EarIO: A Low-power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements

KE LI, Cornell University, USA
RUIDONG ZHANG, Cornell University, USA
BO LIANG, Peking University, China
FRANÇOIS GUIMBRETIÈRE, Cornell University, USA
CHENG ZHANG, Cornell University, USA

This paper presents EarIO, an AI-powered acoustic sensing technology that allows an earable (e.g., earphone) to continuously track facial expressions using two pairs of microphone and speaker (one on each side), which are widely available in commodity earphones. It emits acoustic signals from a speaker on an earable towards the face. Depending on facial expressions, the muscles, tissues, and skin around the ear would deform differently, resulting in unique echo profiles in the reflected signals captured by an on-device microphone. These received acoustic signals are processed and learned by a customized deep learning pipeline to continuously infer the full facial expressions represented by 52 parameters captured using a TruthDepth camera. Compared to similar technologies, it has significantly lower power consumption, as it can sample at 86 Hz with a power signature of 154 mW. A user study with 16 participants under three different scenarios, showed that EarIO can reliably estimate the detailed facial movements when the participants were sitting, walking or after remounting the device. Based on the encouraging results, we further discuss the potential opportunities and challenges on applying EarIO on future ear-mounted wearables.

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

Additional Key Words and Phrases: Facial expression reconstruction, Acoustic sensing, Low-power, Deep learning

ACM Reference Format:
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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 62 (June 2022), 24 pages. https://doi.org/10.1145/3534621

1 INTRODUCTION

Humans use facial movements/expressions to interact with the world by conducting physical activities (e.g., eating, drinking), expressing feelings (e.g., emotion [29]) and communicating non-verbal information (e.g., silent speech recognition [43]). Thus, in order to interpret human behaviors in detail, the first step is to continuously record the detailed facial movements at any location and any time. Existing facial movements tracking technologies are largely based on using a frontal camera to capture the complete face of the user, which is not portable. To address this challenge, multiple wearable-based facial expression recognition technologies have been developed. However, most of them can only recognize discrete facial gestures instead of tracking the full facial movements continuously.

Authors' addresses: Ke Li, Cornell University, Ithaca, USA, kl975@cornell.edu; Ruidong Zhang, Cornell University, Ithaca, USA, rz379@cornell.edu; Bo Liang, Peking University, Beijing, China, rambo@pku.edu.cn; François Guimbretière, Cornell University, Ithaca, USA, fvg3@cornell.edu; Cheng Zhang, Cornell University, Ithaca, USA, chengzhang@cornell.edu.

2474-9567/2022/6-ART62 $15.00
https://doi.org/10.1145/3534621

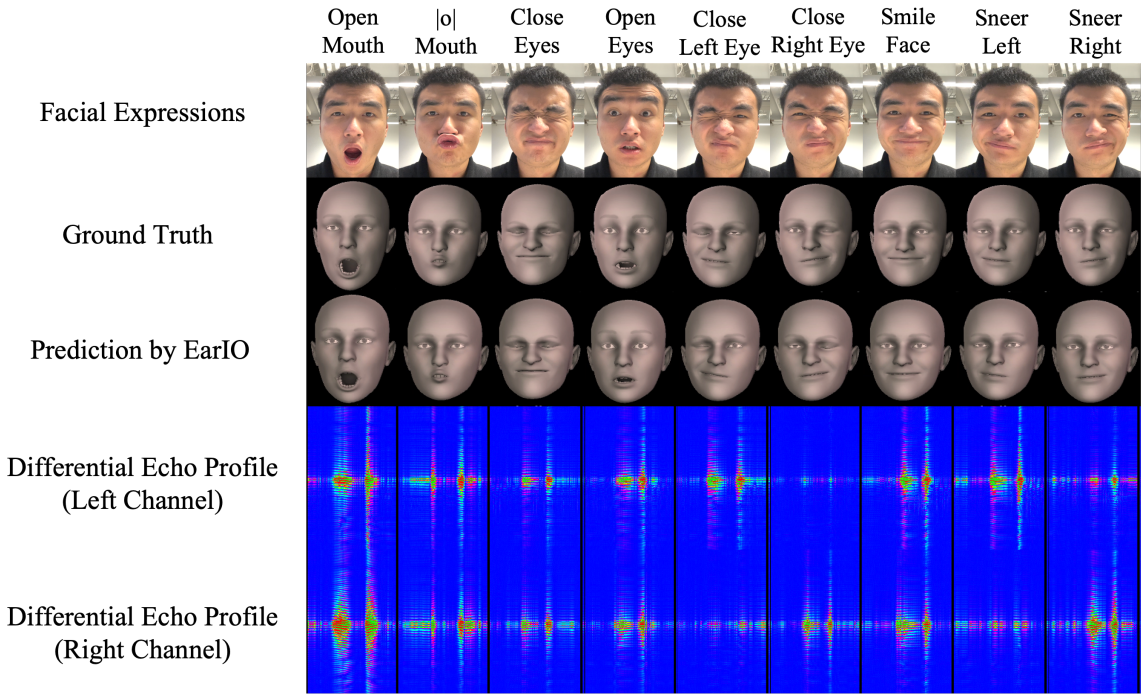


Fig. 1. Designed Facial Expressions and Corresponding Differential Echo Profiles

The most recent work showed the feasibility of continuously tracking facial movements using cameras [4, 5] or bioSensors (e.g., EOG/EMG) [38]. Despite promising tracking performance, it is hard to apply these technologies on a daily wearable device due to their cumbersome dimension (e.g., [38] requires electrodes to be attached on face), or high power consumption [4, 5]. As a result, it is currently impossible to apply these sensing systems on a commodity earable device (e.g., earphone), which has very limited size, sensing modality, and battery life. There is a clear need for a low-power and minimally-obtrusive sensing method that can potentially be embedded on earables without compromising the performance on sensing facial movements.

We present EarIO, the first very low-power and minimally-obtrusive active-acoustic sensing technology, that uses the machine learning method to continuously infer detailed full facial movements from the subtle skin deformations. Our technology only requires one speaker and one microphone on each side of the earable, which are widely available on many modern earphones (e.g., Apple PowerBeats Pro[1]). The speaker on each earable emits encoded acoustic signals (Frequency-Modulated Continuous-Wave, FMCW) towards the user's face. These acoustic signals are reflected differently to a microphone based on the skin deformation associated with different facial movements. The received acoustic signals are first decoded and processed to extract the echo profiles around the face, which are then learned by a customized deep learning model to estimate the full facial expressions. Because our system only requires a pair of microphones, a pair of speakers and a BLE module, it can sample facial movements at 86 Hz with a power signature as low as 154mW, which is 25 times lower than the previous wearable camera-based sensing system [4]. Using a LiPo battery of 110 mAh , our system can last for about 3 hours. A user study with 16 participants showed that EarIO can continuously and accurately estimate facial

[1]https://www.beatsbydre.com/earphones/powerbeats-pro

movements under a variety of daily scenarios, including sitting, walking, and remounting the device. Besides, we also discuss the challenges and opportunities of applying this novel sensing technology on commodity earables, such as earphone, headphone, and glasses.

We summarize the contributions of the paper below:

- We present the first acoustic-based earable sensing method that can estimate the full facial movements continuously from subtle skin deformations.
- We designed a minimally-obtrusive form factor and optimized the power consumption of the system, so that it can operate at 86 Hz with a lower power signature of 154 mW.
- We conducted a user study of 16 participants, of which the results showed that our system can accurately estimate the facial movements under different scenarios including sitting, walking and after remounting the device.
- We discussed the opportunities and challenges on applying this low-power sensing technology on the next generation of earables.

2 RELATED WORK

In this section, we examine previous research projects that are related to EarIO in the following scopes: 1) frontal camera-based facial expression tracking methods, 2) wearable-based facial expression tracking technologies, and 3) acoustic-based wearable human activity tracking systems.

2.1 Frontal Camera-based Facial Expression Tracking

Frontal cameras have been widely used as the major approach to capture human facial expressions by capturing the face of the user. These methods usually deploy cameras [11, 28, 30, 32, 36] in front of the user's face and utilize various computer vision (CV) technologies [6, 9, 12, 14, 16, 19, 20, 27] for facial expression representation such as 2-dimensional (2D) facial landmarks tracking [15, 37]. Recent advancement in CV and AR/VR technologies makes more accurate and subtle 3D tracking possible. For instance, Apple's ARKit provides blendshape-based 3D facial expression representation and visualization API using its TrueDepth camera. We utilize this technology as our ground truth acquisition method which we will elaborate in Subsubsec. 4.4.1.

These frontal camera-based methods have provided reliable tracking performance and have been widely used as ground truth by other facial expression tracking systems [4, 5, 38]. However, they require the users to be present in front of a camera at all times, which clearly does not work when the user is in motion. Furthermore, the performance can be significantly impacted by issues like light conditions, occlusions, etc. The high-energy consumption of cameras makes it challenging for them to be deployed on commodity earables, which have extremely limited battery life and processor power.

2.2 Wearable Facial Expression Tracking

To conquer the challenges that frontal facial reconstruction methods encounter, researchers have put their efforts into developing wearable devices to estimate facial expressions. These methods usually focus on attaching sensors on users' head to track physical or bio signals while users perform different facial expressions. These signals include electromygraphy (EMG) and/or electrooculography (EOG) [8, 38], skin deformation [3–5, 17] or eye movement [10] tracked by cameras, acoustic [13, 39], or capacitive [26] sensors, and canal deformation tracked by barometers [1] or motion sensors [33]. However, most of these technologies can only distinguish several discrete facial expressions instead of continuously tracking detailed facial movements. The most recent work (C-Face [5], NeckFace [4] and BioFace-3D [38]) have shown the possibility of estimating facial expressions continuously on neck- or ear-mounted wearables. However, C-Face and NeckFace use cameras which are not only power-hungry (e.g., NeckFace operates at 4W, 25 times higher than EarIO) but potentially raise privacy concerns by capturing

images of the user and the surrounding environment. In contrast, BioFace-3D consumes relatively lower energy but requires attaching multiple electrodes on the face with a relatively large form factor, making it potentially uncomfortable to wear in daily activities.

Apparently, there is a clear need for a very low-power and minimally-obtrusive earable sensing technology that can estimate full facial expressions continuously in daily activities.

2.3 Active Acoustic Sensing and its Applications in Wearable Human Activity Tracking

Being contact-less, low-power and ubiquitously available on commodity devices, active acoustic sensing has been used in a variety of wearable computing applications. Active acoustic sensing technologies have been demonstrated in applications such as motion tracking [18, 45], silent speech recognition [21, 44], gesture tracking [2, 23, 31, 35, 40–42], breathing detection [34], sleep apnea detection [22], etc. However, using active acoustic sensing to estimate facial expressions is under-explored. The closet work is conducted by Xie et al. [39] demonstrating the feasibility of using acoustic sensors on a smart eyewear to recognize 6 discrete upper facial expressions.

To the best of our knowledge, EarIO is the first earable sensing technology that uses active acoustic sensing to estimate full facial movements from subtle skin deformations continuously. It only needs one pair of speaker and microphone on each side, which are already widely available on commercial wearable devices. Compared with previous work, EarIO provides reliable and high-resolution facial movement tracking performance under different daily scenarios (walking, sitting, remounting) with a very low power signature (86 Hz and 154 mW).

3 THEORY OF OPERATION

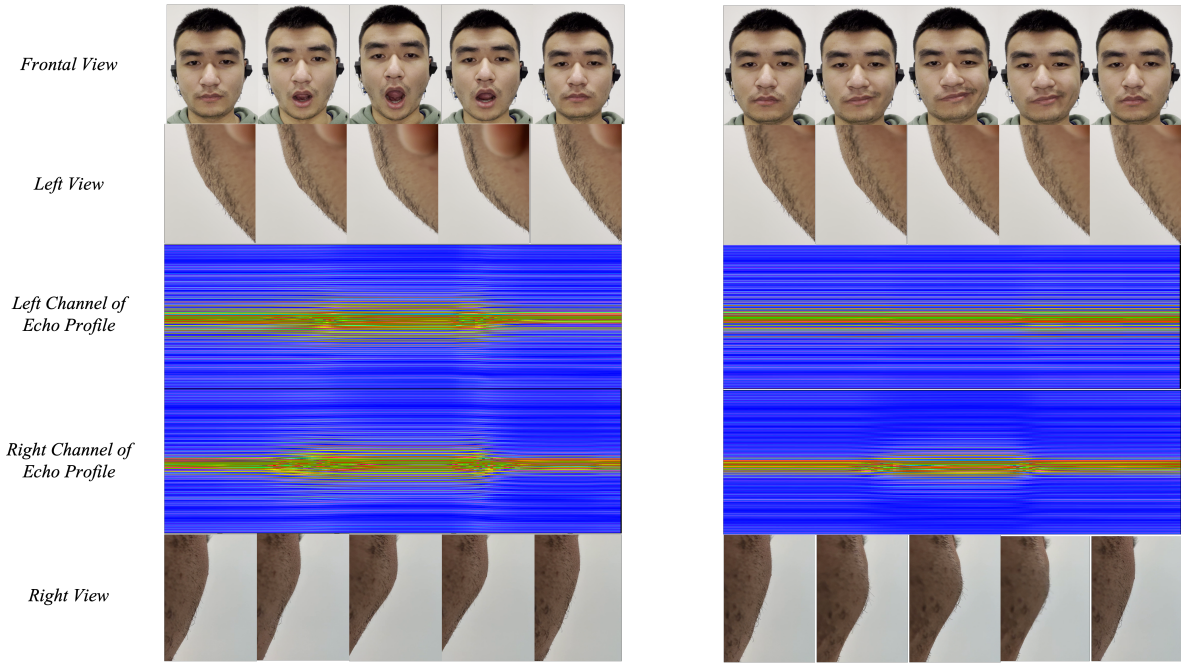
The goal of EarIO is to provide a low-power and minimally-obtrusive sensing method for earables to track detailed facial movements continuously. It is inspired by C-Face [5], which first demonstrated promising performance to infer the full facial expressions from facial contours captured by miniature cameras on earables. However, deploying cameras on earables can be very challenging, because cameras have relatively large size compared to the size of earables (e.g., earphone), consume a significant amount of energy, and require high bandwidth to transmit and process the data in real-time.

EarIO intends to build on the sensing principle demonstrated by C-Face, that the subtle skin deformation can be highly informative to infer the full facial expressions. Instead of using cameras to capture the contour of the face, EarIO proposes to replace cameras with active acoustic sensing units to capture the subtle skin deformations. Using active acoustics to sense the shape or distance of an object has been previously demonstrated on other applications, such as motion tracking [18, 45] and breath detection [34]. Compared to cameras, acoustic hardware (e.g., microphone and speaker) are not only widely available on commodity wearables, but also have much smaller dimension, generate smaller size of data, and consume significantly lower power.

To demonstrate the feasibility of such sensing principle, one researcher performed several facial expressions while attaching a pair of speaker and microphone on each side of earphones. We recorded the skin deformation as captured by cameras and by microphones. We then used one of the methods (Echo Profile on FMCW signal, detailed in later sections) that we explored to analyze the echos. We illustrate such process in Fig. 2. It is clear that, with our data analysis method, the skin deformations during facial expressions lead to clear pattern changes in the captured acoustic echos.

4 DESIGN AND IMPLEMENTATION OF EARIO

In this section, we present the design and implementation on the hardware and algorithms of EarIO. The design principles of EarIO are to design the minimally-obtrusive form factor and optimize the power consumption in the whole system including sensors and data transmission. As a result, the EarIO prototype uses customized printed circuit boards (PCBs) to house the micro-controller, microphones and speakers to optimize the device size



(a) Patterns of Open Mouth (b) Patterns of Snear Right

Fig. 2. Verifying the Theory of Operation of EarIO

and the power consumption. A customized deep learning algorithm is developed to learn the skin deformation patterns from the Echo Profiles extracted from FMCW-based acoustic signals. We detail each component of EarIO in the following subsections.

4.1 Hardware Prototype

In order to minimize the size and optimize the power consumption, we designed our own form factor, fine-tuned the sensing solutions, and customized the PCB for our prototype.

4.1.1 Form Factor Design. The goal of the form factor design is to demonstrate that EarIO can be embedded to a form factor that is highly similar to the commodity ear-mounted wearables. We started by 3D printing the whole earable, with a shape looking similar to PowerBeats Pro. However, we realized that 3D printed materials are relatively rigid and hard to fit different people's ear size. Therefore, instead of 3D printing all parts, we decided to customize the commodity earphones to house the hardware components (battery, micro-controller, sensors) so that we could take advantage of the form factor on commodity earphones which are designed to fit ears with different sizes. We first designed a 3D printed supporting board, where we attached the microphone and speaker on. The 3D printed board was glued to the side of the commodity earphone. Then we replaced the electronic components from the earphone with our customized PCBs (except the battery). As demonstrated in Fig. 3, our final prototype is highly similar to the original earphone except a small extension with the core acoustic sensing unit.

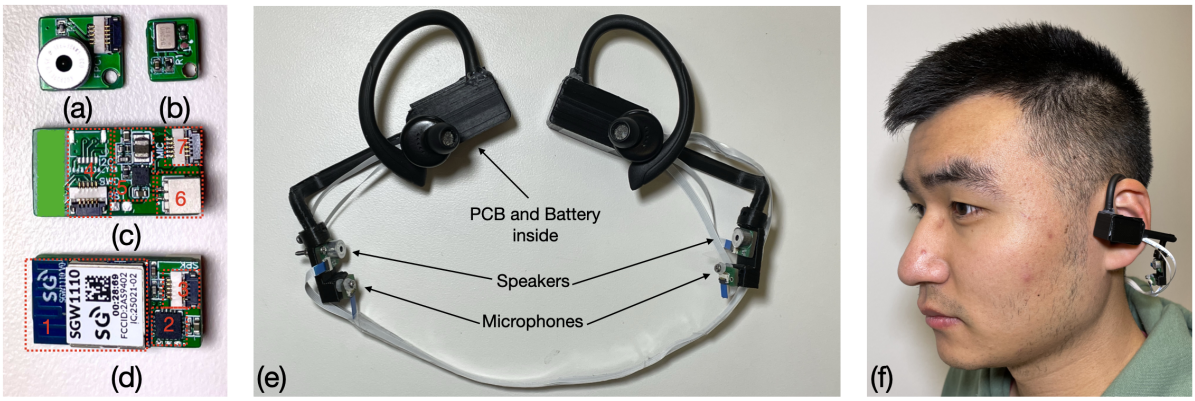


Fig. 3. Customized PCBs and Form Factor of EarIO. (a) Speaker board. (b) Microphone board. (c)-(d): Bottom and top layers of the PCB. (e) Form factor design of EarIO. (f) EarIO as an earable. Components on the customized PCB: (1) SGW1110 BLE module, (2) MAX98357A audio amplifier, (3) Speaker socket, (4) Downloading and debugging port, (5) TPS62065 switch regulator module, (6) Battery socket, (7) Microphone socket.

4.1.2 The Wired Prototype. In order to validate our system as well as providing more flexibility in exploring various configurations, we first designed a wired prototype. The core sensing unit of EarIO consists of a MEMS microphone (ICS-43434[2]) and a speaker (OWR-06944T-16B[3]). We designed a customized PCB with 2 SGTL5000 audio codecs[4] for support of connecting up to 4 speakers and 4 microphones. This customized PCB is connected to the speakers and microphones in our system and then communicates data to the micro-controller (1 Teensy 4.1[5]) through the Inter-IC Sound (I2S) interface operating at 44.1 kHz. An on-board SD card is used to save all acoustic data on the micro-controller.

4.1.3 The Wireless Prototype. The wired prototype is great for verifying the sensing principle. However, in order to simulate the operating environment for a commodity earphone, we implemented a wireless prototype which transmits acoustic data to a smartphone in real-time. In order to minimize the power consumption, we chose Bluetooth Low-Energy (BLE) as our wireless data transmission solution. We use an SGW1110 [6] module with nRF52840 [7] micro-controller for BLE communication, acoustic signal transmission and reception. The micro-controller is connected to two MEMS microphones directly (ICS-43434) and drives two speakers (SR6438NWS-000 [8]) via a MAX98357A audio amplifier [9]. Similar to our wired version, all amplifiers and microphones are connected to the same I2S interface using the clocks provided by the hardware PWM module of the micro-controller, which guarantees that their sampling rates are exactly the same. Due to limited options for clocks, we were not able to obtain a 44.1 kHz or 48 kHz sampling rate. Instead, we configured the system to operate at 50 kHz. The wireless data collection system was implemented on a customized miniature PCB (10.4 mm × 20 mm), powered by a small

[2]https://invensense.tdk.com/products/ics-43434/
[3]https://owolff.com/page140.aspx?recordid140=1278&output=pdf&delay=3000&margin=1cm
[4]https://www.nxp.com/products/audio-and-radio/audio-converters/ultra-low-power-audio-codec:SGTL5000
[5]https://www.pjrc.com/store/teensy41.html
[6]https://www.sgwireless.com/product/SGW111X
[7]https://www.nordicsemi.com/Products/nRF52840
[8]https://www.knowles.com/docs/default-source/model-downloads/sr6438nws-000.pdf?Status=Master&sfvrsn=212b75b1_0
[9]https://www.maximintegrated.com/en/products/analog/audio/MAX98357A.html

LiPo battery. The PCB has an on-board switch-regulator TPS62065[10] for high-efficiency power management. The FMCW sequence length was changed to 600 optimized for BLE transmission.

We evaluated EarIO in both wired and wireless prototypes in our user study, which will be detailed in the later sections.

4.2 FMCW-based Acoustic Sensing for Detecting Skin Deformations

4.2.1 Comparing Active-acoustic Sensing Methods. In order to estimate distance or human activities, researchers have developed various acoustic-based sensing approaches including phase changes [31, 35, 45], angle of arrival (AoA) [7], Doppler effect [2, 21], channel impulse response (CIR) [41, 44], and FMCW [22, 34]. Among these techniques, we compared three types of acoustic sensing signals, which have demonstrated promising performance on other work. They are 1) 20 kHz chirps, 2) CIR with GSM sequence [41, 44], 3) FMCW [22, 34]. In order to evaluate the performance on estimating facial expression from skin deformations, we compared these three types of signals in a preliminary study, where one researcher collected similar facial movement data for each type of signals in different scenarios. These collected data were decoded and processed first, then sent to the same deep learning algorithm to estimate the full facial movements. The details of the algorithms will be described in the next subsections. The results in Tab. 1 evaluated with the metrics introduced in Subsubsec. 4.4.3 showed that FMCW presented the best performance among the three methods, especially in cross-session scenarios where users remount the device. Thus, we decided to use FMCW in EarIO to sense the skin deformations.

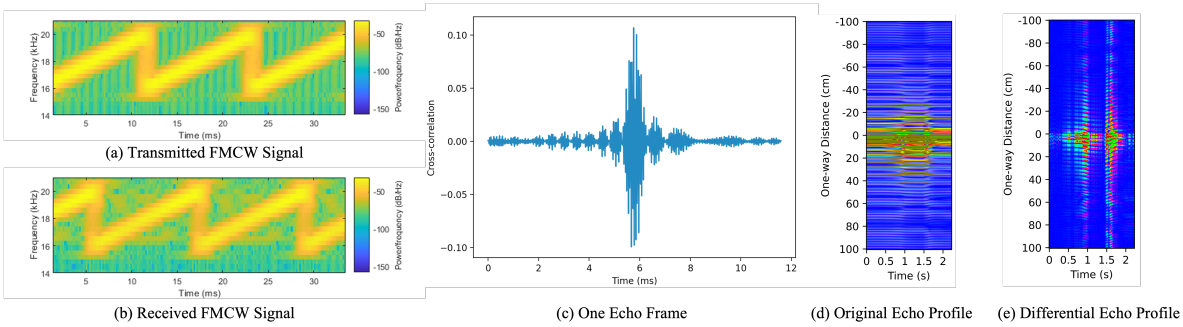
Table 1. Performance of Three Different Transmitted Signals

Transmitted Signal	In-Session					Cross-Session				
	MAE	LMAE	UMAE	PL40	PU60	MAE	LMAE	UMAE	PL40	PU60
Chirp	25.6	18.7	37.6	91.5%	83.2%	27.7	19.6	41.7	90.1%	79.0%
GSM	20.7	16.1	**28.8**	93.7%	**89.6%**	23.8	17.0	35.6	92.0%	85.2%
FMCW	**20.6**	**14.7**	30.9	**96.6%**	88.3%	**19.2**	**12.5**	**30.9**	**95.4%**	**88.0%**

4.2.2 Generating FMCW Signals. Since EarIO uses an active acoustic sensing method, we need to first generate FMCW signals for the speaker. The frequency of FMCW changes with time linearly. Hence, to generate an FMCW signal, we need to determine a frequency range that the signal would be sweeping within. Because we would like to deploy the system on an ear-mounted device and track users' facial expressions in their daily lives, we chose the frequency range 16-20 kHz which is inaudible to most people and also supported by most commodity microphones and speakers. As a result, our EarIO system generates the FMCW signal, sweeping from 16-20 kHz at a sampling rate of 44.1 kHz. To achieve a reasonable resolution of tracking facial expressions, we set the period of FMCW to be 512 samples which is around 11.6 ms (512 samples / 44.1 kHz). In this case, we have around 86 sweeps per second. In the wireless prototype, the period is changed to 600 samples in accordance with its 50 kHz sampling rate and BLE transmission considerations. However, The period of FMCW can be potentially revised to lower the estimation sampling rate of our system to save battery. The spectrogram of the transmitted FMCW signal that we generated is shown in Fig. 4 (a).

4.2.3 Calculating Echo Profile from FMCW Signals. The FMCW signal is fed into the speaker, which emits the signals towards the face. A microphone on the same board with the speaker receives the reflected acoustic signals (echos) via multiple paths. EarIO analyzes the reflected signals by performing cross-correlation between echos and the transmitted signals, which reflects the distance of the reflecting medium to the sensors [34]. For targets

[10]https://www.ti.com/product/TPS62065



(a) Transmitted FMCW Signal
(b) Received FMCW Signal
(c) One Echo Frame
(d) Original Echo Profile
(e) Differential Echo Profile

Fig. 4. Overview of the Transmitted and Received Signals and Echo Profile Analysis

that do not move too fast, the cross-correlation has the same period as the transmitted signal [22, 34]. We denote the cross-correlation in such a period as an *echo frame*. Using cross-correlation-based FMCW as descried in [34], a 44.1 kHz sampling rate allows us to reach a resolution of around 0.77 cm (340 m/s / 44.1 kHz), with a maximum theoretical range of 395 cm [34]. Because the speakers and microphones are co-located, the resolution is actually the minimum round-trip distance that the system is capable of distinguishing. Therefore, the one-way resolution we can get is about 0.385 cm with 197 cm range, which we believe is enough to detect even subtle movements of facial muscles. With sequence length being 512 samples, an echo frame is a 1024-dimension vector (2 channels stacked together). However, only a smaller portion of them contain useful information about the face since the distance between our device and the skin is usually smaller than 15 cm. Furthermore, by limiting the echo profile to 15 cm, we can remove reflections from objects further away which are useless to our system. As a result, we clip an echo frame to 200-dimension to focus on objects at a closer distance. By repeatedly sending FMCW frames, we could get a figure of multiple consecutive echo frames that vary in time. This figure that demonstrates the change of echo frames in the time domain is called *Echo Profile*. Using similar methods as described in [34], we calculate the zero-position using the direct path, which is the shortest and strongest path where the signal travels in the 3D-printed board.

We display the Echo Profiles for two different facial expressions, *Open Mouth* and *Sneer Right* in Fig. 2. As is shown in Fig. 2, the Echo Profiles for different facial expressions are visually different. For instance, if we only move one side of our face, like Fig. 2 (b), we can see a stronger pattern on the right channel of our system. There are also some patterns in the left channel because the muscles on our face are interconnected and when we try to sneer right, the left side will also move, but less significantly than the right side.

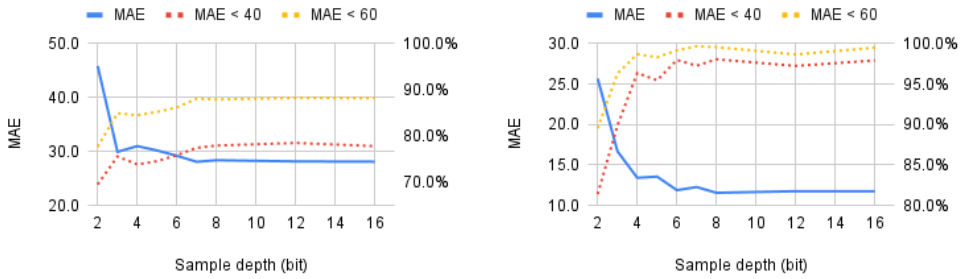
In addition to using the original Echo Profile, we calculate the Differential Echo Profile by calculating the difference between two consecutive Echo Profiles. As Fig. 4 (e) shows, the Differential Echo Profile has higher signal-to-noise ratio (SNR) because it subtracts the static noise in the background and only focuses on the patterns of facial movements on each facial expression. We also discovered that the Differential Echo Profile is especially more effective when the user remounts the device because after the device is remounted, the position of the sensor shifts, which leads to visually different Echo Profiles. However, these differences are usually constant which can be mostly removed by calculating the Differential Echo Profile. Furthermore, we compared the performance of the system in a pilot study under three conditions: 1) original Echo Profile, 2) Differential Echo Profile, 3) original Echo Profile + Differential Echo Profile. The result also showed that Differential Echo Profile has the most robust performance among different scenarios. Thus, we decided to use Differential Echo Profile in EarIO.

The patterns in the Differential Echo Profiles that we observed convinced us that a machine learning algorithm may be able to estimate different facial movements based on Differential Echo Profiles.

4.3 Optimizing Wireless Data Transmission

Among common wireless solutions, WiFi provides highest bandwidth. However, its high power consumption makes it not practical for earables. In contrast, traditional Bluetooth solutions consume less power, but is still high for wearables. For optimized power performance, we chose BLE as our wireless data transmission solution, which is widely used in commodity wearables. However, the challenge of BLE solution is that the bandwidth is relatively limited. Therefore, we have to optimize EarIO data transmission process for BLE.

BLE 5.0 protocol supports 2M PHY, which in practice supports up to about 1.4 Mbps data throughput rate[11]. However, it is still not enough to transmit data for two 16-bit microphones sampling at 50 kHz. Moreover, in order to reduce power consumption as well as allowing sufficient margin over long-term operation, we prefer to limit the bit rate to below 1 Mbps. Therefore, we tried to reduce the depth of the samples hypothesizing that FMCW signals combined with long-sequence cross-correlation algorithm make the system insensitive to sample depth loss. To verify this, we collected 16-bit data using our wired data collection system. Then we downsampled the 16 bits raw signal to 2-12 bits and ran an end-to-end evaluation for each configuration on the facial expression tracking performance using the same metrics as used in our user study which will be described in Subsubsec. 4.4.3. The results in Fig. 5 demonstrate that the performance quickly increases from 2 to 4-5 bits and flattens at 6-7 bits sample depth. Based on these results, we chose 8 bits sample depth, which requires a bit rate of 800 kbps for the real-time data transmission at 50 kHz sampling rate. This transmission setting allows us to transmit data from two microphones using BLE 5.0 without compromising tracking performance.



(a) P6, cross-session

(b) P7, in-session

Fig. 5. Impact of Sample Depth on Tracking Performance. Experiments conducted on P6's cross-session data which is among the worst, and P7's in-session data which has the best performance among all participants. MAE drawn on the left axis, percentages on the right.

4.3.1 BLE Packet Loss Analysis. High throughput BLE data transmission often suffers from packet loss issues. Taking potential bandwidth fluctuations into consideration, we created a long buffer on the micro-controller for a smoother data transmission. We made the most of BLE 5.0's extended ATT MTU size and set each BLE packet at 244 bytes. In order to keep track of all lost packets, we used a 4-byte header in each BLE packet indicating its order. Combined with reduced bandwidth requirement, we were able to keep the packet loss rate at 0.035% (1621 out of 4658373) during all wireless data collection sessions. An analysis on all lost packets shows that 98.7% of data loss were caused by buffer running out, indicating occasional lack of sufficient bandwidth, while only 1.3% were lost during transmission (sent but never received).

[11]https://infocenter.nordicsemi.com/index.jsp?topic=%2Fsds_s140%2FSDS%2Fs1xx%2Fble_data_throughput%2Fble_data_throughput.html&cp=4_7_4_0_16

4.4 The Deep Learning Algorithm

4.4.1 Ground Truth Acquisition Method. To ensure ground truth quality, we adopted the TrueDepth camera on an iPhone with ARKit API as the ground truth acquisition system, which is similar to the prior work [4]. It uses 52 blendshapes to represent the shapes of the mouth, nose, eyes, eyebrows, cheeks, etc. Compared to 2D landmark-based ground truth of facial expressions, this ground truth acquisition method provides higher resolution and richness on representing facial movements, as it can model 3D movements. Furthermore, it allows us to easily render the facial expressions on Unity which helps us to visualize our results for quick analysis and comparison.

4.4.2 Deep Learning Model. Considering that the distance of our module to the skin is usually below 10 cm, also preserving sufficient margin, we first crop out the 100 pixels at the center of each channel (equivalent to about ± 19.75 cm from the module) and remove the rest. We then concatenate two channels vertically. After this step, each echo frame is reduced to a 1x200 vector. Please note that we use the Differential Echo Profile instead of original Echo Profile in training as discussed in the previous subsection. The temporal patterns of echo frames are highly informative to estimate facial movements. Therefore, instead of using the Echo Profile from one frame, we use the Differential Echo Profiles from one second (87 frames) to train the model. This results a feature vector with a dimension of 200×87, which is then sent to an end-to-end convolutional neural network (CNN) with ResNet-34 as the backbone and a fully-connected decoder to learn the facial expressions. The backbone extracts a feature vector with the dimension of 512. This new feature vector then goes through an average-pooling layer, a two-layer fully-connected regression network with dropout probability of 0.5, and output dimension of 52 to estimate the 52 blendshape parameters. To force the model to pay more attention to more active facial expressions (facial expression with more movements), we use mean squared error (MSE) loss so that the model can put heavier weights on the facial expression frames with larger movements (which usually leads to larger error). The mean absolute error (MAE) is used as the main evaluation metric. We use the Adam optimizer and set the learning rate at 0.01. The model is trained for 30 epochs.

4.4.3 Evaluation Metrics. As it was done in the past NeckFace [4] to efficiently evaluate the performance of our EarIO system, we adopt MAE as the main evaluation metric. As explained in Subsec. 4.4.1, we use 52 blendshapes from Apple ARkit API to represent any facial expression. As a result, we calculate the MAE between the 52 parameters predicted by our model and also generated by the TrueDepth camera of an iPhone 12. This is the overall MAE of all frames that we collect. From NeckFace [4] we know that while the MAE is below 40, the prediction is usually highly similar to the ground truth in visual. While NeckFace also reports the Active MAE (AMAE) to reflect the performance of active frames, we discovered that this approach did not work well in practice as it was difficult to settle on a proper activity threshold. Looking closer at our data we also discovered that the correspondence between a given value of MAE and the perceived visual difference of an expression was quite different when considering the lower and the higher parts of the face as shown Fig. 6.

Hence, we decided to separate the 52 blendshapes into two parts. For the upper face part we included all the 19 parameters related to eyes and brows, and the remaining 33 parameters for the lower face. We then calculated different MAEs for lower face parameters and upper face parameters, which are called Lower Face MAE (LMAE) and Upper Face MAE (UMAE). As we can see in the figure, when LMAE is lower than 40, the prediction of the lower face is highly similar to the ground truth visually. With respect to UMAE, this value is close to 60. Through the evaluation section we will use these values a reference point and we will report the Percentage of Frames with LMAE under 40 (PL40) and Percentage of Frames with UMAE under 60 (PU60) as well. We believe that this approach better reflects the actual performance of the system, and also allows for an easy comparison with previous work that also reported the MAE.

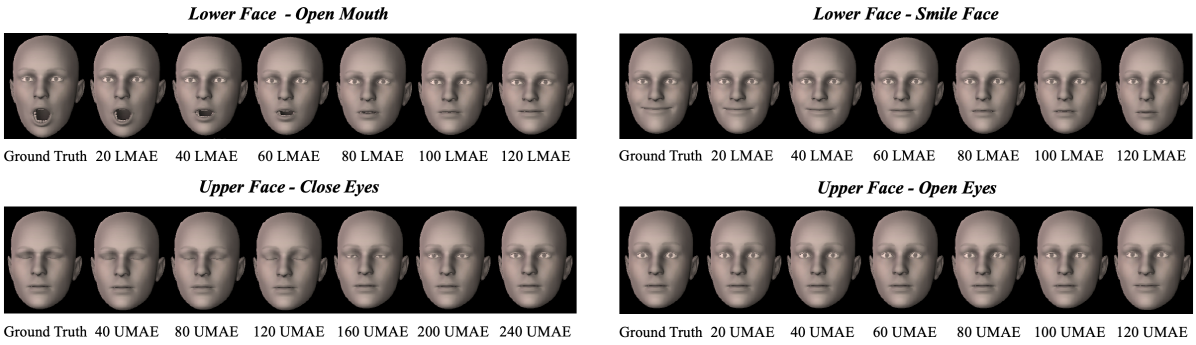


Fig. 6. Visualization Results with Different MAE for both Lower Face and Upper Face. The maximum MAE for closing eyes is much larger than other three expressions so we draw this figure using a different scale.

5 USER STUDY

In this section, we first present the goal and design of the user study. We then talk about the studies that we conducted in detail.

5.1 Study Objective

We conducted a user study in order to evaluate EarIO's effectiveness in continuously tracking full facial expressions. To validate the viability of our system we decided to first evaluate the system in a sitting configuration so that we could have a better control of the different parameters which might affect the performance of the system. We then extended the external validity of our finding by studying our system in a walking scenario to examine how EarIO would perform while the user was in motion.

5.2 Facial Expressions

We designed 9 types of different facial expressions to test EarIO, which we believe most movements of the facial muscles that people would do in their daily lives were covered when users were performing these facial expressions. The 9 different types of facial expressions and their corresponding Differential Echo Profiles are shown in Fig. 1. These expressions are designed to focus on different areas of the face, including lower (e.g., mouth, jaw) and upper (e.g., eyes, eyebrow) face. Note that these expressions are used as instructions to the users, but our goal is to evaluate how EarIO continuously tracks not only the expressions but also the transition between them. In particular, the fact that participants might not follow the provided instructions will have little impact on our evaluation as our system is not focusing on recognizing specific expressions.

5.3 Apparatus

5.3.1 Sitting Scenario. We used the wired prototype described in Subsubsec. 4.1.2 for the sitting scenario study to validate our EarIO system. Before the study, we 3D printed a case to hold the Teensy 4.1 and the power bank which powered the system. The participants needed to wear this case like a necklace during the study so that the whole system was completely wearable.

In order to record the ground truth of facial expressions, and play the instruction video to the participants as well, we used an iPhone 12 with the TrueDepth camera to realize these purposes. A stand was used to hold the iPhone 12 in front of participants. In the instruction video, we repeated the 9 facial expressions that we

introduced in Subsubsec. 5.2 multiple times, in a completely random order. The participants would just follow the instruction video and try to mimic whatever facial expressions they saw.

What is more, we were also using cameras to record videos of the participants from different angles during the study so that we could get a clear view of how the participants were wearing the device and performing different facial expressions, which helped us analyze the results of the study in a better and more efficient way.

5.3.2 Walking Scenario. For the walking scenario, participants would wear the EarIO device the same way as they did in the sitting scenario. The way how we held the iPhone 12 to record the ground truth and play the instruction video was different because the participants would need to wear a chest mount device to keep the smartphone in front of their face while walking. We did not have multiple cameras recording videos of the participants from different angles in the walking scenario because it would be hard to do this while participants were walking.

5.4 Study Procedures

5.4.1 Sitting Scenario. For the sitting scenario, the study was conducted in a small experiment room on campus. After the participants came in, and signed the consent form, we introduced the basic procedures of the study to them. Then participants were asked to sit on a chair with a stand holding the iPhone 12 on a table in front of them. Then the participants were instructed to wear the EarIO device as they usually did when wearing a pair of sports earphones. They also wore a case holding the Teensy 4.1 and the power bank, as a necklace.

If needed, we would adjust the angle of speakers and microphones to point them directly at the skin of the user in order to get stronger reflection of signals and clearer patterns. Because long hair may cover the speakers and microphones, having a severe impact on the experiment results, we asked the participants with long hair to use a hair tie to put their hair up during the study. Furthermore, because sometimes the TrueDepth camera is not capable of tracking the eye movement of users with eyeglasses, we also asked participants wearing glasses to remove them and all these participants reported that they were still able to see the instructions. However, our EarIO system itself does not require users to remove their glasses while in use. After finishing setting up the device, we started the study. Throughout the study, the participants followed the instruction video on the screen of the smartphone. The same smartphone was recording the ground truth of facial movements as well.

Before we started the official study, we had a short 1-minute practice section, which could help participants to get familiar with the study procedure and the facial expressions they were going to perform. Then the study started with the in-session section during which the device was not remounted at all, followed by the cross-session section during which the device was remounted between sessions.

For the in-session section, the participants followed the instruction video to perform multiple facial expressions without remounting the device. In-session section had 6 sessions of facial expressions in total, each one of which contained 6 repetitions of the 9 facial expressions that we designed in Subsec. 5.2. That means we had 324 facial expressions in total (6 sessions $\times$ 6 repetitions $\times$ 9 facial expressions). In our setting, each facial expression lasted 2 seconds in the instruction video, which means each session (54 facial expressions) lasted for 120.6 seconds including some grace time between two expressions. Therefore, the in-session section lasted for around 12 minutes (6 sessions $\times$ 120.6 seconds). Before and after the in-session section, we asked the participants to clap their hands in front of the smartphone which would help us synchronize our EarIO system and the ground truth acquisition system.

After the in-session section, the participants were instructed to take off the device and take a break. Following the break was the cross-session section, which included 20 sessions and each session also had 54 facial expressions, same as what has been described above. In this case, we had 1080 facial expressions (20 sessions $\times$ 6 repetitions $\times$ 9 facial expressions) and 40.2 minutes (20 sessions $\times$ 120.6 seconds) in total for the cross-session section. Before

each session, the participants were also asked to clap their hands for the synchronization purpose. After each session, the participants remounted the device and took a short break.

All these facial expressions were shown in a completely random order and were in a different order for each session of each participant.

5.4.2 Walking Scenario. This part was conducted in a large room next to the small experiment room, which provided a larger space for the participants to walk around while performing the facial expressions. In this part, the participants were given the same instruction and follow the same procedures to collect both in-session and cross-session data, as described above in Subsubsec. 5.4.1. The only difference was that the participants were required to walk around at their normal speed of their everyday lives in a larger experiment room while they were performing the facial expressions. In this part, they wore a chest mount which could hold the smartphone playing the instruction video and collecting the ground truth data in front of them while they were walking.

5.5 Participants

We recruited 12 participants (7 females and 5 males, ranging from 19 to 26 years old) for the sitting scenario. The study was conducted throughout different time of the day, including mornings, afternoons and evenings. We collected 12.43 hours of audio data, and corresponding ground truth files captured at a frame rate of 30 FPS.

Because of the duration of the study, the walking scenario was conducted at a different time from the sitting scenario. Because of the attrition to be expected in this setting, we were only able to capture the data for 7 participants (3 females and 4 males) who took part in both sitting and walking scenarios of our user study. This group is of particular value as it gives us the opportunity to compare performance of sitting and walking scenarios. We also recruited 3 new participants for a total of 10 participants (4 females and 6 males, ranging from 19 to 26 years old) for the walking scenario. 10.41 hours of audio data and corresponding ground truth files were collected from all 10 participants in this part of the study.

5.6 Follow-up Study with Wireless Prototype

We conducted a follow-up study with our wireless prototype to confirm that the wireless version of our EarIO system is capable of working comparably with the wired version with a much lower energy consumption.

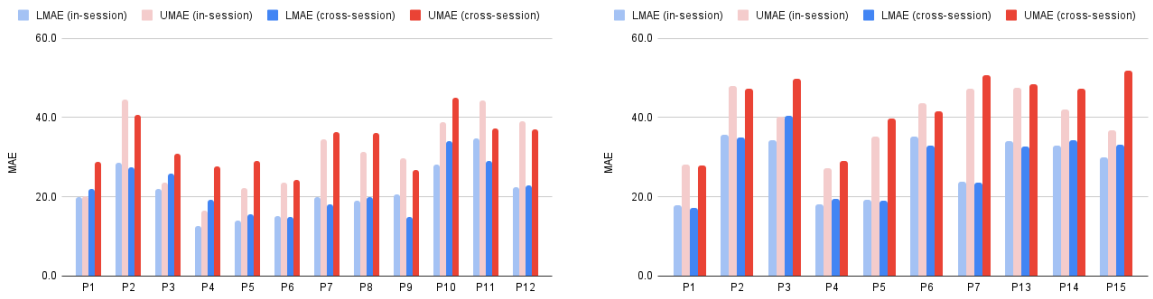
Since the wireless prototype is almost the same as the wired prototype except that the data collection and streaming methods are different, we only ran the study on a small scale of participants and only tested the system in the sitting scenario to confirm that the system was collecting and transmitting data reliably. We used 3 participants (2 participants from the previous study and 1 new participant, 2 females and 1 male). For this follow-up study, we collected 3.11 hours of audio data and corresponding ground truth files.

6 EVALUATION

6.1 Results

In Fig. 7, we plotted both LMAE and UMAE for all participants under all circumstances. As shown in the figure, the LMAE (blue bars) for all participants are always under 40 while the the UMAE (red bars) are always under 60, which means that the prediction of our EarIO system is visually highly similar to the ground truth for all participants according to Subsubsec. 4.4.3. For better reference and comparison, we also listed a summary of the aggregate results for in-session and cross-session performance under both sitting and walking scenarios of our user study in Tab. 2.

6.1.1 In-session Evaluation. We first evaluated our system's performance in in-session settings. We conducted a 6-fold cross-validation on all 6 sessions that we collected without remounting for each participant, and used the evaluation metrics described in Subsubsec. 4.4.3. This means that for each participant, we used about 18090



(a) Sitting scenario

(b) Walking scenario

Fig. 7. Lower-face MAE (LMAE) and Upper-face MAE (UMAE) for All Participants with In-session and Cross-session Setups

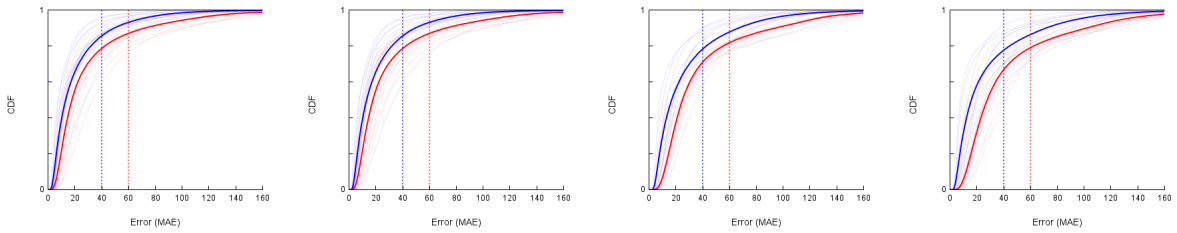
Table 2. Evaluation Results for In-session and Cross-session Performance while Sitting and Walking. Results are presented in the format of Mean | Standard Deviation.

Evaluation Metrics	In-Session		Cross-Session	
	Sitting	Walking	Sitting	Walking
MAE	24.6 | 7.5	32.1 | 7.3	25.9 | 6.0	33.9 | 7.8
LMAE	21.2 | 6.5	27.9 | 7.5	21.8 | 6.2	28.6 | 8.1
UMAE	30.5 | 9.6	39.3 | 7.6	33.1 | 6.5	43.1 | 8.7
PL40	85.8% | 7.9%	78.3% | 8.6%	84.8% | 7.2%	77.7% | 9.7%
PU60	87.0% | 6.3%	81.7% | 5.6%	85.5% | 4.3%	78.8% | 7.0%

samples (5 sessions × 120.6 seconds × 30 FPS), each of which has a dimension of 200 × 87 to train the model (see Subsubsec. 4.4.2 and Subsec. 5.4). In the sitting scenario, results show that the MAE ranges from 13.8 to 38.0 across different participants with an average of 24.6 ($std = 7.5$). LMAE is relatively lower at 21.2 while UMAE is 30.5. In our predictions, PL40 equals 85.8% while PU60 is 87.0% (Fig. 8 (a)). These results demonstrate the capability of EarIO in tracking full facial expressions.

Performance slightly drops but remain consistent in the walking scenario. The MAE ranges from 21.3 to 39.9 across all participants with an average of 32.1 ($std = 7.3$). Besides, LMAE and UMAE are 27.9 and 39.3 respectively. In our predictions, PL40 is 78.3% and PU60 is 81.7% (Fig. 8 (c)). These results demonstrate EarIO's ability to track facial expressions while users are in motion. While the user is walking, user's steps cause the sensor as well as the participant's head to slightly shake and displace. The objects in the background also changes as the user moves. These effects cause noises in the calculated Differential Echo Profiles, thus resulting slightly decreased performance. However, even with these drawbacks, EarIO still achieves consistent performance, demonstrating its adaptability to motion noises.

6.1.2 Cross-session Evaluation. We evaluated EarIO's cross-session performance to examine its robustness to remounting, which is very common in daily usage. Similar to the in-session evaluation, we conducted a 5-fold cross-validation on all 20 sessions that we collected with remounting for each participant. This means that for each participant, we used about 57888 samples (16 sessions × 120.6 seconds × 30 FPS) to train the model. In the sitting scenario, MAE averages 25.9 across 12 participants ($std = 6.0, min = 18.1, max = 37.8$). LMAE and



(a) Sitting (in-session) (b) Sitting (cross-session) (c) Walking (in-session) (d) Walking (cross-session)

Fig. 8. MAE CDF. Blue Lines: LMAE. Red Lines: UMAE. Solid Lines: Overall CDF. Pale Lines: CDF For Each Participant.

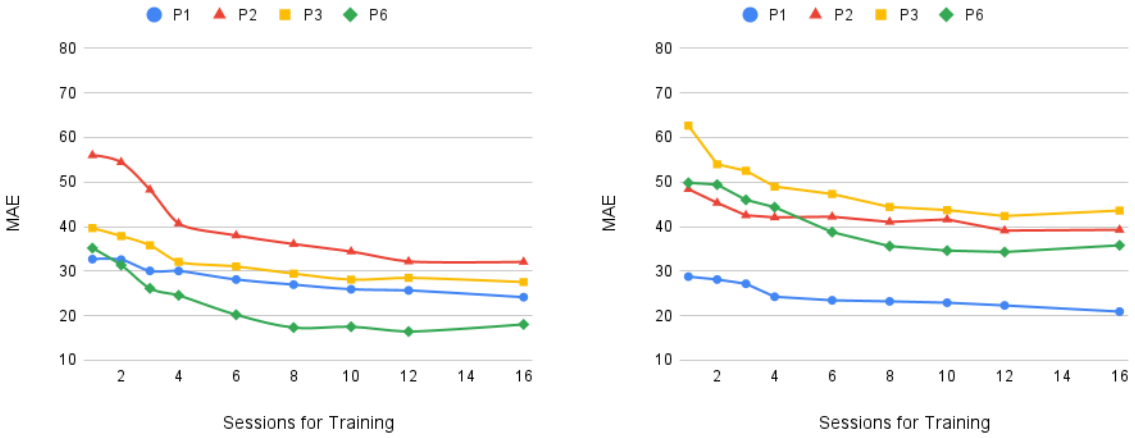
UMAE are 21.8 and 33.1, respectively. PL40 and PU60 are 84.8% and 85.5% respectively (Fig. 8 (b)). We ran a one-way repeated measures ANOVA test to compare the in-session and cross-session performance in the sitting setting and we did not find a statistically significant difference ($F(1, 22) = 1.04, p = 0.33 > 0.05$). Besides, we also calculated the effect size between these two groups. The Cohen's d reflects a very small effect size ($d = 0.196 < 0.2$) leading us to conclude that there is only a negligible performance difference between in-session and cross-session experiments in this setting. This suggests that our approach is resilient to device remounting in our sitting tests.

We found similar results for the cross-session performance while walking. As expected the performance of cross-session decreases compared to the in-session performance. Under this scenario, MAE averages 33.9 for cross-session ($std = 7.8, min = 20.9, max = 43.6$). The LMAE and UMAE are 28.6 and 43.1 respectively, with PL40 and PU60 as 77.7% and 78.8% (Fig. 8 (d)). We ran a one-way repeated measures ANOVA test to compare the in-session and cross-session performance in the walking setting and and did not find a statistically significant difference ($F(1, 18) = 2.93, p = 0.12 > 0.05$). Similarly, we calculated the effect size between these two groups and the Cohen's d ($d = 0.241 < 0.5$) reflects that there is only a small performance difference between in-session and cross-session experiments in this setting. This suggests that our approach is resilient to device remounting while walking, a frequent occurrence in practice.

6.1.3 Sitting versus Walking Comparison. As discussed before only seven participants participated in both the sitting and the walking part of our study. We ran one-way repeated measures ANOVA tests on the data captured for these participants and the results revealed a statistically significant difference between sitting and walking experiments in the in-session setting ($F(1, 12) = 15.95, p = 0.007 < 0.05$) and a statistically marginal difference in the cross-session setting ($F(1, 12) = 7.07, p = 0.04 < 0.05$). Furthermore, we calculated the effect sizes for these two settings. The results ($d = 1.233/1.097 > 0.8$) show that the practical significance of the finding that the tracking performance for sitting and walking scenarios are different is large in both in-session and cross-session settings.

6.2 The Impact of the Amount of Training Data

In Subsec. 6.1, we evaluated the user study results using 16 sessions for training and 4 sessions for testing for the cross-session setting because we would like to fully evaluate the performance of the system for each participant. However, collecting 16 sessions of data can be time consuming. Therefore, we conducted an additional experiment to see how the number of training data impacts the performance of the system. We picked 1 participant with one of the best performance in both siting and walking scenarios (P1), 1 participant with one of the worst performance in both scenarios (P2), and 2 other participants with moderate performance (P3 and P6). Then we used different numbers of sessions to train the model and tested on the same 4 sessions. The results are demonstrated in Fig. 9.



(a) Sitting scenario

(b) Walking scenario

Fig. 9. MAE for 4 Participants with Different Numbers of Training Sessions

As is shown in the figures, for participants with relatively good performance, e.g. P1, P2, and P3 for sitting and P1, P6 for walking, usually 8 sessions of training data are enough to predict their facial expressions accurately cross sessions. Since each session is around 2 minutes, the system needs the collection of 16-minute training data from them before usage. In contrast, for participants with worse performance, the system needs more sessions, e.g., 12 or 16 sessions to reach an acceptable result.

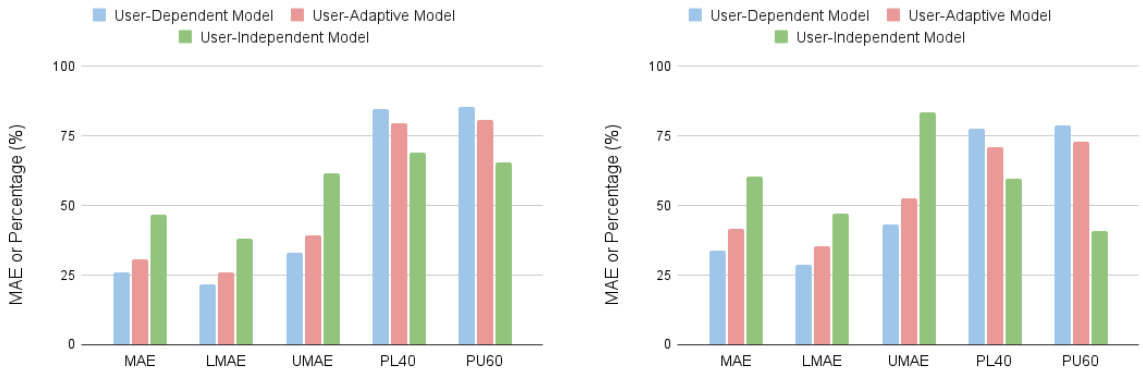
6.3 User Adaptive Model

In the previous subsection, we mentioned that our EarIO system would need 16 minutes or more of data for training before being used. This might be too long for some users in practice. As a result, we conducted a Leave-One-Participant-Out (LOPO) evaluation on all the participants for sitting and walking to see whether training data from other users can be used for new users, which can help decrease the time needed for data collection when new users come. The results are shown in Fig. 10. According to the results, We found that the model was very much user-dependent as the model trained on other people's data did not work very well if being directly used for new participants.

However, it is still possible to use this model to shorten the data collection process for new users if we fine-tune the user-independent model with just 2 sessions of data (4 minutes) collected from new users. We denote such model as the user-adaptive model. We will explain how we determine our user-adaptive model later in Subsec. 7.4. The results in Fig. 10 demonstrate that the user-adaptive model trained with 2 sessions can provide similar performance to the user-dependent model and works much better than the user-independent model. This means that we still need to collect data from new users but the amount of data needed will be much less. This will help make the device even more practical and easy to use with less amount of training data from the user.

6.4 Evaluation Using the Wireless Prototype

In the previous subsections, we validate the system using the wired version of the system. In order to further analyze our system in a more practical and low power consumption setting, we conducted a follow-up study



(a) Sitting scenario

(b) Walking scenario

Fig. 10. Comparison between User-Dependent, User-Adaptive, and User-Independent Models across 5 Evaluation Metrics

with 3 participants (P4, P5 and one new participant P16, 2 females and 1 male) to test our wireless version of the system. We only tested the sitting scenario because we only needed to verify that the data collection and streaming of the wireless system is stable enough to support the facial tracking system. Overall, based on a 6-fold cross-validation on 6 sessions collected for each participant, the MAE, LMAE, and UMAE for in-session experiments are 23.4 ($std = 10.9, min = 14.8, max = 35.6$), 19.4, and 30.2 respectively while PL40 and PU60 are 88.6% and 87.2%. For the cross-session experiments, using a 5-fold cross-validation on 20 sessions collected for each participant, we have 25.6 ($std = 7.9, min = 19.6, max = 34.5$), 21.4, and 32.9 for MAE, LMAE and UMAE respectively while PL40 and PU60 are 86.1% and 85.7%. These results are very similar to the performance of the wired prototype shown in Tab. 2. Hence, by verifying the performance of the wireless system, it is safe to say that our EarIO is capable of tracking facial expressions continuously with significantly low energy consumption. In the next subsection, we would evaluate the specific power consumption of our wireless system.

6.4.1 The Impact of External Noise. Because our system relies on acoustic signals to reconstruct facial expressions, there is a possibility that external noise in users' daily lives could have influence on the system's performance. As a result, we tested the performance of our system with the existence of external noise. Since the effect of the noise should be additive in a linear system, we decided to record the noise with our wireless prototype separately and add the noise into the data we have already collected before in the wireless study. We believe this is a good approximation of the system performance in real life. We tested the system under two different kinds of external noise. Firstly, we wore the wireless system and sat by the roadside. We kept the system on to record the noise of vehicles passing by and the noise of the wind. Then we mixed the noise into the data collected from three participants in our previous wireless study. Because the noise was collected using the same system as the one used in the wireless study, the mixing of noise and collected data has the same effect as testing the system on the street. Using the same method, we used the wireless device to record the noise of users talking on the phone and mixed it into the collected data separately.

Then we adopted 16 sessions data without noise from 20 sessions cross-session data collected from each participant to train a model and tested the system on the remaining 4 sessions under three different settings, without noise, mixed with street noise and mixed with talking noise. The tracking performance of EarIO in these three scenarios is shown in Tab. 3. As is shown in the table, both noises have little impact on the performance

of the system for all three participants. We believe that the noise in our daily lives like those caused by people talking or vehicles passing by usually has low energy in the frequency range we used so it should not cause severe performance drop in our system. Therefore, this experiment proves the robustness of our system in noisy environments.

Table 3. Performance of EarIO with Two Different Kinds of External Noise

External Noise	P4			P5			P16		
	MAE	LMAE	UMAE	MAE	LMAE	UMAE	MAE	LMAE	UMAE
No Noise	20.2	16.7	26.1	20.1	14.0	30.8	34.0	31.0	39.1
Street Noise	19.2	15.8	25.0	20.5	14.0	31.8	34.3	31.6	38.9
Talking Noise	19.7	15.8	26.6	20.4	14.4	30.9	34.7	32.0	39.3

6.5 Power Consumption

We examined EarIO's power signature to evaluate the practicability in deploying it on commodity devices. We used a Current Ranger[12] to measure the current drawn from the battery. Overall, EarIO consumes 153.7 mW (41.5 mA@ 3.70 V) with 2 speakers and 2 microphones on, and data transmission at 800 kbps. The 110 mAh battery lasted 3 hours during our test run. Fig. 11 illustrates the current draw from the battery over time and we can clearly identify the draw from the BLE subsystem. The two speakers consume the most power as the system consumes only 12 mA on average without the speakers connected.

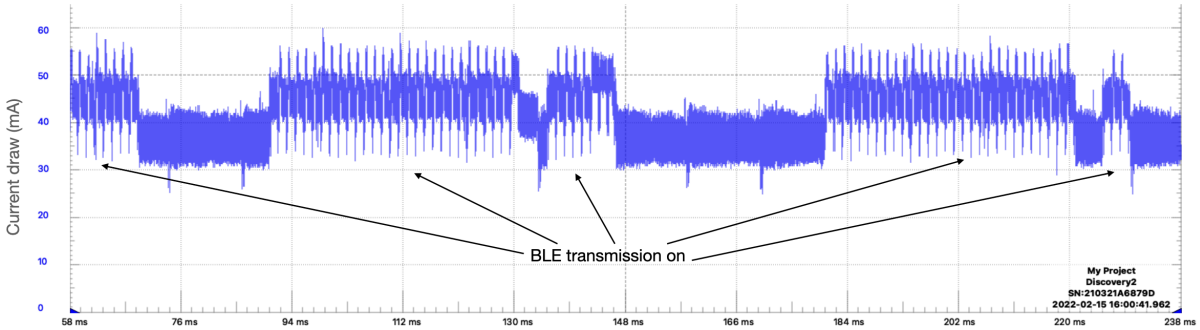


Fig. 11. Power Signature of EarIO Measured by Current Ranger

6.6 User Perception and Comfort

At the end of the study, each participant was instructed to finish a questionnaire to document their demographic information and collect their thoughts and comments on the device. In this questionnaire, they were first asked a question "How comfortable is this wearable device to wear around the face (0 most uncomfortable, 5 most comfortable)?" Among all the 16 participants who took part in at least one of our study, an average score of 3.2 is given for the comfort of the device. In terms of the question "Were there any periods of discomfort?", four participants mentioned that the data collection process was too long while only one participant mentioned that

[12]https://lowpowerlab.com/guide/currentranger/

he noticed the sound emitted from the device and felt not quite comfortable with it. As for how to reduce the amount of time needed for data collection, we have discussed it in Subsec. 6.2 and 6.3. Besides, the operating frequency range of the system can be shifted even higher to make sure that it is completely inaudible to every user, which will be discussed in Subsec. 7.2. Overall, most participants do not regard EarIO as a device which is uncomfortable to wear and do not consider the emitted acoustic signals as a discomfort during the study.

6.7 Summary

In this section, we first validated the performance of our EarIO system based on the results of the user study. Results show that the system can work satisfactorily both with and without remounting while users are sitting and walking. We also evaluate the amount of training data needed for new users and the possibility of adopting the user-adaptive model to significantly reduce the time of data collection. Following that, we conducted a follow-up study with our wireless system and measured its power consumption to verify that our system can work reliably using the wireless system with a remarkably low power consumption compared with previous work. Furthermore, we verified the robustness of our system with the existence of external noise, as well as the user perception and comfort of the system. All the results above demonstrate that our EarIO system is capable of providing a low-power and practical method to track full facial expressions continuously.

7 DISCUSSION

7.1 Comparison with Previous Facial Expression Tracking Technologies

As discussed in Subsec. 2.2, most of the previous wearable facial expression tracking technologies are only capable of distinguishing several discrete facial gestures, including Interferi [13] which also uses acoustic sensing. Among the three recent work which can track facial movements continuously on wearables [4, 5, 38], we picked NeckFace [4] to compare the tracking performance of our EarIO system with because it adopts the same ground truth acquisition method and similar evaluation metrics. The results are shown in Tab. 4. As is shown in the table, EarIO is better than NeckFace in sitting scenarios while worse than it in walking scenarios. Overall, these two technologies are basically comparable to each other in tracking performance. However, compared with NeckFace which is camera-based, EarIO can achieve a much lower power consumption (25 times lower than NeckFace as discussed in Subsec. 2.2). As a result, we believe EarIO is a more practical wearable device with lower power consumption while maintaining comparable tracking performance compared with previous camera-based technologies.

Table 4. Comparison between EarIO and NeckFace [4] under Different Scenarios. Evaluation Metric: MAE. Results are presented in the format of Mean | Standard Deviation.

Projects	In-Session		Cross-Session	
	Sitting	Walking	Sitting	Walking
EarIO	24.6 | 7.5	32.1 | 7.3	25.9 | 6.0	33.9 | 7.8
NeckFace (Necklace)	30.3 | 6.3	25.4 | 6.4	34.1 | 9.5	/
NeckFace (Neckband)	25.6 | 5.1	22.6 | 5.2	28.4 | 7.9	/

7.2 Frequency Range

In the system implementation and user study, we used the frequency range of 16-20 kHz. This range was empirically determined since it is not audible to most people, meanwhile it falls well within the limits with decent margin of the 44.1 kHz sampling rate and the operating limits of the speaker and microphone. This range might

be audible to some people. However, this frequency range can be easily modified based on the characteristics of the speakers and microphones and sampling rate. We do no anticipate impact on performance by shifting the frequency range higher as long as it is supported.

7.3 Noise Exposure Measurement

Because our system is continuously emitting sounds during the tracking process, there might be a concern that this device can do harm to the hearing of users. As a result, we performed a noise exposure measurement of our wireless system to make sure that our device is safe to use. According to the noise exposure limits recommended by Centers for Disease Control and Prevention (CDC), a person can continuously be exposed to 85 $dB(A)$ over 8 hours in the work space before reaching the maximum allowable daily dose [24]. Besides, for general environmental noise outside the work space, a report from the U.S. Environmental Protection Agency (EPA) in 1974 recommended a 70 $dB(A)$ over 24-hour and 75 $dB(A)$ over 8-hour for average exposure limit [25].

Because dB(A) is measured while taking various sensitivities of human ears to different frequency ranges into account, we believe this scale is very suitable for the noise exposure measurement of our EarIO system. Hence, we used a smartphone app provided by CDC[13] to measure the noise level of our wireless system. We did two testings. First, we turned on the speaker and attached the smartphone directly onto the speaker. The app gave us a measurement of 67.9 $dB(A)$. Then we measured the distance from the speaker to our ears while wearing the device, which was around 6 cm. We put the device on table and place the smartphone around 6 cm away from the speaker. The measure was around 63.5 $dB(A)$. Because when users wear the device, there is some occlusions between the device and their ear canal, the 63.5 $dB(A)$ is actually the upper bound of the noise level that users could hear. Comparing this result with recommendation from CDC and EPA, we are confident that our device is safe to wear and can work all day with little concern of damaging the hearing of users.

7.4 Analysis of User-Adaptive Model

As mentioned in Subsec. 6.3, we applied a user-adaptive model to reduce the amount of time needed for data collection process while new users come. In this subsection, we want to justify our choice of parameters for the user-adaptive model. To determine how many sessions are enough to fine-tune the model, we experimented using 1-4 sessions and the results show that all evaluation metrics converge at 2 sessions of data and using more sessions will not increase the performance significantly.

Besides, we would like to make sure that the user-adaptive model can make a difference compared with the performance not using a pre-trained model from other users. Hence, we selected the same 4 participants as Subsec. 6.2 did and used 2 sessions from them respectively to fine-tune the user-adaptive model and also to directly train a model for prediction. The 5 metrics MAE, LMAE, UMAE, PL40, and PU60 of the prediction outputted by the user-adaptive model for sitting are 31.8 ($std = 6.1, min = 26.8, max = 40.5$), 28.4, 37.8, 77.6%, and 81.8% while those of the prediction outputted by the directly trained model are 39.1 ($std = 10.6, min = 31.4, max = 54.4$), 36.1, 44.2, 68.9%, and 77.4%. For walking, the 5 metrics are 42.2 ($std = 10.1, min = 28.0, max = 50.2$), 38.0, 49.5, 66.9%, and 76.5% of the prediction outputted by the user-adaptive model while those of the prediction outputted by the directly trained model are 44.2 ($std = 13.2, min = 28.1, max = 54.0$), 41.0, 49.8, 63.1%, and 75.3%. By comparison, it is safe to say that the user-adaptive model can significantly improve the performance for new users with just a small amount of training data collected, especially under the sitting scenario.

7.5 Privacy

Although we have made our efforts to improve the practicability of our system to a large extent, there are still some challenges remaining before directly putting this device into immediate deployment. For instance, the

[13]https://blogs.cdc.gov/niosh-science-blog/2014/04/09/sound-apps/

recorded audio may still incur some privacy concerns from users. One direction to address this issue is to perform on device filtering before transmitting the data through BLE. This can play an important role in alleviating the privacy concern of EarIO.

7.6 Deployment of EarIO Algorithms on Smartphone

In order to enable our EarIO system to run in real time, we plan to deploy the data processing and machine learning pipeline of EarIO on the smartphone. This is possible with the support of PyTorch Mobile[14]. With this implemented, the collected data can be transmitted to the user's personal phone and processed on it as well. Only the estimated blendshapes will be forwarded to other devices for potential interaction with other users so that the privacy of the user can be better protected. The estimated facial movements can be used in many applications like rendering avatars for users in video chats in real time. Further experimentation will be needed to evaluate this approach in practice.

7.7 Limitation

Similar to other wearable devices, our EarIO system also has some limitation which we should put efforts in to resolve in future work. In this subsection, we list five major limitation of the system from our perspective.

7.7.1 Limited Choices of the Size of Earbuds. In Subsec. 6.1, we demonstrated that the system worked with an acceptable performance for all participants. However, there are some participants who have relatively worse prediction results than others, such as P10 in the sitting scenario. Based on the videos we recorded during the experiment, we find that the earbuds we were using were too big to fit into the ear canal of P10 so the device was actually very unstable on P10's ears. Although our system was built upon commercial earphones and we did have three sizes of the earbuds to change for different participants, even the smallest earbuds were too big for P10. Hence, this is actually a drawback of the commercial earphones. However, in the future, we planned to design our own earphone so that the size of the earbuds has more flexibility to switch between users.

7.7.2 Hair Blocking. As we mentioned in the study procedure part in Subsubsec. 5.4.1, we asked all participants with long hair to put their hair up using a hair tie because we believe long hair has a possibility of blocking the device and preventing it from tracking facial expressions accurately. This can limit the usage of the device in real life because users do not always bring a hair tie with them and it will be inconvenient for them to put their hair up every time they use this device. One way to solve this issue is to improve the form factor of our hardware design, enabling more degrees of freedom to adjust the angles and lengths of different parts of the form factor. In this way, users have the options to adjust the position of the device if they find their hair is blocking the device. Besides, deploying the system using in-ear speakers and microphones can also help us avoid this blocking issue.

7.7.3 Running. In Sec. 6, we showed that our EarIO system could achieve a satisfactory performance in both sitting and walking scenarios. Apart from these results, we also tested the system in a situation where the user is moving faster. Using a setting similar to the cross-session section in the study described in Subsec. 5.4, one researcher collected 16 sessions of cross-session data while keeping still for training. Then another 4 sessions of data with the researcher keeping still and 4 sessions of data with the research running were collected for testing. While the researcher was collecting the running data, he used a setup similar to the procedures of the walking scenario introduced in Subsubsec. 5.4.2. He adopted a chest mount to hold the iPhone 12 in front of him to collect ground truth and followed the instruction video to perform facial expressions while running in the large experiment room. The testing result for 4 still sessions is 27.2 for MAE while that for 4 running sessions is 62.6 for MAE. Even though using one running session to fine-tune the model can improve the MAE for running

[14]https://pytorch.org/mobile/home/

sessions to 48.9, the performance is still not good compared with sitting and walking scenarios. We believe this decrease of performance is mainly caused by the rapid movement of the device and the user's body. While sitting and walking are the most common circumstances when users will use this device, it is possible that users may also want to have their facial expressions tracked while running so it is crucial to perform future study, such as improving the stability of the device, collecting a larger amount of data in this scenario, and applying more complex noise-removal algorithms to better understand how to improve the performance of the system in more dynamic environments.

7.7.4 User Dependency. Currently, our EarIO model is still user-dependent, which means new users still need to collect training data before using this system. Even though we designed a user-adaptive model to shorten this data collection process, the result was not completely the same as the user-dependent model. Especially for some participants with worse performance, we may need to collect more training data from them. Considering that different people have different shapes of faces, it might be helpful if we can take this into account for prediction when we try to develop a completely user-independent model in the future. If we can make an effort to make the system user-independent, then the EarIO system would be as convenient as the frontal camera-based methods, without requiring the data collection process before usage for any user.

7.7.5 Requirements on Changing the Hardware Setting on Commodity Earphones. EarIO only needs one pair of microphone and speaker on each side, which most commodity earphones already have. However, it is true that our setup (e.g., position, orientation) of the microphone and speaker may need to be adjusted in order to deploy EarIO. Exploring the feasibility of deploying EarIO on commodity earables is what we plan to explore in the next step.

8 CONCLUSION

This paper presents EarIO, a low-power, minimally-obtrusive and practical acoustic sensing method to track full facial expressions continuously. By using acoustic signals and the customized PCBs to transmit data via BLE, the system operates at a very low power consumption at 154 mW while maintaining a comparable tracking performance compared with previous work. A user study with 16 participants in total under three different scenarios validated our system while participants are sitting, walking and after remounting the device. A follow-up study verified the stability and practicability of our wireless system to work at a low energy consumption state. There still remains several challenges before we can put the device into immediate deployment but we also discussed several directions to go for future work to address these issues.

ACKNOWLEDGMENTS

This work is supported by the Ann S. Bowers College of Computing and Information Science at Cornell University. We would like to thank all the lab members who have provided meaningful feedback on the project and the manuscript at Cornell SciFi Lab. Furthermore, we also appreciate all the participants who took part in the study and offered valuable comments on the project after the study during the ongoing global pandemic.

REFERENCES

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

[2] Md Tanvir Islam Aumi, Sidhant Gupta, Mayank Goel, Eric Larson, and Shwetak Patel. 2013. DopLink: using the doppler effect for multi-device interaction. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 583–586.

[3] Jaekwang Cha, Jinhyuk Kim, and Shiho Kim. 2016. An IR-based facial expression tracking sensor for head-mounted displays. In *IEEE SENSORS*. IEEE, 1–3.

[4] Tuochao Chen, Yaxuan Li, Songyun Tao, Hyunchul Lim, Mose Sakashita, Ruidong Zhang, François Guimbretière, and Cheng Zhang. 2021. NeckFace: Continuously Tracking Full Facial Expressions on Neck-mounted Wearables. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, Vol. 5. 1–31.

[5] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. 2020. C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-mounted Miniature Cameras. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology (UIST)*. 112–125.

[6] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18, 1 (2001), 32–80.

[7] Lloyd E Emokpae, Stephen DiBenedetto, Brad Potteiger, and Mohamed Younis. 2014. UREAL: Underwater reflection-enabled acoustic-based localization. *IEEE Sensors Journal* 14, 11 (2014), 3915–3925.

[8] Anna Gruebler and Kenji Suzuki. 2010. Measurement of distal EMG signals using a wearable device for reading facial expressions. In *Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 4594–4597.

[9] Shan He, Shangfei Wang, Wuwei Lan, Huan Fu, and Qiang Ji. 2013. Facial expression recognition using deep Boltzmann machine from thermal infrared images. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 239–244.

[10] Steven Hickson, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan Essa. 2019. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1626–1635.

[11] Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. 2015. Unconstrained realtime facial performance capture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1675–1683.

[12] Earnest Paul Ijjina and C Krishna Mohan. 2014. Facial expression recognition using kinect depth sensor and convolutional neural networks. In *International Conference on Machine Learning and Applications*. IEEE, 392–396.

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

[14] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al. 2013. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the ACM on International Conference on Multimodal Interaction*. 543–550.

[15] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10 (2009), 1755–1758.

[16] Ying-Hsiu Lai and Shang-Hong Lai. 2018. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 263–270.

[17] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–9.

[18] Jie Lian, Jiadong Lou, Li Chen, and Xu Yuan. 2021. EchoSpot: Spotting Your Locations via Acoustic Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–21.

[19] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. 2014. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1749–1756.

[20] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. 2014. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1805–1812.

[21] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. Lip reading-based user authentication through acoustic sensing on smartphones. *IEEE/ACM Transactions on Networking (TON)* 27, 1 (2019), 447–460.

[22] 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.

[23] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1515–1525.

[24] 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/

[25] U.S. Environment Protection Agency Office of Noise Abatement and Control. 1974. Information on levels of environmental noise requisite to protect public health and welfare with adequate margin of safety. *EPA/ONAC 550/9-74-004* (1974). http://nepis.epa.gov/Exe/ZyPDF.cgi/2000L3LN.PDF?Dockey=2000L3LN.PDF

[26] Ville Rantanen, Pekka-Henrik Niemenlehto, Jarmo Verho, and Jukka Lekkala. 2010. Capacitive facial movement detection for human–computer interaction to click by frowning and lifting eyebrows. *Medical & biological engineering & computing* 48, 1 (2010), 39–47.

[27] Marc'Aurelio Ranzato, Joshua Susskind, Volodymyr Mnih, and Geoffrey Hinton. 2011. On deep generative models with applications to recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2857–2864.

[28] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. 2012. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*. Springer, 808–822.

[29] James A Russell. 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological bulletin* 115, 1 (1994), 102.

[30] Nicu Sebe, Michael S Lew, Yafei Sun, Ira Cohen, Theo Gevers, and Thomas S Huang. 2007. Authentic facial expression analysis. *Image and Vision Computing* 25, 12 (2007), 1856–1863.
[31] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom)*. 591–605.
[32] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.* 34, 6 (2015), 183–1.
[33] Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. 2021. ExpressEar: Sensing Fine-Grained Facial Expressions with Earables. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, Vol. 5. 1–28.
[34] 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.
[35] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom)*. 82–94.
[36] Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. An anatomically-constrained local deformation model for monocular face capture. *ACM transactions on graphics (TOG)* 35, 4 (2016), 1–12.
[37] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. 2018. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2129–2138.
[38] Yi Wu, Vimal Kakaraparthi, Zhuohang Li, Tien Pham, Jian Liu, and Phuc Nguyen. 2021. BioFace-3D: Continuous 3d Facial Reconstruction through Lightweight Single-Ear Biosensors. In *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom)*. 350–363.
[39] Wentao Xie, Qian Zhang, and Jin Zhang. 2021. Acoustic-Based Upper Facial Action Recognition for Smart Eyewear. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, Vol. 5. 1–28.
[40] Xuhai Xu, Haitian Shi, Xin Yi, Wenjia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K Dey. 2020. EarBuddy: Enabling On-Face Interaction via Wireless Earbuds. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.
[41] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the Annual International Conference on Mobile Systems, Applications, and Services*. 15–28.
[42] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Ruichen Meng, Sumeet Jain, Yizeng Han, Xinyu Li, Kenneth Cunefare, Thomas Ploetz, Thad Starner, et al. 2018. FingerPing: Recognizing fine-grained hand poses using active acoustic on-body sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–10.
[43] Ruidong Zhang, Mingyang Chen, Benjamin Steeper, Yaxuan Li, Zihan Yan, Yizhuo Chen, Songyun Tao, Tuochao Chen, Hyunchul Lim, and Cheng Zhang. 2021. SpeeChin: A Smart Necklace for Silent Speech Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 5, 4 (2021), 1–23.
[44] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing Acoustic-Based Imaging for Issuing Contact-Free Silent Speech Commands. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 4, 1 (2020), 1–26.
[45] Yunting Zhang, Jiliang Wang, Weiyi Wang, Zhao Wang, and Yunhao Liu. 2018. Vernier: Accurate and fast acoustic motion tracking using mobile devices. In *IEEE International Conference on Computer Communications (INFOCOM)*. IEEE, 1709–1717.