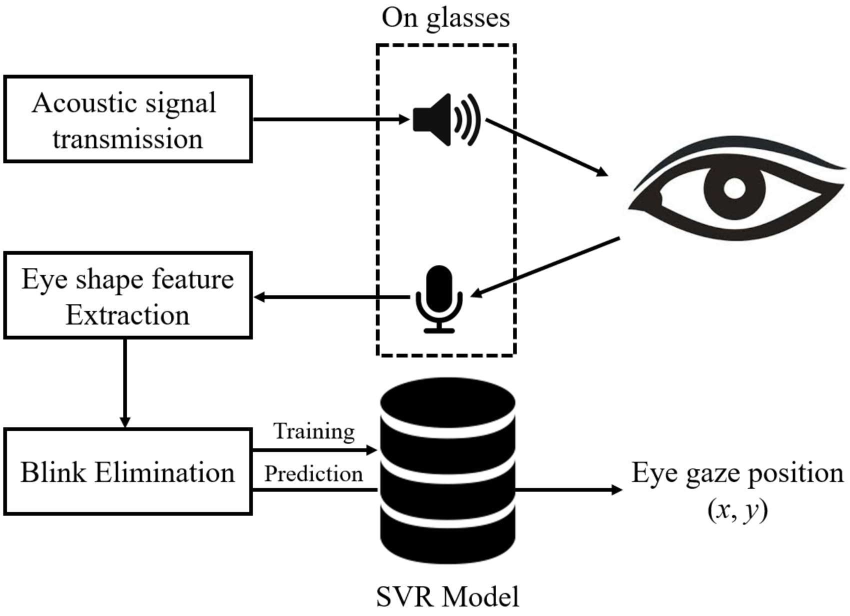

On glasses
Acoustic signal transmission
Eye shape feature Extraction
Blink Elimination
Training
Prediction
SVR Model
Eye gaze position (*x*, *y*)

**Figure 3. System diagram.**

**Challenges.** To achieve high-accuracy eye tracking using the above methodology, we face the following three main challenges:

1. Due to the multi-path effect, the received signal is the superposition of signals propagating through different reflection paths. Thus, the phase computed by our system is caused by the eyeball reflection as well as other reflection paths (*e.g.*, reflected by the eyelid). Consequently, to infer the eyeball movement from the received signal, our system needs to consider the influence of multi-path effect.
2. Unlike a camera, the microphone array of our system does not provide any spatial resolution within its field-of-view, but only a feature vector with only a few elements. Thus, a minor eye movement may result in negligible

differences in the feature vector computed by the system, which severely limits the system accuracy and resolution.

3. In practical use, blink happens frequently. Since our system infers the eye movement by sensing the variation of eye shape, eye blink severely influences the eye movement prediction because it largely changes the shape of the reflection surface. Consequently, blink events should be accurately detected and discarded.

In the next section, we will describe our solutions to addressing the above three challenges.

# 4. EYE TRACKING USING ACOUSTIC SIGNALS

In this section, we elaborate on the detail design of our on-glasses eye-tracking system. We first present the algorithm we use to derive the eye shape information. Then, we show how our system detects the blink events and eliminates the influence of eye blink on our system. Finally, we present our approach to infer the eye movement from the eye shape information.

## 4.1 Sensing the Eye Surface Shape with Acoustic Signals

Our objective is to build an eye-tracking system with high spatial sensing resolution so that the system is capable of differentiating the tiny motion of eye movement. Since our system uses speakers and microphones to build a sonar system and each pair of speaker and microphone covers a region of the eyeball, a straightforward way to improve the system performance is to increase the sensor number (number of speakers and microphones). In our system, each pair of microphone and speaker covers a region of the eye surface and generates an eye shape feature. With $M$ speakers and $N$ microphones, the hardware can provide $M \times N$ eye-shape features. As is shown in Figure 1, we mount two speakers on the top and lower edge of the frame of the glasses and we mount four microphones at the four corners of the frame. Thus, our system generates 8 eye-shape features each time.

### 4.1.1 Acoustic Signals Design

In practical use, our system should work without making any noise, thus, the transmitted signal should be inaudible (≥ 17kHz). However, the sampling rate of a normal audio device is 48kHz, the maximum frequency accessible to the system is 24kHz according to the Nyquist Theorem. Thus, the acoustic signals sent by the system should be within 17kHz and 24kHz. There are 2 speakers in our system, we let one speaker send 18kHz CW signal and let the other one send 22kHz CW signal to avoid interference.

### 4.1.2 The Sound Propagation Model

When the system is running, the speakers send acoustic signals. The signals propagate in the air and later be reflected by the user's eye, and microphones finally receive the signals. The following presents the detailed math model of the above propagation process. For a pair of speaker and microphone, suppose that the speaker transmits the following signal:

$$S(t) = cos(2\pi f t), \tag{1}$$

where *f* is the frequency of the transmitted signal. After being transmitted by the speaker, a signal propagates through multiple reflection paths before being received by the microphone (*e.g.*, reflected by the cornea and reflected by eyelid). Assume that there are *K* reflection paths and the *i*th reflection path is associated with a delay $\tau_i(t)$ and a fading factor of $A_i(t)$ where $\tau_i(t) = 2d_i(t)/v_s$ where $d_i(t)$ represents the half round-trip distance of the *i*th reflection path and $v_s$ denotes the speed of sound in air. Note that the delay and the fading factor is a function of time because when the eye is moving, the reflection strength as well as the delay of each reflection path will change. Under the above assumption, the microphone receives the following signal:

$$R(t) = \sum_{i=1}^{K} A_i(t) cos(2\pi f(t - \tau_i(t))) = \sum_{i=1}^{K} A_i(t) cos(2\pi f t - 2\pi f \tau_i(t)). \tag{2}$$

In Equation (2), the phase $2\pi f \tau_i(t)$ is proportional to the round-trip distance of the reflection path and the amplitude $A_i(t)$ represents the reflection strength of the reflection path. Consequently, $\tau_i(t)$ and $A_i(t)$ are the two features which are capable of identifying a reflection path and by collecting these features of all the reflection paths, we are able to model the eye surface shape. However, these two kinds of important information are modulated into the transmitted signal $S(t)$ by amplitude modulation (AM) and phase modulation (PM) respectively. The following section presents our method to extract $\tau_i(t)$ and $A_i(t)$.

### 4.1.3 Eye Shape Extraction

Our system derives the eye shape features, *i.e.*, $\tau_i(t)$ and $A_i(t)$, from the received signal $R(t)$ using quadrature demodulation. The following shows the quadrature demodulation process of our system. *Firstly*, we multiply the received signal $R(t)$ with its in-phase carrier $cos(2\pi ft)$ with a gain of 2, *i.e.*,

$$T_I(t) = R(t) \cdot 2cos(2\pi ft). \tag{3}$$

By applying the Prosthaphaeresis Equation to Equation (3), we get

$$T_I(t) = \sum_{i=1}^{K} A_i(t)\big(cos(4\pi ft) + \cos\big(2\pi f\tau_i(t)\big)\big). \tag{4}$$

*Secondly*, we multiply $R$(t) with its quadrature carrier $(-sin(2\pi ft))$ with a gain of 2, *i.e.*,

$$T_Q(t) = R(t) \cdot 2\big(-sin(2\pi ft)\big). \tag{5}$$

Again, with the Prosthaphaeresis Equation,

$$T_Q(t) = \sum_{i=1}^{K} A_i(t)\big(-sin(4\pi ft) + sin\big(2\pi f\tau_i(t)\big)\big). \tag{6}$$

From Equation (4) and (6), both $T_I(t)$ and $T_Q(t)$ consists of two components: a high frequency $A_i(t)cos(4\pi ft)$ (or $A_i(t)sin(4\pi ft)$) and $A_i(t)cos(2\pi f\tau_i(t))$ (or $A_i(t)sin(2\pi f\tau_i(t))$), and the second component is what we want to keep because it contains $A_i(t)$ and $\tau_i(t)$. Because eyeball rotation does not have very high frequency (*e.g.*, saccade, the fastest type of eye movement, lasts around 200 *ms* on average [20] with the velocity up to 700 °*/s* [18]). Thus, the second component does not have the comparable high frequency as the first component. Consequently, we extract the second component $cos(2\pi f\tau_i(t))$ and

$sin(2\pi f\tau_i(t))$ from $T_I(t)$ and $T_Q(t)$ by using a low-pass filter. After filtering, the remaining signals are

$$I(t) = \sum_{i=1}^{K} A_i(t)\, cos\big(2\pi f \tau_i(t)\big), \tag{7}$$

and

$$Q(t) = \sum_{i=1}^{K} A_i(t)\, sin\big(2\pi f \tau_i(t)\big). \tag{8}$$

We use a vector $\boldsymbol{\Phi}(t)$ to join the in-phase $I(t)$ component and the quadrature component $Q(t)$, *i.e.*,

$$\boldsymbol{\Phi}(t) = \big(I(t), Q(t)\big). \tag{9}$$

We define a feature vector of the *i*th reflection path as

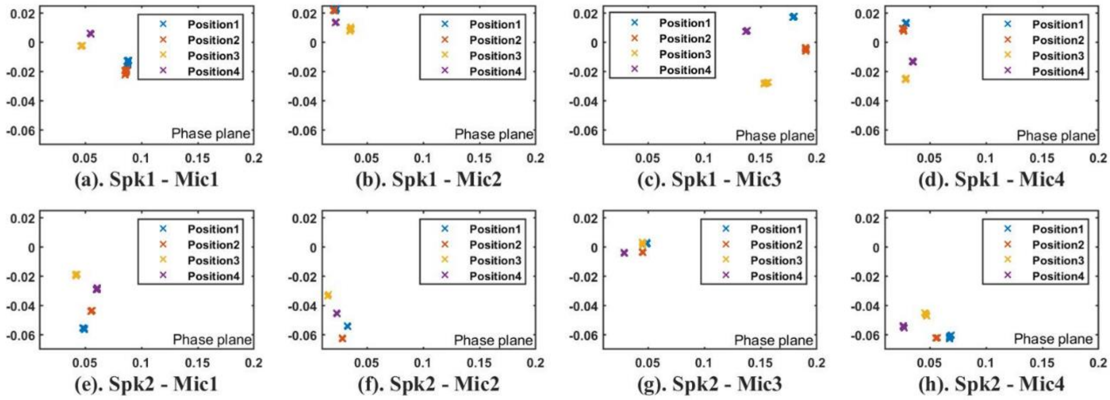
$$\boldsymbol{\varphi_i}(t) = A_i(t) \cdot \Big(cos\big(2\pi f \tau_i(t)\big), sin\big(2\pi f \tau_i(t)\big)\Big), \tag{10}$$

so that $\boldsymbol{\varphi_i}(t)$ can contains all the information we want for identifying a single reflection path, *i.e.*, the magnitude of $\boldsymbol{\varphi_i}(t)$ represents the reflection strength $A_i(t)$ of the reflection path and the angle of $\boldsymbol{\varphi_i}(t)$ represents the delay $\tau_i(t)$ of the reflection path. Consequently, we can rewrite $\boldsymbol{\Phi}(t)$ according to Equation (10):

$$\boldsymbol{\Phi}(t) = \sum_{i=1}^{K} \boldsymbol{\varphi_i}(t). \tag{11}$$

In this way, $\boldsymbol{\Phi}(t)$ is the net feature vector of one pair of speaker and microphone and it is the summation of feature vectors of all the reflection path covered by this pair of speaker and microphone. We use $\boldsymbol{\Phi}(t)$ to represent the eye surface shape.

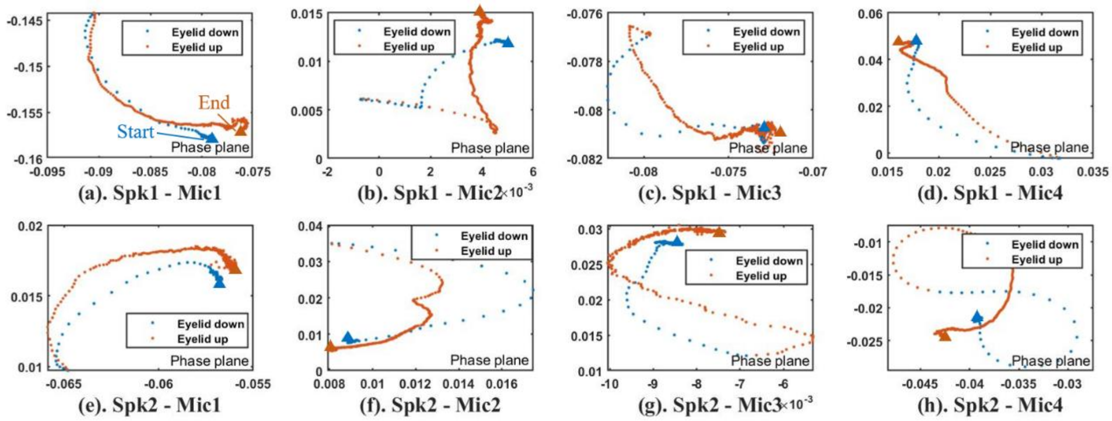
We conduct an experiment to examine the feasibility of using $\boldsymbol{\Phi}(t)$ to represent the eye surface shape. We let a subject wear our prototype and ask the subject to stare at objects located at *upper left*, *upper right*, *lower left* and *lower right* direction of his/her eye respectively. We compute $\boldsymbol{\Phi}(t)$ when the subject stares at these four directions. Figure 4 shows the result and it is evidently that when the subject watches different directions, the eye shape feature $\boldsymbol{\Phi}(t)$ of each pair of miccophone and speaker have distinct value in the phase plane. Consequently, $\boldsymbol{\Phi}(t)$ is capable of indicating the eye surface shape.



(a). Spk1 - Mic1 (b). Spk1 - Mic2 (c). Spk1 - Mic3 (d). Spk1 - Mic4

(e). Spk2 - Mic1 (f). Spk2 - Mic2 (g). Spk2 - Mic3 (h). Spk2 - Mic4

**Figure 4. Eye shape feature *Φ(t)* when the subject stares at different positions.**



(a). Spk1 - Mic1 (b). Spk1 - Mic2 (c). Spk1 - Mic3 (d). Spk1 - Mic4

(e). Spk2 - Mic1 (f). Spk2 - Mic2 (g). Spk2 - Mic3 (h). Spk2 - Mic4

**Figure 5. Eye shape feature *Φ(t)* during a blink.**

## 4.2 Detecting Eye Blinks

Since our system uses the surface shape of the human eye to infer the eye movement, our system should have the ability to eliminate the impact of eye blinks. We achieve this objective by detecting the blink event and discard the data during a blink.

Although the eye shape feature $\boldsymbol{\Phi}(t)$ should be discard during a blink, it is useful in identifying a blink. When eye blink happens, the eyelid is closed which will completely cover the eye surface thus largely changing the surface shape. Besides, a blink interval is short (a blink lasts 100 to 400 *ms* [20]). Thus, the eye shape feature $\boldsymbol{\Phi}(t)$ changes dramatically during a blink. Figure 5 shows an example of $\boldsymbol{\Phi}(t)$ during a blink.

In Figure 5, each dot denotes a sampling point of $\boldsymbol{\Phi}(t)$.

A straightforward way to detect a blink is to compare the gradient of eye shape feature (*i.e.*, $||\boldsymbol{\Phi}(t) - \boldsymbol{\Phi}(t-1)||$) to a fixed threshold. However, to set a proper threshold is tough as it is related to the current noise level which varies both temporally and spatially. We address this problem by applying the constant false alarm rate detection (CFAR). CFAR is widely used in radar system to detect a sharp signal rise or fall in a context with the varying noise level. In general, CFAR estimates the current noise level by examining $m$ reference samples around the current test sample which excludes $n$ guard samples adjacent to the current test sample. The guard samples are set to avoid corrupting the noise level estimation with the current testing sample. Let $\boldsymbol{\Phi}(t)$ denotes the current eye shape feature sample. Let $\boldsymbol{R}(t)$ denotes the reference samples which contains $\boldsymbol{\Phi}(t - n_l - m_l), \ldots, \boldsymbol{\Phi}(t - n_l - 1), \boldsymbol{\Phi}(t + n_r + 1), \ldots, \boldsymbol{\Phi}(t + n_l + m_r)$ , where $n=n_l+n_r$ and $m=m_l+m_r$. Then a blink is detected if the following holds for at least half of the speaker-microphone pairs:

$$\max_{i} ||\boldsymbol{\Phi}(t) - \boldsymbol{R}(i)|| > \alpha \cdot \max_{j,k} ||\boldsymbol{R}(j) - \boldsymbol{R}(k)||, \tag{12}$$

where $\alpha$ is the threshold factor. We set $n_l$, $n_r$, $m_l$, $m_r$ and $\alpha$ to be 240, 960, 480, 480 and 2 respectively considering that the sampling frequency of our system is $48kHz$. Note that traditionally, CFAR is aimed at detecting dramatic changes in one-dimensional data. Here we extend the algorithm so that it can detect sharp changes in multi-dimensional data. Once detecting a blink, the system discards the following 250 *ms* data given that a blink lasts around 250 *ms* on average [20].

We conduct an experiment to examine whether our system can accurately detect a blink and eliminate its impact. We ask a user wears our glasses and instruct him/her to stare at a moving ball on the screen. Since human eye blinks frequently in nature, we collect several blinks in this test. Figure 6 shows the $\boldsymbol{\Phi}(t)$ vectors before blink elimination. Note that since $\boldsymbol{\Phi}(t)$ is a 2-dimenssional vector, it is hard to represent a series of $\boldsymbol{\Phi}(t)$ over time. Thus, we separate its $x$ value and $y$ value. It is obvious that blinking leads to a dramatic change of $\boldsymbol{\Phi}(t)$ (those sharp rises and falls in Figure 6). After applying the CFAR algorithm we can locate each blink in the raw data and discard the data as shown in Figure 7. From this experiment, we can draw a conclusion that our CFAR algorithm successfully eliminates the impact of blinks.

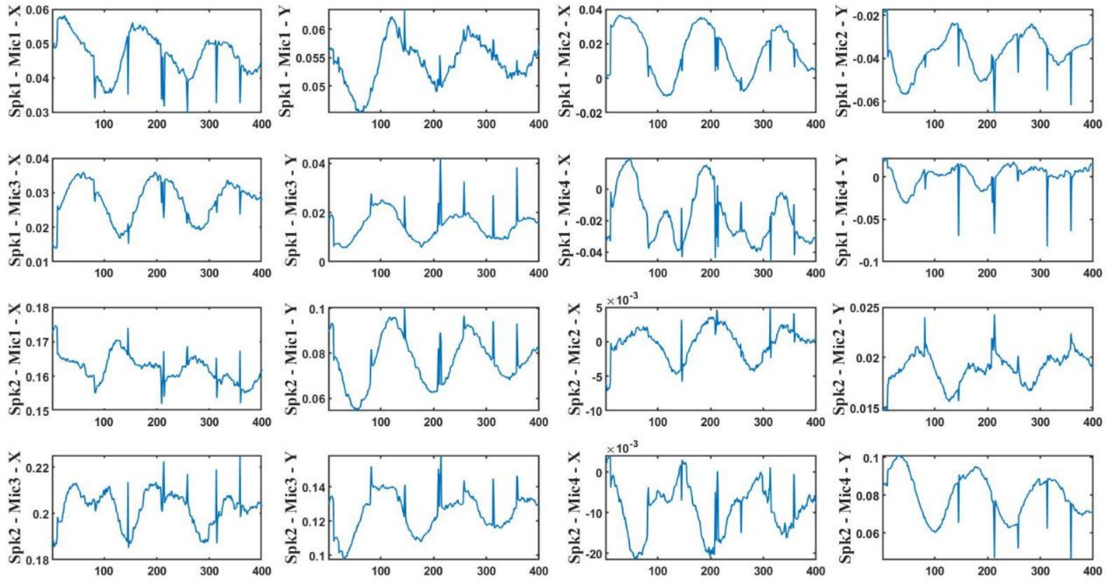


**Figure 6. *Φ(t)* vectors before blink detection.**

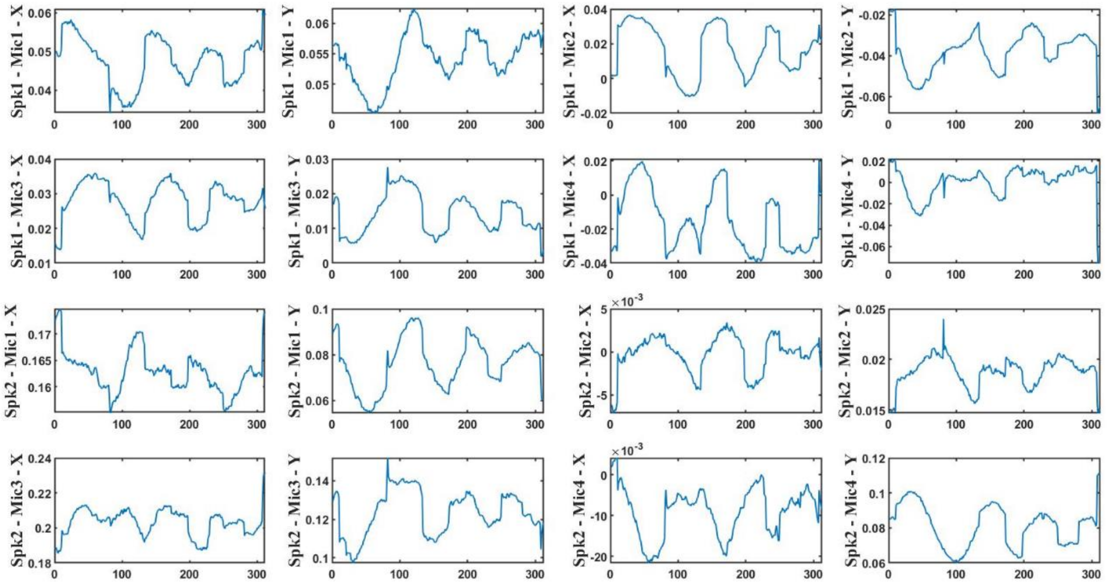
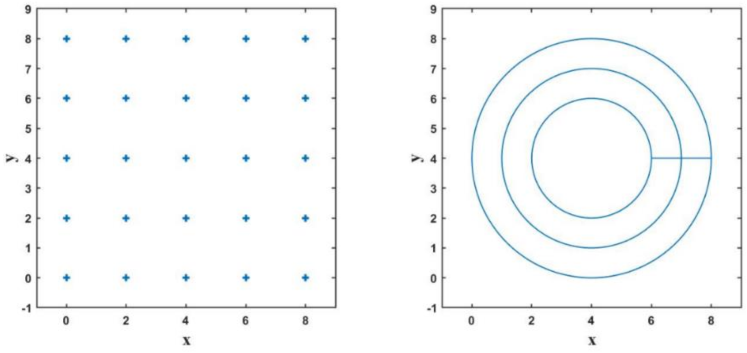


**Figure 7. *Φ(t)* vectors after blink detection.**

## 4.3 Inferring Eye Gaze Position

With $M \times N$ data points (eye shape features) from the microphones at a time instance, the next step is to infer the 2D coordinate of the eye gaze position. We achieve this objective with supervised learning to train personalized models capturing the relationship between the sensing data and eye gaze status. With the trained models, we then compute the eye gaze position based on the current sensing data. We use Support Vector Machine Regression (SVR) with Gaussian kernel and we train separate models for eye gaze position in *x* and *y* axis.



**(a) Training step** **(b) Prediction step**

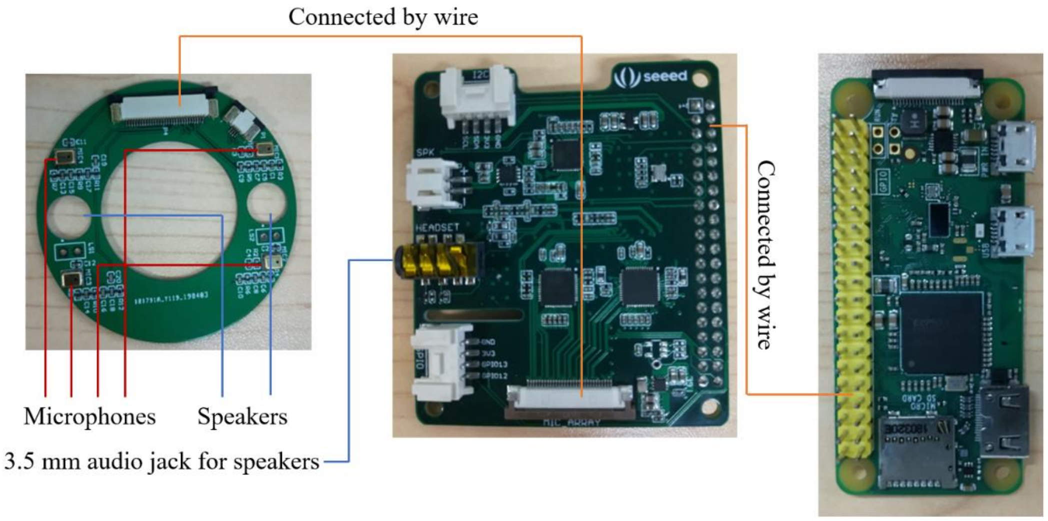
**Figure 8. Trajectories of the moving ball on the screen.**

**Offline training.** To train the models for a new user, we collect 190-second data where we ask the user to wear our glasses and stare at a ball on a computer screen. The ball randomly jumps among 25 pre-defined positions on the screen (the pre-defined ball positions are shown in Figure 8(a)). Meanwhile, our system continuously collects eye shape features @ 480*kHz*. Later, we record the actual positions of the ball to obtain the ground truth of eye gaze's 2D positions. Since there are $M \times N$ eye shape features and each feature has 2 data ($\boldsymbol{\Phi}(t)$ is a two-dimensional vector), the feature vector $\mathbb{F}(\mathrm{t})$ of the current eye shape has $2 \cdot M \cdot N$ elements which is 16 in our system. Specifically, let $\boldsymbol{\Phi}_{i,j}(t)$ denotes the eye shape feature computed by the *j*th microphone with the corresponding *i*th speaker at time *t*. We compute $\mathbb{F}(\mathrm{t})$ as:

$$\mathbb{F}(\mathrm{t}) = \{\boldsymbol{\Phi}_{i,j}(t) | 1 \leq i \leq M, 1 \leq j \leq N\}. \tag{13}$$

Then, we use the 16-element $\mathbb{F}(\mathrm{t})$ as well as the corresponding ground truth 2D positions to train the SVR model.

**Online inference.** As the reflected signals arrive on the fly, we compute the feature vector $\mathbb{F}(\mathrm{t})$ as Equation (13), and feed the vector to the SVR model to infer the current eye gaze positions.



(a) Speakers and microphones (Rebuild from Seeed Respeaker 4-mic linear array).

(b) Data acquisition unit (Part of Seeed Respeaker 4-mic linear array).

(c) Signal processing unit.

**Figure 9. Each unit of our prototype.**

# 5. PROTOTYPE IMPLEMENTATION

We have built a simple prototype using COTS hardware components. The prototype composed of three units as shown in Figure 9. The first unit is the **sound transmission/receiving unit**. We rebuild the signal receiving module of the Seeed Respeaker 4-Mic Linear Array[2] to fit the scenarios of our system. Originally, the signal receiving module is a long stripe of circuit board mounted with 4 microphones. We rebuild this module so that it fit the shape of a round glasses frame (Figure 9(a)). We also reserve two interfaces for speakers (i.e., the two circular holes on the board) where we simply use a pair of Xiaomi earphones 2nd gen[3] as the speakers. The second unit is the **data acquisition unit**. We directly use the data acquisition module of the Seeed Respeaker 4-Mic Linear Array to drive the speakers and microphones (Figure 9(b)). The third unit is the **control unit**. We use the Raspberry Pi Zero W[4] to control the signal transmission and reception as well as process the received signals (Figure 9(c)). Note that since the speakers and microphones share one crystal oscillator (located on the Raspberry Pi

[2] Seeed Respeaker 4-Mic Linear Array. http://wiki.seeedstudio.com/ReSpeaker_4-Mic_Linear_Array_Kit_for_Raspberry_Pi/

[3] Xiaomi Earphones 2nd Gen. https://www.mi.com/quantie2/

[4] Raspberry Pi Zero W. https://www.raspberrypi.org/products/raspberry-pi-zero-w/

Zeros W board), there is no frequency offset between the speakers and microphones which usually exists in acoustic systems because of the unperfect clocks between the speaker and microphone [8].

# 6. EVALUATION

In this section, we conduct several experiments to evaluate the performance of our system. We recruit four volunteers to participate in our experiments.

## 6.1 Experimental Setup

The experimental scenario is shown in Figure 10 where we mount a head holder on the table to fix the head position and we put a computer screen 20 *cm* away from the holder which directly faces the subject. With the above settings, we collect training and testing data from each subject.

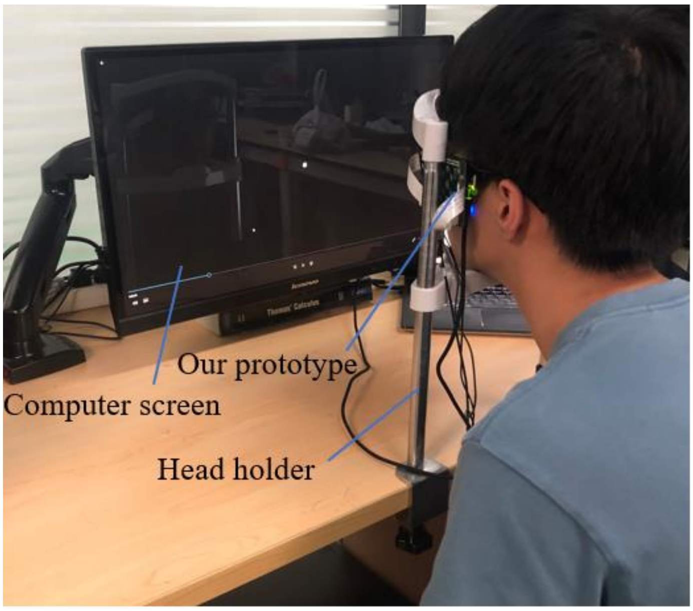


**Figure 10. Experiment scenario**

**Training phase.** The training phase lasts around 3 minutes, where each participant wears our glasses, sits in front of the screen and stares at a moving ball on the screen. The ball is programmed to jump between 25 pre-defined screen positions where the ball stays 1.5 seconds in each position and each position appears 5 times in the training phase. The background color of the screen is set to be black and the ball is white. We collect $9 \times 10^6$ raw feature vectors with the system sampling frequency of 48kHz. We first down-sample the collected data 100 times (sampling rate becomes 480Hz) to reduce the data redundancy. Next, we discard the data when the ball switches from one position to another. Specifically, since the ball stays in a position for 1.5 seconds, we retain the middle 0.5 seconds of data and discard the rest of it. In this way, we collect $3 \times 10^4$ feature vectors and use the data as well the labels to train SVR models. The model training lasts around 30 seconds.

**Testing phase.** We design a test to evaluate the accuracy of our system. We ask the participants to stare at a moving ball on the screen where we programmed the ball to move along three concentric circles (Figure 8(b)). Similar to the training phase, we resample the sensing data to 480kHz to lower the data redundancy. We compare the eye gaze positions predicted by our system with the ground truth ball trajectories (i.e., the de facto ball coordinates on the computer screen).

**Evaluation metrics.** We use two metrics to evaluate system performance. (i) *Distance deviation*. The distance deviation is defined as the absolute distance difference between the inferred and ground truth eye gaze position. (ii) *Angular deviation*. Since the trajectories of the moving ball on the computer screen are concentric circles, the angle of the inferred position with respect to the center point also indicates the system accuracy. The angular deviation is defined as the angular difference between the inferred position and the ground truth position. Note that angular deviation is a relatively loose metric compared to distance deviation because it only measures the angular motion of the eye gaze position without exact coordinate information.

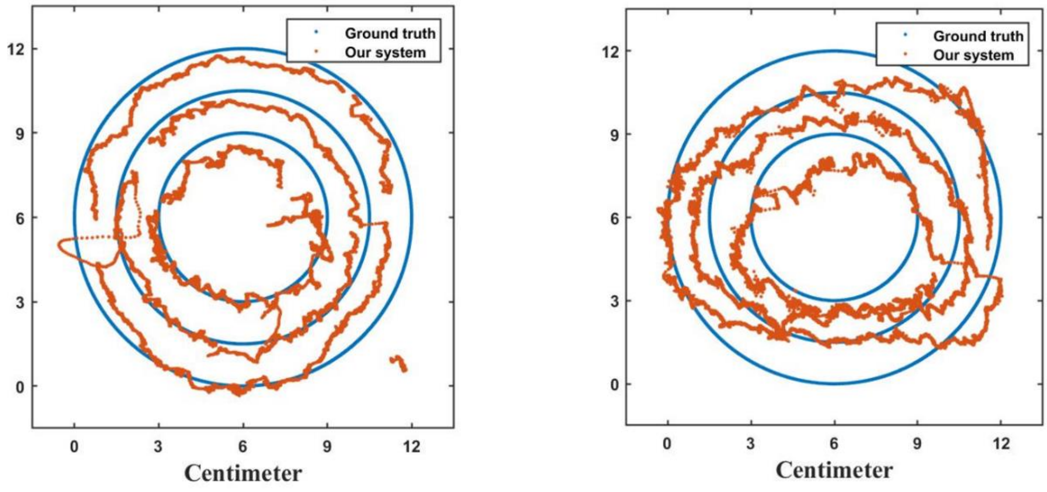
## 6.2 Experimental Result

In this section, we present our experimental result.

### 6.2.1 Eye Gaze Inference Accuracy

We first show the eye gaze inference accuracy of our system. Two representative results are shown in Figure 11 where the blue lines represent the ground truth eye gaze positions and the red lines represent the eye gaze positions inferred by our system. Figure 11(a) shows an example of a good inference result whose trajectory is basically consistent with the ground truth. Figure 11(b) shows an example of a poor inference result where the inferred eye gaze positions are largely different from the ground truth, but still, we could clearly observe three concentric circles which means our system can give meaningful result anyway.

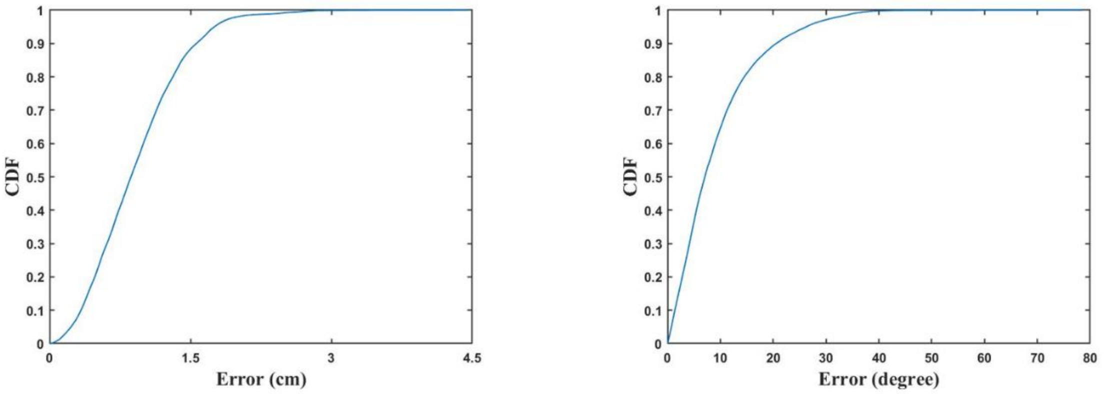
Figure 12 shows the eye gaze tracking accuracy in our experiments. Figure 12(a) shows the CDF (Cumulative Distribution Function) of distance deviation. The mean distance error is 0.92 cm, the standard deviation of distance error is 0.49 cm and the maximum distance error is 4.46 cm. Figure 12(b) shows the CDF of angular deviation. The mean angular error is 9.30 degree, the standard deviation of the angular error is 8.00 degree and the maximum angular error is 78.55 degree. Overall our system can infer the current eye gaze position with distance accuracy around 1 cm. Although sometimes the inferred gaze position is largely different from the ground truth position, the angular error is small which means the inferred eye movement pattern is still consistent with the ground truth ball trajectory but the final result is out of shape for some reasons (Figure 11 (b)). The following paragraph gives the possible reasons for the out-of-shape inference of our system.



(a) Good result

(b) Poor result

**Figure 11. Eye gaze inference results.**



**(a) Distance error.** **(b) Angular error.**

**Figure 12. Accuracy of eye gaze tracking.**

There are two possible reasons for the out-of-shape inference: (i) *Head movement.* On default, we assume that the subject's head is stationary during our experiments. However, the subject's head may still have unconscious tiny motion even with the head holder. Thus, staring at the same position may result in different eye shapes due to the head movement which will give rise to errors in both the model training and eye gaze position inference. (ii) *Impact of eye closure state.* This is an observation from daily lives. When someone stares at the same position, the eye closure state may change over time. In the training phase of our experiment, the subject needs to stare at each position on the screen five times, and each time the subject's eye closure state may be different from others. Since our system infers the eye gaze position through sensing the eye shape, different eye closure states result in different eye shape feature $\boldsymbol{\Phi}(t)$. Thus, errors are introduced in the model training phase and the situation is similar during the prediction phase.

### 6.2.2 Blink Detection Accuracy

Since the system performance highly relies on the elimination of eye blinks (Section 4.2), we also evaluate the blink detection accuracy of our system. We use all the eye shape feature data collected in the above experiment to do the evaluation. When the subject is doing the experiments, we use a camera to record the user's eye state. Later, we manually locate each blink by watching the video captured by the camera. Figure 13 shows the experimental result.

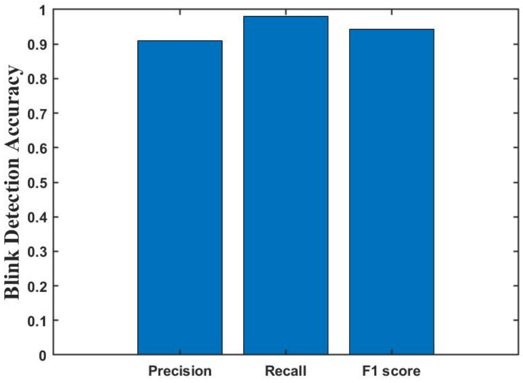


**Figure 13. Blink detection accuracy.**

The experimental result shows that our system accurately detects eye blink. Specifically, the precision is 90.7%, the recall rate is 98.2% and the F1 score is 94.3%. Note that the recall rate is much higher than the precision which means the system seldom misses a blink but may miss-classify non-blink data into a blink then be discarded by the system. This is an expected result due to a relatively low threshold factor $\alpha$ (*i.e.*, $\alpha = 2$). We think it is sensible to do so because reducing the data amount is better than including wrong data to the training set.

## 7. DISCUSSION

This work explores the feasibility of eye gaze tracking using acoustic signals. The experimental result shows that the overall system accuracy is good. However, there are some limitations of the current design: (i) we use the eye shape to infer the current eye gaze position. However, since the eye closure level is not exactly the same when a person stares at the same object, errors are included in the model training process as well as the eye gaze position inference process; (ii) we use the Raspberry Pi Zero W as the control unit of our system. Since it is a not a special designed device for our system but rather a general platform for developers, the energy consumption is still high. In the future, we may continue this work and mainly focus on the above limitations.

## 8. CONCLUSION

In this work, we designed and implemented an acoustic-based wearable eye tracker using COTS hardware. The total cost of the hardware is within 250 RMB. We use acoustic signals to capture the eye shape feature, and we leverage this eye shape feature to train a Support Vector Machine Regression (SVR) model which is used to infer the current eye gaze position. We design several experiments to evaluate the performance of our system. The experimental result shows that our system can accurately track the eye gaze with accuracy around 1 *cm*. We also analyze the possible causes of errors which may be the direction of our future works.

# REFERENCES

[1] IACONO W G, MOREAU M, BEISER M, et al. Smooth-pursuit eye tracking in first-episode psychotic patients and their relatives.[J]. Journal of Abnormal Psychology, 1992, 101(1):104

[2] WANG W, LIU A X, SUN K. Device-free gesture tracking using acoustic signals[C]//Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking. [S.l.]: ACM, 2016: 82-94.

[3] NANDAKUMAR R, IYER V, TAN D, et al. Fingerio: Using active sonar for fine-grained finger tracking[C]//Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. [S.l.]: ACM, 2016: 1515-1525.

[4] NANDAKUMAR R, TAKAKUWA A, KOHNO T, et al. Covertband: Activity information leakage using music[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2017, 1(3):87.

[5] YUN S, CHEN Y C, ZHENG H, et al. Strata: Fine-grained acoustic-based device-free tracking[C]//Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services. [S.l.]: ACM, 2017: 15-28.

[6] CHEN H, LI F, WANG Y. Echotrack: Acoustic device-free hand tracking on smart phones[C]//IEEE INFOCOM 2017-IEEE Conference on Computer Communications. [S.l.]: IEEE, 2017: 1-9.

[7] ZHANG Y, WANG J, WANG W, et al. Vernier: Accurate and fast acoustic motion tracking using mobile devices[C]//IEEE INFOCOM 2018-IEEE Conference on Computer Communications. [S.l.]: IEEE, 2018: 1709-1717.

[8] WANG T, ZHANG D, ZHENG Y, et al. C-fmcw based contactless respiration detection using acoustic signal[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018, 1(4):170.

[9] YUN S, CHEN Y C, QIU L. Turning a mobile device into a mouse in the air [C]//Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services. [S.l.]: ACM, 2015: 15-29.

[10] MAO W, HE J, QIU L. Cat: high-precision acoustic motion tracking [C]//Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking. [S.l.]: ACM, 2016: 69-81.

[11] ZHANG Y, WANG J, WANG W, et al. Vernier: Accurate and fast acoustic motion tracking using mobile devices[C]//IEEE INFOCOM 2018-IEEE Conference on Computer Communications. [S.l.]: IEEE, 2018: 1709-1717.

[12] OHNO T, MUKAWA N, YOSHIKAWA A. Freegaze: a gaze tracking system for everyday gaze interaction[C]//Proceedings of the 2002 symposium on Eye tracking research & applications. [S.l.]: ACM, 2002: 125-132.

[13] HUANG M X, KWOK T C, NGAI G, et al. Building a personalized, autocalibrating eye tracker from user interactions[C]//Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. [S.l.]: ACM, 2016: 5169-5179.

[14] ZHANG L, LI X Y, HUANG W, et al. It starts with igaze: Visual attention driven networking with smart glasses[C]//Proceedings of the 20th annual international conference on Mobile

computing and networking. [S.l.]: ACM, 2014: 91-102.

[15] MAYBERRY A, HU P, MARLIN B, et al. ishadow: design of a wearable, real-time mobile gaze tracker[C]//Proceedings of the 12th annual international conference on Mobile systems, applications, and services. [S.l.]: ACM, 2014: 82-94.

[16] MAYBERRY A, TUN Y, HU P, et al. Cider: Enabling robustness-power tradeoffs on a computational eyeglass[C]//Proceedings of the 21st Annual International Conference on Mobile Computing and Networking. [S.l.]: ACM, 2015: 400-412.

[17] LI T, LIU Q, ZHOU X. Ultra-low power gaze tracking for virtual reality[C]//Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems. [S.l.]: ACM, 2017: 25.

[18] LI T, ZHOU X. Battery-free eye tracker on glasses[C]//Proceedings of the 24th Annual International Conference on Mobile Computing and Networking. [S.l.]: ACM, 2018: 67-82.

[19] FISCHER B, RAMSPERGER E. Human express saccades: extremely short reaction times of goal directed eye movements[J]. Experimental Brain Research, 1984, 57(1):191-195.

[20] SCHIFFMAN S, H.R., PERCEPTION. Average duration of a single eye blinks[EB/OL]. https://bionumbers.hms.harvard.edu/bionumber.aspx?&i d=100706&ver=4.

[21] TIAN Y, LEE G H, HE H, et al. Rf-based fall monitoring using convolutional neural networks[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018, 2(3):137.

[22] HSU C Y, AHUJA A, YUE S, et al. Zero-effort in-home sleep and insomnia monitoring using radio signals[J]. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 2017, 1(3):59:1-59:18.

[23] YUE S, HE H, WANG H, et al. Extracting multi-person respiration from entangled rf signals[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018, 2(2):86

# ACKNOWLEDGMENTS

First of all, I am very grateful to my supervisor, Professor Jin Zhang, for her careful guidance of my graduation thesis in the past several months, which greatly improved my understanding of research and academic writing skills. Every time when I have problems, she discusses with me enthusiastically and gives her detailed suggestions. Specifically, when I ran into a bottleneck, she always kindly encourages me to overcome the challenges. Besides, she has been supportive of my new ideas and thoughts all the time. Without her help, there is no way that I can finish this graduation thesis on time.

Secondly, I would like to thank my lab members. I have been joined the lab for more than two years, and I can still remember the countless nights that I spent with them struggling on some problems desperately but fulfilling. Here I want to express my special thanks to Miss Kaiyue Zhang who has always been supportive of my final project and we spent a lot of time working late into the night together and encouraging each other. I also want to thank Mr. Runxin Tian for his extensive discussion with me about this final project.

This graduation thesis is not an end of my academic career. The four years of undergraduate study has taught me that there are always too many unknows waiting for me to explore. With what I have learned in writing this graduation paper, I am confident that I will make further progress in my future research and study.