

CochlearMotion: Head Gesture Recognition Leveraging Ear Canal Deformation Sensing

Youngone Lee
Trinity University
ylee5@trinity.edu

Sheng Tan
Trinity University
stan@trinity.edu

Zi Wang
Augusta University
zwang1@augusta.edu

Abstract

Hands-free interfaces have become increasingly popular due to the growing demands for convenient interaction with mobile and wearable devices. Among all of hands-free interfaces, head gesture interaction has shown great potential in providing alternatives in various real-world scenarios, such as interfaces for people with disabilities and Virtual/Augmented Reality applications. However, existing head gesture recognition systems require either Line-Of-Sight or specialized/customized hardware. Additionally, some approaches could raise potential privacy concerns. In this work, we propose CochlearMotion, a novel in-ear wearable system that achieves head gesture recognition by utilizing off-the-shelf earbuds with a built-in microphone. Specifically, we leverage sonar-like techniques to sense the unique deformation of the ear canal, which closely correlated with each head motion for cross-user head gesture recognition. Our extensive experimental evaluation shows that our system can achieve over 95% recognition accuracy for six typical head gestures and works well in various real-world environments and scenarios.

Keywords: wearable, human computer interaction (HCI), head gesture, Internet-of-Things (IoT), mobile computing.

1. Introduction

Up until recently, Human Computer Interactions (HCIs) on mobile devices have been dominated by contact interactions, including touching the screen or pressing physical buttons. Due to technological advancements in hardware and the rapid evolution of ubiquitous computing, a growing number of mobile and

wearable devices, such as smart glasses, the Internet of things (IoT), virtual reality/augmented reality (VR/AR) devices, earables (i.e., in-ear wearables), have been developed. To better facilitate control over these emerging devices, more hands-free interaction approaches have been proposed, including gaze tracking Morimoto and Mimica (2005), voice interaction Rebman Jr et al. (2003), brain wave control Ho and Sasaki (2001), and head posture recognition K. Chen et al. (2022). Among these novel interaction approaches, head gesture recognition has shown great potential in providing alternatives for various real-world applications. For example, people with certain disabilities can leverage head gestures to interact with mobile and wearable devices K. Chen et al. (2022). Moreover, head gestures can be utilized when both hands of the user are not available (e.g., when the user is driving or running). Additionally, such an approach can be used to control head-mounted VR/AR devices, where the head gestures can be a natural way to interact with the system Mustafa et al. (2018).

Much research effort has been dedicated to developing different techniques for head gesture recognition. Traditionally, Computer Vision (CV) based approaches utilize cameras that capture the image of the user's head motion to achieve gesture recognition Mardanbegi et al. (2012). However, such a solution cannot work under Non-Line-Of-Sight (NLOS) scenarios and suffers from performance degradation in poor lighting conditions. Moreover, CV-based approach could raise serious privacy concerns if the image data of the users is not properly managed.

Another body of work leverages motion sensors or Radio Frequency (RF) devices worn by the users to achieve head gesture recognition K. Chen et al. (2022) and Yang et al. (2023). The motion

sensor-based approach mainly relies on sensors such as accelerometers and gyroscopes to infer the user's head motion speed and direction. On the other hand, RF devices (e.g., Radio-frequency Identification (RFID) tags, WiFi transceivers) mounted on the users or the headsets are leveraged to measure the relative distance to the access point to distinguish different head gestures. However, those approaches require specialized hardware, which incurs non-negligible deployment costs. Additionally, some users might be reluctant or feel uncomfortable wearing additional devices.

In this paper, we introduce CochlearMotion, a head gesture recognition system aims to resolve the aforementioned issues by taking advantage of a single earable device. We utilize only the Commercial Off-The-Shelf (COTS) earbud with microphone to infer various head movements. This is achieved by sensing the unique ear canal deformation that is closely associated with distinctive head motions. The proposed system does not require any specialized or additional hardware. Furthermore, our system works well in various real-world environments and scenarios and is unobtrusive to the user during the recognition process.

CochlearMotion exploits the acoustic sensing approach that can detect the unique ear canal deformation caused by head motion. The proposed system utilizes sonar-like techniques that can be implemented using any COTS earbud with a built-in microphone, as shown in Figure 1. To achieve this, the earbud speaker continuously emits an inaudible acoustic signal through the ear canal when the user is performing a head gesture. The inward-facing microphone captures the signal reflections that encompass unique ear canal deformation information, which can be further analyzed to distinguish various head movements. However, accurately discerning the head gestures can be challenging. For example, due to the imperfection of the COTS hardware and the changing environments, the desired signal reflections are mixed with noises from various sources. Without properly addressing, such noises can hamper the later recognition process. Additionally, there exists individual diversity among users, including movement speed, physical characteristics (e.g., different head sizes and ear canal shapes). Even for the same user, they might perform the same head gestures slightly different from time to time due to inconsistency.

To mitigate the potential noises, we propose using a band-pass filter combined with wavelet based denoising techniques to remove out-of-band and high frequency noises while retaining sufficient signal details that can differentiate head gestures. To deal with individual diversity and inconsistency, similar systems

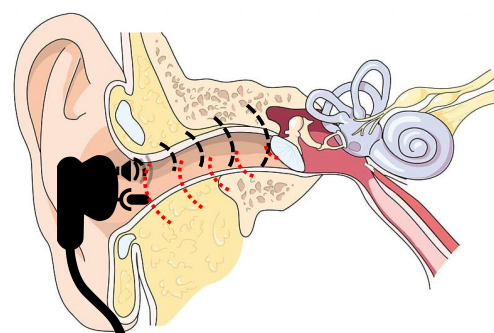


Figure 1. Illustration of CochlearMotion.

usually conduct per-user based profile/model building to avoid the potential impact of the issues discussed before. However, this approach requires individual data collection process and an additional profile/model building phase which is unattainable given limited resources and greatly limit the scalability of the system with large number of users. Therefore, we adopt a CRNN-based learning framework combined with a continuous learning module to achieve cross-user gesture recognition and unsupervised individual model updates after the initial training phase.

We experimentally evaluate CochlearMotion in various environments and scenarios with six typical head gestures: turn left, turn right, look up, look down, clockwise rotation, and counter-clockwise rotation. Results show that our system can achieve overall recognition accuracy of over 95% and is resilient to noises and individual diversity/inconsistency. Additionally, our system can maintain high performance at different locations with real-world use cases. The contributions of our work are summarized as follows:

- We demonstrate that COTS earbuds with built-in microphones can be reused to capture the dynamic deformation information of the ear canal. Such deformation is unique for each head motion and can be used to achieve gesture recognition and can be used to achieve gesture recognition without dedicated or specialized devices, working under various real-world scenarios.
- We propose CochlearMotion, a head gesture recognition system utilizing a single COTS earable device. It leverages wavelet-based denoising technique to mitigate noises caused by hardware imperfections and surrounding environments. Additionally, the proposed system adopts a CRNN-based learning framework combined with a continuous learning module to address individual diversity and inconsistency

issues without requiring per-user training.

- We conduct extensive experiments to evaluate the system's performance. Results show that CochlearMotion can achieve overall accuracy of 95% for six typical head gestures and works well under various environments and scenarios.

2. Related Work

2.1. Head Gesture based Interactions

The problem of expanding the current HCI landscape to accommodate various hardware, users, and scenarios has been well studied. Many solutions have been proposed to enrich existing interfaces through novel interaction methods. Among them, head gestures are considered preferable in many scenarios compared to traditional methods due to their easy-to-use and hands-free features.

For instance, Jia et al. (2007) proposed a control system for intelligent wheelchairs (IWs) based on visual recognition of head gestures. Jackowski et al. (2017) conducted a usability study on adaptive head motion control to increase the autonomy of motion impaired people. By utilizing wearable millimeter-wave radar, Headar developed by Yang et al. (2023) can sense head gestures to enable dialog confirmation interaction on smartwatches. With the growing popularity of VR and AR applications, several studies Mustafa et al. (2018) and Yan et al. (2020) have leveraged head gestures to provide alternative hands-free interactions, authentication and assistance on those devices. Additionally, other work Yi et al. (2016) has explored the potential of using head gestures as a mechanism to interact with the system on smart glasses.

2.2. Head Gesture Sensing Approaches

As a natural and non-verbal communication method, head gestures provide an ideal alternative to traditional HCI mechanisms such as hand gesture, touch, voice control, and so on. Such an approach can be more advantageous compared to others under specific scenarios or for users with special needs.

Traditionally, computer vision-based approaches have been adopted to enable effective head gesture recognition Mardanbegi et al. (2012). With the fast advancement of embedded IMU technologies, motion sensors can be easily integrated into other devices (e.g., smartglasses, earables) and have been widely utilized to distinguish various head movements Yi et al. (2016). Yang et al. (2023) exploited millimeter-wave/infrared signals to achieve head motion sensing. Study has

leveraged electromyography to identify various head gestures Y. Chen et al. (2015).

3. Preliminary

3.1. Ear Canal Deformation

The head movement system has several intrinsic specializations that lead to various side effects compared to other motor systems. The muscles responsible for moving the head have very distinctive characteristics Abrahams (1977). Among all those muscles, the sternocleidomastoid (SCM) muscle is a thick, paired muscle that runs along the front of the neck and behind either side of the ear canal, as shown in Figure 2. Additionally, the ear canal of an adult is a roughly S-shaped elliptical cylinder about 30 mm in length. Due to its location concerning the head, the shape of the ear canal is very sensitive to any head or face motion.

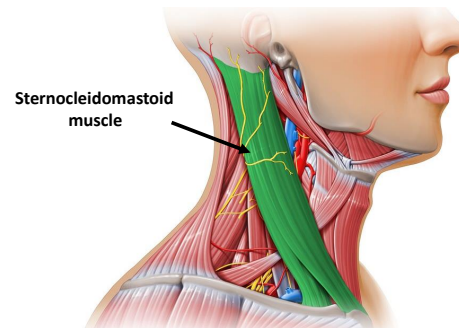


Figure 2. The SCM muscle connected to the ear canal.

In general, when users move their heads, the musculoskeletal also changes, affecting the shape of the ear canal. Previous study conducted by Zhu et al. (2019) has shown that more than 70 facial and neck muscles could affect the geometry of ear canals directly or indirectly. Specifically, when users is moving their heads, the SCM muscle that connects the clavicle to the back of the ear, will either contract or relax depending on the movement. Such contraction or relaxation will compress or expand the ear canal, respectively, eventually causing its deformation.

3.2. Acoustic Sensing Approach

As discussed previously, the ear canal is sensitive to any head movements, which can be utilized to distinguish different head motions. However, measuring the subtle variations in shape can be very challenging due to the complex nature of the ear canal geometry. In particular, the ear canal consists of several distinctive sections where changes in the cross-section can be

difficult to capture directly. Moreover, head motions can cause various deformation of the ear canal in different directions making the combined results difficult to infer.

This motivates us to explore other approaches to capture the ear canal deformation caused by different head motions. In this work, we propose using acoustic sensing to achieve that. Specifically, a speaker will first emit near-ultrasonic signal that can propagate through the entire ear canal when the user is performing a head gesture. The emitted signal will be partially reflected or absorbed by different sections of the ear canal walls and the eardrum. An inward-facing microphone will then be utilized to capture the echoed signals. These signals contain crucial information about the unique ear canal deformation caused by head movements.

4. System Design

4.1. System Overview

Our system utilizes COTS earbud integrated with an in-ward facing microphone for head gesture recognition. Figure 3 shows the system's overview, which comprises four main components: *Acoustic Echo Collection*, *Signal Enhancement*, *Gesture Pattern Extraction*, and *Gesture Recognition*. The recognition process can be activated continuously or on-demand based on the application, shown as *HGI (head gesture interface) Activation* in Figure 3.

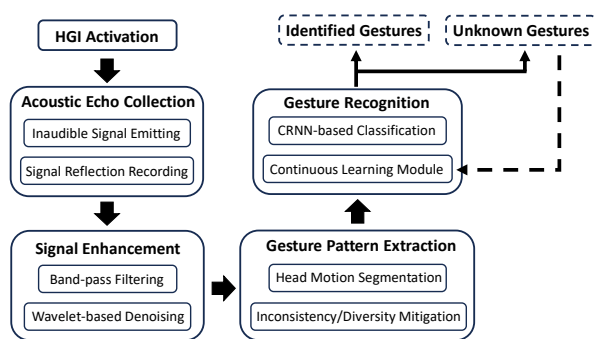


Figure 3. Overview of system flow.

In *Acoustic Echo Collection*, the earbud speaker coupled with the in-ward facing microphone serves as a sonar-like device. Once activated, the speaker emits inaudible acoustic signals, and the microphone collects the signals reflected from the ear canal when the user is performing head gestures.

During *Signal Enhancement*, we leverage a Butterworth filter to remove the out-of-band noises caused by surrounding environments or hardware imperfection. Then, our system employs wavelet-based denoising to further mitigate unwanted noises while

retaining the signal components containing head motion information.

The denoised signals then go through *Gesture Pattern Extraction* for further processing. We first conduct head motion segmentation to separate all potential movement frames using a dynamic threshold. Next, we utilize dynamic time warping techniques to address the diversity and inconsistency issues in segmented samples.

Lastly, our system adopts a CRNN-based learning model in the *Gesture Recognition* component. Additionally, we leverage a continuous learning module to enable the update of existing head motion information and adaptation to the new head gestures.

4.2. Acoustic Echo Collection

To maximize the ability of our system for head gesture recognition, several considerations are needed in designing the probe signal. Firstly, as an interactive system, it is important to make the process transparent to the users. Therefore, we choose a chirp signal with a frequency range above 16kHz since it is inaudible to most adults. Additionally, the audio distortion of the COTS hardware is more serious above 23kHz which narrows the upper bound of the frequency range to 20kHz. Moreover, the chosen frequency range is also sensitive to subtle changes in the ear canal deformation, further enhancing our system's sensing ability. Secondly, we utilize a 192kHz sampling frequency that enables our system to extract more fine-grained head gesture information from the reflected signals. During the process, the earbud speaker continuously emits inaudible signals and the inward-facing microphone captures the signals reflected from the ear canal while the user is performing head gesture.

4.3. Signal Enhancement

4.3.1. Band-pass Filtering Since our system is designed to function in various environments (e.g., indoor or outdoor) and scenarios (e.g., listening to music, answering a phone call, and running) with many noises, we utilize a band-pass filter to remove the out-of-band noises. Considering the frequency band for music and speech ranges from 20Hz to 6kHz and the emitting signal range of our system is from 16kHz to 20kHz, we choose a 2-order Butterworth filter with a pass frequency from 15kHz to 21kHz. This way, we can obtain signal reflections with minimal interference from surrounding environments.

4.3.2. Wavelet-based Denoising Noises come from different sources such as nearby electric devices or speaker and microphone internal noise, cannot be easily removed using a band-pass filter alone. Thus, we proposed using wavelet-based denoising to further mitigate the noises from those sources. This approach is based on Discrete Wavelet Transform (DWT) Tzanetakis et al. (2001) and does not make any assumption about the nature of the signal.

Specifically, we first conduct the signal decomposition recursively by multiple levels (e.g., 3) using DWT, yielding both approximation coefficients and a sequences of detailed coefficients. Next, we apply dynamic thresholding to each level of the detailed coefficients to mitigate the noisy components. Lastly, we reconstruct the final denoised signals by combining the approximation coefficients and the resulting detailed coefficients without the noises together using inverse DWT. The reconstructed signals retain the detailed information of head motions while further mitigating the noise components.

4.4. Gesture Pattern Extraction

4.4.1. Head Motion Segmentation Our system requires the user to have a short pause between performing different head gestures to serve as the sentinel signal. Intuitively, we can segment the motions by detecting a pause followed by a huge spike in the envelope signals. We take the advantage of the fact that the first derivatives of the spikes are high and the first derivatives of the pauses are low and stable. To achieve segmentation, we apply a certain threshold to the first derivatives of the signal. If the results exceed the threshold, it is consider the beginning of a movement. On the other hand, if it is below the threshold for a certain period, it is deemed the end of a movement.

However, it is very difficult to set a suitable threshold for all the users because of movements diversity and environmental uncertainty. Moreover, besides head gestures, other facial or involuntary head movements could also cause spikes. Therefore, the threshold must be sufficiently small to capture all head gesture while large enough to exclude other facial or involuntary head movements. Based on our observation, the intended head gestures usually have a higher entropy level compared to facial or involuntary head movements. Therefore, we leverage a percentile-based dynamic threshold. In particular, we first calculate the intensity distribution of the first derivative of the signal, then based on the empirical study, we set the threshold for the

cutoff (e.g., 70). It is worth noticing that this threshold can be automatically adjusted with new data.

4.4.2. Inconsistency/Diversity Mitigation The head gesture recognition is subject to individual diversity and gesture inconsistency. Different people have different head and ear canal sizes, movement paces and habits of performing head gestures. Additionally, it is possible for users to perform the same gesture slightly differently every now and then because of inconsistency. For example, one user can perform the same gesture much faster/slower or with higher/lower motion intensity compared to the other user. Such individual diversity and gesture inconsistency could seriously lower the recognition accuracy and affect the robustness of our system.

To address the issue of varying motion intensities, our system performs normalization to scale different signals traces into the same range (e.g., 0 to 1). With all the signals traces normalized into same scale, it could effectively mitigate the potential problem caused by users performing head gestures with various intensities. To overcome the problem of users performing head movements at different paces. We propose using Dynamic Time Warping (DTW) to align the signals from various samples into a standardized temporal frame. This is a crucial step when dealing with individual diversity problem. It allows us to overcome pace differences and focus on the shifts in the time series data as evidenced by previous work Muda et al. (2010). Additionally, DTW can help in smoothing the features overtime, reducing noises and variability that might hinder the later recognition process.

4.5. Gesture Recognition

4.5.1. CRNN-based Gesture Classification We build a CRNN-based deep learning framework to identify various head gestures. It has been shown in previous work that convolutional neural networks (CNNs) demonstrate superior capability in dealing with audio signal data Cheng et al. (2020). Because it is necessary to consider the intrinsic sequential characteristics of the signal reflections caused by head motion, we utilize three-layer 1D convolution which demonstrates exceptional performance while incurring much lower computational cost compared to other convolutions. Additionally, as a special type of Recurrent Neural Network (RNN), Long Short-Term Memory network (LSTM) excels in learning long-term temporal dependency features, handling variable-length sequences, and enhancing

other models when combined Cheng et al. (2020). Since it is crucial to learn the dependencies from both directions, we adopt two bidirectional LSTM layers for higher level of abstraction and further enhance the recognition ability. After passing through a fully connected layer with 16 units and an output layer using the *Softmax* function, the result will be shown as either one of the six typical head gestures or unknown gestures that can be later registered by the user.

4.5.2. Continuous Learning Module It is very common for users to adjust the way they perform head gestures due to various reasons such as injury. Additionally, the geometry of the ear canal can also gradually change because of aging. These differences can lead to lower recognition accuracy if the deep learning model stay the same over an extended period. Therefore, we propose including a continuous learning module to tackle this problem. This is done by leveraging unsupervised learning to enrich the training datasets and adapting to new data using Learning without Forgetting (LwF) technique Li and Hoiem (2017). The basic idea of LwF is to continuously modify the CRNN-based model with limited new training data while protecting or maintaining the existing model.

In particular, we first sort the new samples based on their probabilities of being classified as recognized gestures and construct the new training dataset by empirically select top K (e.g., 70%) positive samples with the highest probabilities. This value can be adjusted depending on the model and data. Then, we proceed to update the CRNN-based model using the training dataset built from the previous step. Here, we consider learning the new and old dataset are two different task M and N , respectively. The idea of LwF is to constrain the change of the most important parameters of the task N , which can be done by adding a quadratic penalty on the difference between the loss functions.

5. Evaluation

5.1. Experimental Methodology

Prototype Implementation. Although in-ward facing microphones are getting embedded into many existing commercial earbuds, most manufacturers do not allow direct access to the raw data. Therefore, we developed our prototype system leveraging only COTS hardware, which costs less than 9 dollars for all the parts and materials. Compared to other earbuds with in-ward facing microphones that usually cost around 200 dollars, our system is much more affordable, which can be easily integrated into many existing HCI applications

and adopted by a wider range of users. Specifically, our system uses the most common in-ear earbud found on the market with a 12mm speaker, 3.5mm audio jack, and a microphone chip. We modify the microphone to attach in front of the speaker and kept in the center area of the speaker. To enable the system activate/deactivate operations, we designed a special movement for that, which is opening and closing the mouth multiple times. The ear canal deformation caused by such motion can be easily differentiated from the head gestures. It is worth noticing that the user could also choose their own motions for system activate/deactivate with other facial/head movements.

Head Gestures Selection. Because head gestures are not commonly used compared to traditional hand gestures, we need to carefully design the head movements from both the user and system perspectives. This means the proposed head movements should feel natural and can be easily learned or performed by the user. Additionally, the head motion candidates should be fairly distinguishable from each other. There are two critical considerations in the process of head gestures design.

Human Anatomy Limitation: The head movements have limited freedom in 3D space, especially for users with neck injuries or lack of flexibility (e.g., most human cannot look vertically or horizontally over 180 degrees). Therefore, we need to choose gestures that can be done without requiring extensive ranges of motion to accommodate a wider range of users.

Daily Gesture Conflict: There are already a large set of existing head movements commonly used in daily life. For example, nodding or shaking the head to signify either agreement or disagreement. When we explore the potential head gestures, it is important to select the motions that feel natural but avoid those that can be accidentally triggered by daily activities.

Guided by the principles discussed above, we designed six head gestures for our system. It is worth noticing that users can always add new movements into the system by registering them. Moreover, to better facilitate the recognition process, we designate the neutral position as the starting as well as the ending points for all the movements, which is the natural head position when the user rests the head and facing front/forward without extremely tilting. These head gestures includes turn left, turn right, look up, look down, clockwise rotation, and counter-clockwise rotation shown in Figure 4. To avoid the accidental activation or triggering by other similar daily movements, the first four head gestures are required to be repeated exactly twice consecutively and performed to the user's best range of motion without

feeling uncomfortable.

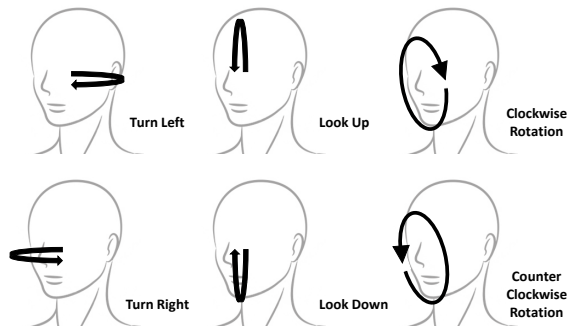


Figure 4. Illustration of six head gestures.

Data Collection. We recruit 20 participants for the experimental evaluation, including 12 females and 8 males, aged 19 to 48 with an average age of 28 years old. The participants were asked to wear our prototype system in their habitual ways during the process. They were informed about the goal of the experiments and given a walk through of the details, including the guidelines for performing head gestures. Then, they were asked to repeat the six head gestures five times while the system emitted probe signals. Because the prototype only supports single microphone, only single ear signal reflections were recorded, and the participants could freely choose either the left or right ear. These samples were used to train the model. After the training phase, each participant was asked to perform eight gesture including the six head gestures and two of their own choices ten times under various scenarios and locations. The experiments were conducted in different locations (e.g., an office, a park, a grocery store, or inside a of vehicle) and scenarios (e.g., the participant is sitting, standing, walking or talk to other people naturally) to simulate real-world use cases. Among all the samples, the training samples were all collected in the office while the user is sitting down without other body motions.

Evaluation Metrics. We use two different metrics to evaluate the performance of our system.

Confusion Matrix: Each column represents the head gesture that was classified by our system, and each row represents the head gesture ground truth performed by the user. Each entry in the confusion matrix represents the percentage of the head gestures in the row that was classified as the gestures in the column.

Recognition Accuracy: The percentage of the head gestures correctly classified/identified by our system.

5.2. Overall Performance

Figure 5 shows the confusion matrix of our system's overall performance. We observe that our system achieves an overall recognition accuracy of over 95% with the standard deviation of around 2%. Among all the gestures, only 2% of the user chosen head gestures have been mistakenly recognized as pre-defined gestures. By comparing the details of each head gesture in the confusion matrix, we find that the classification accuracy distribution is similar across all head gestures. Clockwise and counter-clockwise rotations have the highest recognition accuracy, while looking up and looking down have the lowest accuracy. This is possibly because of the relatively larger range of motion involved in clockwise and counter-clockwise rotation. Consequently, more head motion details could be captured by the signal reflections from the ear canal deformation. The above results demonstrate that our system could provide high accuracy in identifying head gesture using only COTS earbuds. The recognition accuracy could be further improved by using data extracted from both ears.

Left	0.95	0.03	0.01	0.01	0	0
Right	0.01	0.96	0.03	0	0	0
Up	0.02	0.02	0.92	0.02	0	0.02
Down	0.01	0.01	0.03	0.93	0.02	0
Clockwise	0.01	0	0	0	0.98	0.01
Counter-Clockwise	0.01	0.01	0	0	0.01	0.97
	Left	Right	Up	Down	Clockwise	Counter-Clockwise

Figure 5. Confusion matrix of overall head gesture recognition.

5.3. Impact of Training Size

Since the training data size influences the CRNN-based deep learning network's performance, we further evaluate our system using various volumes of data for training. Figure 6 shows the results of the impact of training data size on the performance of our system. Specifically, we study this impact by varying the training data size from 30% to 80% using a 5% interval. Overall, we observe that our system could achieve considerable accuracy with very limited training data size. In particular, with only 30% training data size, our system achieves around 80% recognition accuracy. Moreover, the accuracy is over 90% with only 50% training data size. This result shows that our

system could provide accurate head gesture recognition with limited training data and hence doesn't incur high overhead in training data collection, especially considering the training is done across all users instead of per-user.

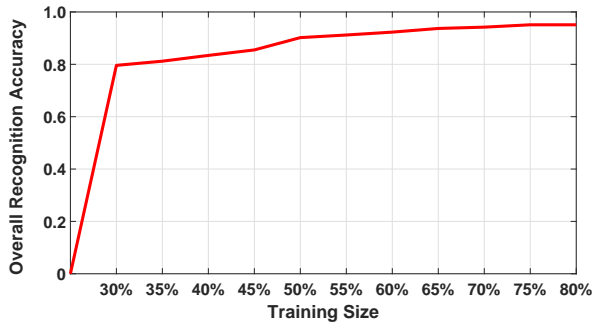


Figure 6. System performance under different training size.

5.4. System Performance under Different Scenarios

5.4.1. Environmental Noise Impact Study It is natural that users could wear our system at different locations with various environmental or background noise levels. Therefore, it is important to study such impacts on the performance of our system. We chose the following locations to represent different background noise levels: office, inside of a vehicle (the vehicle was not in motion but the engine was on), park and grocery store. The noise levels are 45 dB, 60 dB, 55 dB, and 95 dB respectively. We have collected 800 samples (200 samples for each environment) from 10 volunteers who participated in this study. The results are shown in Figure 7. It is easy to observe that our system can achieve similar recognition accuracy in various environments. Even at the grocery store location, where the noise level can be over 90 dB, our system can still maintain high recognition accuracy around 90%. This demonstrates the robustness of our system against background noises and its usability across various environments.

5.4.2. Body Motion Impact Study When users are using our system for head gesture recognition, it is very likely that some voluntary or involuntary body motions will be involved during the process. We consider four scenarios with various body motion intensities and studied their impacts on our system's performance: minimal (i.e., sitting), low (i.e., standing), medium (i.e., turning the head to engage in conversation), and high

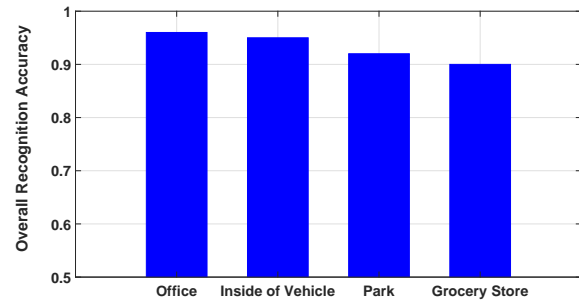


Figure 7. System performance at different locations with various background noises.

(i.e., walking or jogging casually). We have collected 600 samples (150 samples for each body motion level) from 5 volunteers who participated in this study. As shown in Figure 8, we can observe that our system works well across various ranges of motions. In particular, the overall recognition accuracy is above 90%. Although, the recognition accuracy is slightly lower in the last category due to the larger range of motion and higher intensity level compared to others, our system can still maintain high recognition accuracy and resilient to different body motions.

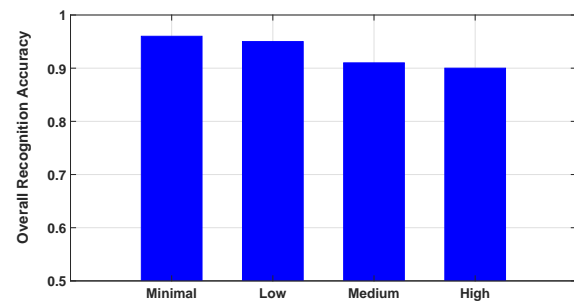


Figure 8. System performance under different body motion levels.

5.4.3. Wearing Position Impact Study The captured signal reflections caused by the ear canal deformation from the same user and head gesture could still vary slightly from time to time due to the variation in the wearing position. Next, we evaluate the impact of different wearing positions of the earbud on our system's performance. We conducted the study by asking the participants to wear the device with three different rotation degrees compared to their habitual position: 10°, 20°, and 30°. The measured degree is facing towards the earlobe with respect to the original position. We have collected 300 samples (100 samples for each

wearing angle) from 2 volunteers who participated in this study. The results are shown in Figure 9. Overall, our system achieves comparable performance with different rotation degrees. Specifically, the accuracy for both 20°, and 30° rotation angle is over 93%. This demonstrates our system is insensitive to the variation in the wearing position.

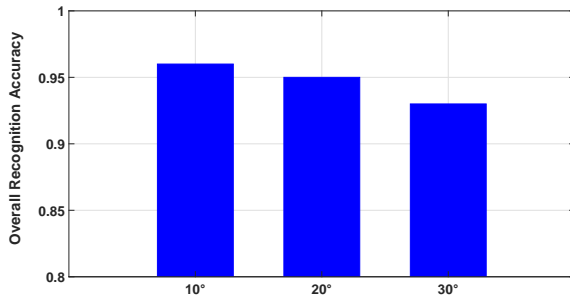


Figure 9. System performance with different wearing angles.

5.4.4. Long-term Performance Study It is important that our system can maintain high performance over an extended period. To evaluate this, we study our system's long-term performance over three months. In particular, we evaluated the system after the periods of 1, 2, 4, 8, and 12 weeks, respectively following the initial training process. We have collected 700 samples (200 samples each for 2 and 4 weeks, and 150 samples each for 8 and 12 weeks) from 4 volunteers who participated in this study. We show the recognition accuracy of our system with and without applying the continuous learning module in Figure 10. We observe that our system maintains high performance even without the continuous learning module, but improves rapidly after 8 weeks with the continuous learning module. This is because the continuous learning module can help the system adapt and make use of more individual data collected after initial training phase. These results show the consistency of our system over the long term and effectiveness of the continuous learning module.

6. Discussion

6.1. Safety of Ultrasound-based Sensing

Existing studies have shown long-term, persistent exposure to high intensity ultrasound can lead to potential damage to human body Moyano et al., 2022. According to the recommendations of the Centers for Disease Control and Prevention (CDC) Murphy, 2018,

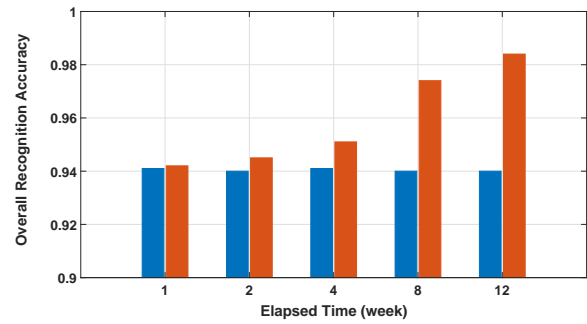


Figure 10. Long-term system performance.

a person can continuously be exposed to 85 dB over 8 hours or 70 dB over a 24-hour time period. The power level of our system is less than 60 dB at medium volume when activated, which is less than both recommendation limits. Additionally, the actual average exposure is likely much lower because the device will only be used under the scenario where tracking is needed. Therefore, we believe our system would not cause adverse effects to user health. But we would still recommend users take breaks after several hour of usage and lower the volume to prevent potential health risk.

6.2. Comparison with Existing Work

For the task of head gesture recognition, prior studies have primarily utilized IMU (Inertial Measurement Unit) sensors Yi et al., 2016 such as accelerometers and gyroscopes that embedded in head-mounted devices or cameras Mardanbegi et al., 2012 that directly point to the user's head region. Recently, mmWave (millimeter-wave), infrared Yang et al., 2023 and EMG (Electromyography) Y. Chen et al., 2015 worn by the user have also been used to achieve head gesture recognition. Compared to systems utilizing IMU, mmWave, and EMG, our approach leverages only COTS earphone, which do not require specialized hardware. Therefore, it has a much lower cost and can be easily adopted by users with different needs. Additionally, our system can work under NLOS conditions and does not have the privacy concerns associated with camera-based approaches, while maintaining high recognition accuracy across various scenarios.

7. Conclusion

In this work, we propose CochlearMotion, a head gesture recognition system utilizing a single COTS earable device to sense the unique canal deformation caused by each head gesture. We leverage wavelet-based denoising to mitigate noises

stemming from various sources. We also adopt CRNN-based learning framework combined with a continuous learning module to address the individual diversity and inconsistency issue without requiring per-user training. Extensive experiments show that our system is highly accurate in recognizing six typical head gestures. Additional results also show that the system performs well under various environments and scenarios and can be easily utilized by real-world applications.

References

- Abrahams, V. (1977). The physiology of neck muscles; their role in head movement and maintenance of posture. *Canadian journal of Physiology and Pharmacology*, 55(3), 332–338.
- Chen, K., Wang, F., Li, M., Liu, B., Chen, H., & Chen, F. (2022). Headsee: Device-free head gesture recognition with commodity rfid. *Peer-to-Peer Networking and Applications*, 15(3), 1357–1369.
- Chen, Y., Yang, Z., & Wang, J. (2015). Eyebrow emotional expression recognition using surface emg signals. *Neurocomputing*, 168, 871–879.
- Cheng, Y.-H., Chang, P.-C., Nguyen, D.-M., & Kuo, C.-N. (2020). Automatic music genre classification based on crnn. *Engineering Letters*, 29(1).
- Ho, C. K., & Sasaki, M. (2001). Brain-wave bio potentials based mobile robot control: Wavelet-neural network pattern recognition approach. *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236)*, 1, 322–328.
- Jackowski, A., Gebhard, M., & Thietje, R. (2017). Head motion and head gesture-based robot control: A usability study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(1), 161–170.
- Jia, P., Hu, H. H., Lu, T., & Yuan, K. (2007). Head gesture recognition for hands-free control of an intelligent wheelchair. *Industrial Robot: An International Journal*, 34(1), 60–68.
- Li, Z., & Hoiem, D. (2017). Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12), 2935–2947.
- Mardanbegi, D., Hansen, D. W., & Pederson, T. (2012). Eye-based head gestures. *Proceedings of the symposium on eye tracking research and applications*, 139–146.
- Morimoto, C. H., & Mimica, M. R. (2005). Eye gaze tracking techniques for interactive applications. *Computer vision and image understanding*, 98(1), 4–24.
- Moyano, D. B., Paraiso, D. A., & González-Lezcano, R. A. (2022). Possible effects on health of ultrasound exposure, risk factors in the work environment and occupational safety review. *Healthcare*, 10(3), 423.
- Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*.
- Murphy, W. J. (2018). Cdc grand rounds: Promoting hearing health across the lifespan. *MMWR. Morbidity and Mortality Weekly Report*, 67.
- Mustafa, T., Matovu, R., Serwadda, A., & Muirhead, N. (2018). Unsure how to authenticate on your vr headset? come on, use your head! *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, 23–30.
- Rebman Jr, C. M., Aiken, M. W., & Cegielski, C. G. (2003). Speech recognition in the human-computer interface. *Information & Management*, 40(6), 509–519.
- Tzanetakis, G., Essl, G., & Cook, P. (2001). Audio analysis using the discrete wavelet transform. *Proc. conf. in acoustics and music theory applications*, 66.
- Yan, Y., Shi, Y., Yu, C., & Shi, Y. (2020). Headcross: Exploring head-based crossing selection on head-mounted displays. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 4(1), 1–22.
- Yang, X., Wang, X., Dong, G., Yan, Z., Srivastava, M., Hayashi, E., & Zhang, Y. (2023). Headar: Sensing head gestures for confirmation dialogs on smartwatches with wearable millimeter-wave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3), 1–28.
- Yi, S., Qin, Z., Novak, E., Yin, Y., & Li, Q. (2016). Glassgesture: Exploring head gesture interface of smart glasses. *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, 1–9.
- Zhu, M., Huang, Z., Wang, X., Zhuang, J., Zhang, H., Wang, X., Yang, Z., Lu, L., Shang, P., Zhao, G., et al. (2019). Contraction patterns of facial and neck muscles in speaking tasks using high-density electromyography. *2019 13th International Conference on Sensing Technology (ICST)*, 1–5.