I am an Earphone and I can Hear my Users Face: Facial Landmark Tracking using Smart Earphones

SHIJIA ZHANG*, Penn State University, USA
TAITING LU*, Penn State University, USA
HAO ZHOU, Penn State University, USA
YILIN LIU, Penn State University, USA
RUNZE LIU, Penn State University, USA
MAHANTH GOWDA, Penn State University, USA

This paper presents EARFace, a system that shows the feasibility of tracking facial landmarks for 3D facial reconstruction using in-ear acoustic sensors embedded within smart earphones. This enables a number of applications in the areas of facial expression tracking, user-interfaces, AR/VR applications, affective computing, accessibility, etc. While conventional vision-based solutions break down under poor lighting, occlusions, and also suffer from privacy concerns, earphone platforms are robust to ambient conditions, while being privacy-preserving. In contrast to prior work on earable platforms that perform outer-ear sensing for facial motion tracking, EARFace shows the feasibility of completely in-ear sensing with a natural earphone form-factor, thus enhancing the comfort levels of wearing. The core intuition exploited by EARFace is that the shape of the ear canal changes due to the movement of facial muscles during facial motion. EARFace tracks the changes in shape of the ear canal by measuring ultrasonic channel frequency response (CFR) of the inner ear, ultimately resulting in tracking of the facial motion. A transformer based machine learning (ML) model is designed to exploit spectral and temporal relationships in the ultrasonic CFR data to predict the facial landmarks of the user with an accuracy of 1.83 mm. Using these predicted landmarks, a 3D graphical model of the face that replicates the precise facial motion of the user is then reconstructed. Domain adaptation is further performed by adapting the weights of layers using a group-wise and differential learning rate. This decreases the training overhead in EARFace. The transformer based ML model runs on smartphone devices with a processing latency of 13 ms and an overall low power consumption profile. Finally, usability studies indicate higher levels of comfort of wearing EARFace’s earphone platform in comparison with alternative form-factors.

CCS Concepts: · Human-centered computing → Ubiquitous and mobile computing; · Hardware → PCB design and layout.

Additional Key Words and Phrases: Wearable Sensing, Mobile Computing, Earable Computing, Facial Reconstruction, IoT

1 INTRODUCTION

This paper presents a system called EARFace (Earphone Acoustics based Reconstruction of 3D Face), that shows the feasibility of reconstructing 3D facial motion using sensor embedded smart earphones, which are gaining in...
popularity. This enables innumerable applications in the areas of facial expression recognition, emotional well being monitoring, affective computing, animation rendering, augmented and virtual reality, user interfaces, etc. For instance, the mental state of people with autism and depression can be constantly assessed, which can provide critical feedback to healthcare providers [57, 111]. Accessibility applications such as controlling wheel-chair equipment can also leverage facial motion based user-interfaces [95, 105]. In the context of augmented and mixed reality applications, a more immersive experience can be created by gauging the interest and attention levels of each user based on their emotions [32, 33].

Recent works in computer vision [90, 97] show the ability of tracking fine grained 3D facial landmarks with high accuracy. However, camera based solutions are susceptible to ambient lighting, resolution, occlusions, and raise privacy concerns. Moreover, the camera needs to be present in front of the user with a clear focus on the face at all times which might be impractical for applications like wheel-chair equipment control, constant monitoring of emotional states of people with autism where the user can move freely. In contrast, EARFace tracks facial landmarks using wearable smart-earphones that are robust to ambient lighting, occlusions, and offers ubiquitous and portable tracking, anytime and anywhere without the dependency on external infrastructure.

Prior works in the area of wearable based facial tracking includes works like EarFs [82] that inserts electrodes in the ear. In response to facial expressions, the electrodes can sense changes in electrical fields with electromyography (EMG), electrooculography (EOG), and capacitive sensing. EarFs classifies five facial expressions. CanalSense [22] measures air pressure changes in response to facial expressions for classifying eleven expressions. Similarly, ECTF [20] uses acoustic sensors in the earphone for classifying facial expressions. In contrast to such works that classify predefined discrete facial expressions, EARFace performs continuous tracking of facial landmarks for 3D facial reconstruction, which can be used by any generic application including facial expression classification. Our work is inspired by recent works such as BioFace3D [112] and EarIO [69] that track facial landmarks. However, based on the usability study in Section 6, their form-factor can cause discomfort because the sensing happens outside the ears with the sensors protruding onto the back of the head and slightly on the face. In contrast to outside-ear sensing in these works, EARFace performs in-ear sensing to sense the shape of ear canal in response to facial motion. The acoustic sensors are embedded in a natural and small earphone form-factor, thus securing higher usability ratings. Such a form-factor is gaining in popularity with a number of applications in healthcare, emotion and human activity recognition, AR/VR, etc [36, 51, 102].

Fig. 1 illustrates the high-level overview of EARFace. The core idea is simple. The shape of the ear canal changes during motion of facial muscles involved during facial motion of lips, eyes, nose, etc (details in Secion 3). EARFace senses the changes in ear canal shape using ultrasonic reflections of the ear canal as captured by the acoustic sensors embedded in the earphones. Towards measuring changes in the shape of the ear canal, the channel frequency response (CFR) of the ear canal is computed and LFCC [122] features are extracted. ML models based on transformers are then designed to capture rich spectral and temporal relationships in the ear-canal CFR data. The models are trained to predict facial landmarks which are later converted into a 3D facial reconstruction by fitting the landmarks into the parameters of FLAME [70], a popular model for facial graphics and animation. Fig. 2 shows examples of 3D facial reconstruction in EARFace which shows the ability to monitor various parts of the face across a universal group of facial motions. A sample demo is included in the link [37] where it can be seen that even eye blinks can be captured. We believe these results are promising.

The problem of facial landmark tracking is challenging for a number of reasons: (i) The changes in ear canal shape due to facial activity is very subtle. (ii) There is no well formed equation or a straightforward relationship between facial expressions and ear canal shape. (iii) The overhead of training data must be minimal and yet the
Fig. 1. Overview of EARFace: The ultrasonic reflections from ear canals is used for extracting the CFR. Features are extracted from the CFR and processed by a 2D transformer based ML model that captures spectral and temporal dependencies in the ear canal CFR to predict facial landmarks, which are ultimately used for reconstructing the 3D face model of the user.

ML models designed must be generalizable to a diverse pool of users and facial motion. (iv) The form factor of the sensing device must be small enough to ensure comfortable level of wearing.
**EARFace** exploits a number of opportunities in hardware and machine learning to solve these challenges: (i) **EARFace** captures CFR within the ear canal using ultrasonic frequencies with a large range up to 40 kHz. The high-frequency CFR provides sufficient resolution to capture subtle variations in-ear canal shape. (ii) **EARFace** designs ML models based on transformers to efficiently learn both spectral and temporal relationships that map the high-resolution acoustic CFR data into facial landmarks. Furthermore, **EARFace** incorporates FLAME model parameters which enable efficient reconstruction of the 3D face using the estimated facial landmarks. (iii) **EARFace** performs domain adaptation to achieve a sweet spot between training overhead and model customizability to an individual user. A group-wise learning rate is designed for domain adaptation where different layers are updated with different learning rates for efficient adaptation. (iv) **EARFace** integrates ultrasonic speakers and microphones into a natural earphone with a small form-factor for capturing the ear-canal shape using ultrasonic CFR data.

Because of requirement of long range of sound frequencies until 40 kHz, **EARFace** develops its own earphone platform. We embed a Sonion EST65DB01 speaker and Sonion P11AC03 microphone operating in ultrasonic frequency ranges up to 40 kHz for capturing acoustic CFR using which the ear canal shape and hence the facial landmarks can be estimated. The ML models in **EARFace** are implemented on smartphones using TensorFlowLite. Evaluated over several experiments across diverse users, **EARFace** achieves an accuracy of 1.83 mm in facial landmark tracking. Furthermore, our experiments validate robustness to natural variation in earphone wearing positions, body and head motion, and the ability to track different parts of the face (eyes, mouth, nose). Therefore, we believe **EARFace** offers a practical solution.

Summarizing the above possibilities, we enumerate our contributions below: (i) Feasibility of tracking facial landmarks for 3D facial reconstruction using in-ear sensing with a earphone form-factor— a first such attempt to the best of our knowledge. (ii) Design of transformer based ML model to capture the rich spectral and temporal relationships across acoustic data to track facial landmarks. (iii) Domain adaptation where various layers are adapted with different learning rates depending on their importance levels. This decreases the training overhead in **EARFace**. (iv) Design of an earphone embedded with acoustic sensors with a form-factor that is comfortable for wearing. (v) Implementation and experimentation across diverse users to demonstrate the feasibility and robustness.

2 RELATED WORK

Table 1 contrasts **EARFace** in the context of key related works. This section elaborates on the details.

**Vision:** OpenPose [30] is a widely popular system that can capture 2D facial landmarks corresponding to eyes, mouth using monocular camera images. More recent works [90, 97] have shown the feasibility of tracking 3D facial landmarks using camera images. The facial features extracted from cameras are fit into the FLAME model [70] parameters that can precisely represent the 3D facial shape, pose, and expression. However, camera based solutions are susceptible to ambient lighting, resolution, and occlusions, in addition to raising privacy concerns. Moreover, the camera needs to be present in front of the user with a clear focus on the face at all times which might be impractical for applications like wheel-chair equipment control, constant monitoring of emotional state of people with autism. C-Face [34] can track facial motion using earphone cameras, but it still needs clear lighting conditions, and can be privacy-sensitive since the placement of the camera is such that it can capture people in the surroundings. Moreover, wearable cameras can be power hungry for implementation. In contrast to vision based solutions, **EARFace** uses smart earphones that are robust to ambient conditions (lighting/occlusions), and offers ubiquitous and portable tracking, anytime and anywhere.
Table 1. Summary of Main Related Work. For brevity, several other works not in the table are discussed in Section 2. The accuracy metric NME is a normalized mean error in percentage (lower the better), defined formally in Section 6. Portability indicates the ability to sense anywhere without external infrastructure. Robustness to Ambience indicates the ability to be unaffected by ambient conditions such as lighting, occlusions, etc.

<table>
<thead>
<tr>
<th>System</th>
<th>Sensor</th>
<th>Earphone Form-factor</th>
<th>Face Landmark Tracking</th>
<th>3D Face Rendering</th>
<th>Robustness to Ambience</th>
<th>Portability</th>
<th>Accuracy (NME)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAN GT [40]</td>
<td>Camera</td>
<td>N/A</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✗</td>
<td>3.98 %</td>
</tr>
<tr>
<td>SDM [124]</td>
<td>Camera</td>
<td>N/A</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✗</td>
<td>5.67 %</td>
</tr>
<tr>
<td>CFSS [113]</td>
<td>Camera</td>
<td>N/A</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>4.87 %</td>
</tr>
<tr>
<td>BioFace3D [112]</td>
<td>EMG and EOG</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.38 %</td>
</tr>
<tr>
<td>EarIO [69]</td>
<td>Acoustic</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>FaceListener [99]</td>
<td>Acoustic</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>No Tracking</td>
</tr>
<tr>
<td>ECTF [20]</td>
<td>Acoustic</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>No Tracking</td>
</tr>
<tr>
<td>CanalSense [22]</td>
<td>Pressure</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>No Tracking</td>
</tr>
<tr>
<td>EarFS [82]</td>
<td>Electric Fields</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>No Tracking</td>
</tr>
<tr>
<td>EARFace (Ours)</td>
<td>Acoustic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.14 %</td>
</tr>
</tbody>
</table>

**Earables:** Smart earphones have been a popular platform for facial activity detection. EarFs [83] classifies five facial expressions by embedding electrodes into the ear that can sense electric field changes due to facial expressions. CanalSense [22] measures air pressure changes in response to facial expressions for classifying eleven expressions. ExpressEar [108] uses the eSense [56] earphone platform with IMU sensors for facial expression classification. ECTF [20] uses acoustic sensors in the earphone for classifying facial expressions. FaceListener [99] uses headphones to capture ultrasonic reflections from the facial surface to classify facial expressions. PPGFace [35] uses PPG reflections inside the ear to classify seven facial expressions. In contrast to such works that classify predefined discrete facial expressions, EARFace performs continuous 3D tracking of facial landmarks, which can be used by any generic application including facial expression classification. Most similar to our work, BioFace3D [112] tracks facial landmarks. However, the form factor can cause discomfort (usability study in Section 6) because the sensors are placed outside the ears to sense EMG and EOG signals from the skin surface. The sensor protrudes on to the face and head. In contrast, EARFace performs in-ear sensing using a natural earphone form-factor that can be more comfortable. Most recently, EarIO [69] shows the feasibility of tracking facial landmarks. However, EARface also senses outside-ear skin deformation due to facial motion, and therefore, the sensors are placed outside the ears resulting in a bulky format. Thus, the problem of potential discomfort still remains as validated by usability studies in Section 6. Moreover, the sensing can be difficult for users with long hairs as reported by EarIO. In contrast to these works on outside-ear sensing, EARFace performs in-ear sensing to sense the shape of the ear canal, which changes due to facial motion. This allows embedding of the sensors into a natural and small earphone form-factor, thus securing higher usability ratings.

**Other Wearables and Biosensors:** EMG signals are captured by attaching electrodes to the surface of the face at various locations such as cheeks, nose bridge, eyebrows, for classifying various emotions and facial expressions [98]. Similarly, electroencephalography (EEG) electrodes can be attached to the head for capturing brain signals which can be analyzed for emotion and facial expression classification [86]. Electrocardiography (ECG) signals have also been used for capturing the emotional state of the user [55]. However, the electrodes are known to be uncomfortable for wearing under daily usage conditions [19]. Work in [80] designs smart glasses with optical and inertial sensors capable of detecting facial gestures for applications in hands-free user interfaces. CapGlasses [84] embeds transparent capacitive sensors into smart-glasses for detecting facial and head related gestures for applications in tracking emotional and physical well being. Work in [81] embeds 17 photo reflective sensors capable of measuring proximity changes between skin surface and the sensors, thus being able to detect 8 facial
expressions. SonicFace [42] uses an external acoustic speaker and microphone array for capturing reflections from the face for detecting six facial expressions. In contrast to such works, which focus on predefined facial gesture classification, EARFace shows the feasibility of tracking continuous 3D motion of the face, thus enabling a broader class of applications. Furthermore, EARFace uses earphones which are gaining in popularity as a sensing, interfacing, and entertainment platform because of high comfort levels of wearing [44].

Other Applications of Earable Sensing: Photoplethysmography (PPG) sensor integrated in the ear has been exploited for heart rate and blood pressure monitoring [27]. Speaker and microphones in earphones have been exploited for tracking the human hand for applications in user-interfaces [29]. Applications in chewing and eating activity monitoring have also been explored [75]. Speech enhancement and silent speech recognition have been studied using earphone embedded IMUs and acoustic sensors [54, 59, 100, 102, 120]. Tongue gesture sensing for interactive applications using in-ear pressure sensors has been performed [79]. Human activity recognition and exercise monitoring [51, 52], teeth activity sensing [92], augmented reality [115] are some of the other recent applications. Our work is inspired by such prior works that show innumerable sensing opportunities in the area of earable sensing. We extend the literature by enabling an application in 3D facial reconstruction using completely in-ear sensing with a natural earphone form-factor – a first such attempt to the best of our knowledge.

Transformer based ML models: Transformer-based ML models with self-attention mechanism are popular in vision and NLP [39, 41, 107] that exploit relationships not only between various parts of the images or audio but also between inputs and currently decoded outputs. The popular BERT language model [39] uses the transformer architecture and is trained by masking words in sentences, and having the model predict them from context. Transformers have been used in the area of image processing for extracting the spatial relationships across images for applications like automatic captioning [47, 68], super-resolution [71, 78, 114]. Transformers are also popular in applications in automatic speech recognition [31, 85] and speech enhancement [60, 116] because of their ability to extract rich relationships across different speech segments and contextual information. In contrast to prior work, EARFace designs attention-modules to simultaneously capture dependencies across frequencies and time.

Domain Adaptation: Transfer-learning based domain adaptation is popular in vision and speech processing. For example, AlexNet model [63] pretrained on ImageNet database [38] was fine-tuned for classifying images in medical domain[123], remote sensing [45] and breast-cancer [87]. Similarly, a pre-trained BERT language model [39] was fine-tuned for tasks such as text-summarizing [117], question answering [94], etc. This significantly reduces the burden of training for a new task. Domain adaptation techniques are recently gaining popularity in the area of wearable sensing including smartphones, earphones, smartwatches, wearable IMU, etc. For example, IMUTube [64] shows the feasibility of adapting video-based training data for inferring on IMU across multiple users for human activity recognition. As a similar approach, ZeroNet [73] harvests training data from videos and adapts it across users for finger motion tracking with IMU. Work in [26] adapts wrist EMG data across different days and users by using the idea maximum-independence domain adaptation which transforms features based on the Hilbert–Schmidt independence criterion. Likewise, there is another study [72, 74] that leverages wrist EMG data from diverse users to develop multi-user models and employ domain adaptation techniques to unseen new target user. SWL-Adapt [50] proposes domain adaptation model with sample weight learning for cross-user Wearable Human Activity Recognition (WHAR) which could achieve a parameterized network on the new user. AdaptNet [21] propose a semi-supervised bilateral domain adaptation method in human activity recognition which enables information fusion of two different data domains using both unlabeled and labeled data. To predict freezing of gait in Parkinson’s disease and generalize the model to other patients, [106] uses domain adaptation algorithms to address the domain disparity between data from different patients. By keeping
the CNN layers unchanged and by adapting some hidden and fully connected layers, work in [58] increases the robustness of human activity detection across users and devices. Liquidmeter [119] shows the feasibility of fine tuning batch normalization layers to minimize the training data overhead for users for an application in liquid intake monitoring using smart earphones. In a similar spirit, we use a pretrained model from one user and fine-tune it by employing differential layer-wise learning rate and applying stochastic weight averaging for a different user to significantly decrease the training overhead without losing much accuracy.

3 BACKGROUND
We begin with a brief background about the following: (i) Facial motion and the muscles involved. (ii) Facial muscles and their relationship with the shape of the ear canal. (iii) Capturing ear canal shape changes using ultrasonic reflections. (iv) 3D model for representing a human face.

3.1 Facial Muscles and Expressions
Human beings are emotional creatures whose state of mind can usually be observed through their facial expressions. With 43 different muscles, our faces are capable of making more than 10,000 different expressions [7]. The muscles for facial expressions (depicted in Fig. 3a), which are also named as mimetic muscles, can generally be divided into three categories: orbital, nasal, and oral [4]: (i) The orbital facial muscles include three main muscles: Frontalis, Orbicularis oculi and Corrugator supercilii. They mainly control the movement of the eyelids and play a role in protecting cornea from potential injury. (ii) The nasal facial muscles, which are typically composed from three muscles called nasalis, procerus, and depressor septi nasi, are responsible for the motions of the nose and its surrounding skin. (iii) The oral facial muscles, as the name suggests, can manage the movements of the lips and mouth. For example, temporalis is a fan-shaped muscle that helps your jaw close; lateral pterygoid is a fan-shaped muscle that helps your jaw open. These muscles broadly originate from the surface of the skull and insert onto facial skin. Their contraction pulls on the facial skin tending to make various facial expressions on our faces. For instance, facial expressions like fear or surprise consist of eye movements, which are controlled by orbital facial muscles; oral facial muscles contribute to happiness, sadness or contempt by changing shapes of the mouth [118].

3.2 Relationship between Facial Motion and Ear Canal Shape
Each human facial expression can consist of multiple groups of facial muscle movements, and the shape of our ear canal changes according to it. Fig. 3b depicts human ear anatomy. There is a strong relationship between shape changing (distortion) of the ear canal and the mandibular condylar (shown in Fig. 3b) movements [93]. Some Facial expressions such as fear or surprise contain mouth opening motions, which will cause the mandibular
condylar to slide ahead and create a small void. This whole process changes the volume of the ear canal and deforms the tissue inside it, thus changing its shape. Besides, the ear canal is connected with the temporalis muscle through the mastoid (shown in Fig. 3b) [20]. Temporalis is the largest muscle on our head (Fig. 3a), and can even forward mechanical artifacts from other muscle groups, for example, orbital and nasal muscle groups, towards the ear canal [83]. By exploiting these properties, EARFace can detect movements like raising the eyebrow, wrinkling the nose, or motion of the mouth by observing changes in the shape of the ear canal. In the next subsection, we outline the process for tracking changes in the shape of ear canal via ultrasonic reflections, which will ultimately lead to 3D reconstruction of the human face.

3.3 Capturing Ear Canal Shape via Acoustic Reflections

Air column resonance is caused due to superimposition of incident and reflected waves inside a tube. When we wear earphones, our ear canal with ear buds can be regarded as a short air tube with a length of 2-3 cm, so generally, the frequency of the first harmonic of the resonance is 5 to 7 kHz, and then can have further harmonics of 10-14 kHz, 15-21 kHz and even higher frequency band [104]. The strengths and attenuation of the resonance can change according to the differences in the shape of our ear canal, and because of this, distinct ear canal shapes can be extracted by transmitting a wide frequency range of swept signal and analyzing the reflections [6, 48]. Accordingly, EARFace computes the CFR of the ear canal which captures such effects over a wide range of frequencies. CFR computation is outlined in Section 5.1, which is ultimately used by the ML models in EARFace for facial landmark tracking. We measure the variation in Channel frequency response (CFR) of the ear canal reflections as a function of facial expressions across different users. Based on the experiments in Fig. 4, we note that the CFR is consistent over time for the same facial expression for same user and different across different facial expressions. We believe this is promising evidence of the feasibility that the CFR can be used for tracking facial landmarks. Later experiments in Section 6.3 confirm the high accuracy of tracking the facial landmarks using the CFR. Across different users, although there are some differences in CFR due to the differences in shapes across individual users, domain adaptation techniques are able to handle such differences and exploit the similarity across users for consistently tracking facial landmarks across different expressions and users.

3.4 3D Facial Modeling using FLAME

EARFace uses the popular Faces Learned with an Articulated Model and Expressions (FLAME) [70] model for representing the face in 3D. FLAME is a statistical model that allows expressing complex facial geometry and articulations mainly using the following three simple sets of parameters in three highly compressed spaces respectively: (i) Shape Parameters ($\beta \in S = R^{||\beta||}$): Defines the deformations in the face due to the unique identity of the subject. (ii) Pose Parameters ($\theta \in P = R^{||\theta||}$): Defines the deformation in the face because of the rotation of the head around the neck or motion of the jaw. (iii) Expression Parameters ($\bar{\psi} \in E = R^{||\bar{\psi}||}$): Defines the deformation in the face due to facial muscle activation that can change the expression of the face to happy, sad, anger, etc. Ultimately, the FLAME model, maps these three sets of parameters into 3D locations of meshes (N = 5023 vertices) that form the face as indicated in the below mathematical model.

$$FLAME(\tilde{\beta}, \tilde{\theta}, \tilde{\psi}): R^{||\tilde{\beta}|| \times ||\tilde{\theta}|| \times ||\tilde{\psi}||} \rightarrow R^{3N}$$

Each of the three sets of parameters above (shape $\tilde{\beta}$, pose $\tilde{\theta}$, and expression $\tilde{\psi}$) are represented in a highly compressed space, the bases (i.e., $S$, $P$, and $E$) of which are determined by performing PCA on a dataset of 33,000 3D facial scans, spanning a wide range of ages, ethnicities, genders, etc. In details, FLAME starts from a neutral template mesh $T \in R^{3N}$. To take the shape variation of different identities into consideration, a linear shape blendshape function $B_\Sigma(\tilde{\beta}; S) = \sum_{n=1}^{||\tilde{\beta}||} \beta_n S_n$ adds the offsets to the template $T$, where $\tilde{\beta} = [\beta_1, \beta_2, ..., \beta_{||\tilde{\beta}||}]^T$ denotes
the coefficients for shape variations and $S = [S_1, S_2, ..., S_J]$ denotes the orthogonal shape basis. Similarly, the offsets calculated from a pose blendshape function $B_P(\tilde{\theta}; \mathcal{P})$ and an expression blendshape function $B_E(\tilde{\phi}; \mathcal{E})$ will be added into the template for correcting the pose deformations and facial expressions, respectively. As a result, the reconstructed 3D mesh will be the neutral template with added shape, pose, and expression variances. This allows for compactly representing complex facial geometry as well as articulation. Fig. 5a shows an example where the action of each of the three FLAME parameters ($\tilde{\beta}, \tilde{\theta}, \tilde{\phi}$) is depicted. Despite FLAME being precise and accurate and compatible with existing rendering methods [28, 91], FLAME decomposes faces into three highly compressed spaces, thus more appropriate for real-time applications. Described in Section 5.4, EARFace uses FLAME for converting the facial landmarks (depicted in Fig. 5b) predicted by the ML model into a 3D reconstruction of the face.

3.5 Difference between Facial Reconstruction and Facial Landmarks Tracking

Human face has a complex geometry and it is capable of sophisticated articulation. Accordingly, the facial landmarks (eyes, nose, mouth, etc, with total 51 landmarks) tracked in EARFace provides a convenient way of representing the 3D facial geometry in a compressed space [25, 103, 112] thus enhancing the accuracy and efficiency of tracking them using machine learning with limited training data. Finally, we reconstruct the 3D facial model of the user using these landmarks for animation and qualitative analysis purposes. Accordingly, we provide results of 2D facial landmark tracking (quantitative) as well as 3D facial reconstruction (qualitative) in the paper. While landmark tracking can provide a quantitative measure of the performance of our system,
Fig. 5. (a) Parameterization in FLAME [70]. Starting from a neutral face with no expression, deformations of shape, pose, and expression can be captured by the corresponding parameters (b) Key Facial Landmarks used in EARFace for 3D facial reconstruction.

Fig. 6. Smart earphone platform: Earphone module

3D facial reconstruction is more qualitative and it is suitable for applications in animations, 3D avatars, scene reconstruction, etc [65, 67]. We leave a thorough investigation of the application space for future research.

4 PLATFORM DESIGN

Because of requirement of long range of sound frequencies until 40 kHz, EARFace develops its own earphone platform. In this section, we outline the design and implementation details of a new portable platform in the form-factor of an earphone to capture the shape of the ear canal. Fig. 7 shows the architecture of the earphone platform which consists of a earphone module and a data acquisition module. The details are elaborated in subsequent paragraphs.

**Earphone Module:** The earphone depicted in Fig. 6 is embedded with an ultrasonic speaker (Sonion EST65DB01 [10]) with a frequency of operation from 10 Hz to 70 kHz. The speaker illuminates the ear canal to be able to capture the shape via reflections. Towards capturing these reflections, we embed a MEMS microphone (Sonion P11AC03 [11]) with a frequency of operation ranging from 18 Hz - 80 kHz as shown in Fig. 6. The earbud is designed to conveniently fit the varying size of ear canals of different users', developed using 3D printing material Thermoplastic polyurethane (TPU) material for flexibility and comfortable skin contact, as shown in Fig. 6. The depth of the earphone placement is completely natural and it can accommodate different users’ comfort and habit.
thus allowing an overall comfortable experience of wearing. To assess the robustness of the earphone placement, we conducted a test where we removed and remounted the earphones to simulate natural variations in position, as described in Section 6.3. This facilitates capturing of ear-canal shape with sufficient resolution which is used for facial motion tracking.

**Data Acquisition Module:** As shown in Fig. 7 the core of the data acquisition module is a Teensy 4.0 [16] micro-controller. It has a single-core ARM Cortex-M7 that runs at 600 MHz. We add Teensy Audio board [14] to control the speaker. We integrate a Texas Instruments OPA344 [13] rail-to-rail precision amplifier to amplify audio signal from the microphone. Next, the micro-controller’s inbuilt Analog to Digital Converters (ADC) converts the ultrasonic reflections captured by the microphone into samples with a 12-bit resolution at a sampling rate of 80 kHz. To wirelessly stream collected data, we add ESP32 [2] as a co-processor with built-in WiFi module. The collected data is streamed over WiFi to a smartphone for running the ML models that track facial motion. We use a Li-Po 500 mAh battery to power the data acquisition module and the earphone module. The power consumption of the module is about 245 mA and its form-factor is similar to a smartphone and fits in the pocket.

**Software:** The software side of the acquisition module includes three main components: (i) Controlling the speaker; (ii) Collecting audio data from the microphone; (iii) Streaming audio data from acquisition module to smartphone. The software is implemented on the Teensy 4.0 board using Teensyduino [15] and appropriate library for playing MP3, DA/AD converting data and streaming audio data to smartphone by WiFi [17, 101, 109]. The audio data is collected by Teensy 4.0 with high sampling rate 80 kHz and streamed by WiFi using AESP32 libraries WiFiNINA [17] to smartphone for data processing. More details on data processing are discussed in Section 5.

5 ACOUSTIC SENSOR DATA TO 3D FACIAL RECONSTRUCTION

In this section, we provide an overview of signal processing and machine learning components involved in transforming the acoustic data from earphones into a 3D reconstruction of the user’s face. It consists of four main components. (i) Estimating the CFR of ear canal reflections and feature extraction. (ii) Design of a transformer based ML model to estimate the facial landmarks using CFR data. (iii) Efficient domain adaptation for decreasing the training overhead. (iv) 3D facial reconstruction using the landmarks predicted by the ML model.
5.1 Extracting Ear Canal CFR

Channel Frequency Response Estimation: We capture the channel frequency response (CFR) of the ear canal in the acoustic domain to track changes in the shape of the ear canal, which can capture the facial motion (discussed in Section 3). Towards this end, a linear sine-sweep signal [12] over 16-40 kHz of length 4000 samples (0.05 seconds) is first generated and played through the embedded speaker in the earphone. The SPL level is about 60 dB which is pretty much below the thresholds for health and safety as our speaker component was tested for compliance with safety regulations by the manufacturer (more details in Section 7). The 16-40 kHz signal ensures that the transmitted sound is within the ultrasound frequencies to eliminate audible noise for the user as well as avoid interference from background noise. On the other hand, with high end of the frequency extending till 40 kHz, this provides enough resolution to track fine-grained changes in the ear canal shape thereby enabling high precision tracking of facial motion. The microphone will record the reflections of the speaker’s transmitted sound from the ear canal. With the knowledge of the known sine-sweep sequence, the CFR is estimated as follows.

\[
H(k) = \frac{Y(k)}{X(k)} \quad \forall \ k \in \{1, 2, \ldots, 4000\},
\]

where \(X\) denotes the FFT of the transmitted sine-sweep signal, and \(Y\) denotes the FFT of the received sine-sweep signal. \(H\) is the estimated CFR. The CFR is tracked over a sliding window (step size of 0.033 seconds) continuously with a frequency of 30 Hz. LFCC features as elaborated next are extracted from the CFR for further processing.

Feature Extraction: CFR extracted in the previous step is processed further to extract Linear Frequency Cepstral Coefficients (LFCC) features as explained below. Frequency filters as depicted in Fig. 8 are applied to the CFR. Each filter extracts energies in the appropriate parts of the spectrum. This provides a compact representation of the high resolution CFR data, thus helping decrease the number of parameters in the ML model. We note that typical speech recognition applications use non-linear filters (with log scale in MFCC [62] features) to mimic the auditory response of the human ear with more discriminative power at lower frequencies. In contrast, with a different application in-ear canal shape detection, EARFace adopts linear scale with LFCC filters so as to capture sufficient information from all frequencies, and particularly at higher frequencies which can capture finer changes. The LFCC features thus computed serve as input to ML models described next for tracking facial motion.

5.2 Facial Landmark Estimation

The high level architecture of the ML model is depicted in Fig. 9. While the model is developed based on the popular transformer architecture, our design fuses a spectral transformer (for exploiting rich relationships in the data across frequencies) with a temporal transformer that exploits relationships and dependencies in the data over time. The various components of the architecture are elaborated next.

Input: The LFCC features extracted from CFR serve as the input to the model. 9 successive frames of 100 LFCC features (hyperparameters chosen based on grid search) at 16-40 kHz from both earphones (dimension 9 \(\times\) 100 \(\times\) 2) are used as inputs instead of a single frame (each frame is of dimension 100 \(\times\) 2). This helps the model exploit rich relationships across time in performing the predictions of facial landmarks.

Spectral Transformer: Each of the individual frames is first passed through a spectral transformer, where the LFCC features are converted into intermediate representations that encode the spectral relationships across different frequencies. Depending on the facial motion or expression being performed by the user, different muscles in the nasal, orbital, and oral regions move at a different rate, and this manifests as a characteristic spectral pattern in the changes in the shape of the ear canal. The spectral transformer extracts such dependencies in its
representations. The encoded representations serve as input to the temporal transformer.

**Temporal Transformer:** The temporal transformer takes as input representations from individual times which capture dependencies across frequencies. The representations are further enhanced by exploiting the relationships across time. Similar to natural languages where surrounding words offer context about what the next word could be, a lot of contextual information can be exploited across time. For example, when user is smiling or expressing surprise, there will be a progressive change in various groups of facial muscles which follow a specific trend. The temporal transformer encodes such rich context into the representations, which can be further utilized for robust inference.

**Transformer Encoder:** The individual input CFR frames of dimension $100 \times 2$ are passed through a linear projection layer that outputs a tensor that is of size $100 \times 1$. The input is then passed through two transformer encoders. Within each encoder, positional embeddings are first added to the data which are later utilized in
the self-attention layer to encode dependencies (across frequencies and time) into the learned representations. The encoder consists of a self-attention module (elaborated next) within the multi-head attention layers, and a multilayer perceptron (MLP), in addition to normalization layers. Residual connections are incorporated at both multi-head attention and MLP layers to boost the convergence of the ML models [46].

The Role of Self-Attention Layers: The heart of the transformer encoder lies in the self-attention module, which creates representations for the input that captures dependencies over time and frequencies. We explain the self-attention layers in the context of the temporal transformer that exploits dependencies across time. The action of the spectral transformer that exploits dependencies across frequencies is analogous. Towards capturing temporal dependencies, the self-attention module in the temporal transformer firstly computes query ($q_i$), key ($k_i$), and value ($v_i$) vectors for $i^{th}$ input frame’s embedding by multiplying it by weight matrices $W_q$, $W_k$, and $W_v$ respectively obtained during training. Secondly, an attention-score is computed for each input frame by performing a dot product between the query and key values. Specifically, the dot product $q_i \cdot k_j$ represents the attention-score between $i^{th}$ and $j^{th}$ frames. These scores capture the temporal dependency between various parts of the input. Finally, the output of an attention-head for input frame $i$ is the normalized sum of attention-scores weighted by the value vector ($\sum_{i=1}^{N} (q_i \cdot k_i)v_i$). Similar to how multiple convolutional kernels can capture different patterns, we use six attention heads with different projected versions of queries, keys, and values, and fuse their attention scores together in a combined representation. This captures a representation for $i^{th}$ input frames in a given attention module. The two attention modules in the two transformers (spectral and temporal transformers in Fig. 9) capture dependencies over time and frequencies respectively for all input frames.

Regression Head and Final Output: The output representations from the temporal transformer are passed through a fully connected layer which converts the representations into facial landmarks as depicted in Fig. 5b. These landmarks are critical for 3D facial reconstruction. The actual rendering of the 3D model of the face using these landmarks is elaborated in the next subsection.

Loss Function and Normalization: The ML model is trained by minimizing the mean squared error (MSE) loss of the predicted landmarks by the model and those predicted by a camera based ground truth [24]. The locations of facial landmarks as captured by the vision ground truth can vary due to changes in head location and pose. Towards consistently tracking the facial expressions independent of the head location and pose, we normalize the extracted facial landmarks to a consistent frame of reference by performing the following sequence of transformations on the landmarks captured by the ground truth. (i) The tip of the nose is chosen as the origin. (ii) The direction from the centre of one eye to the centre of the other eye is considered as the direction of x-axis. (iii) The landmarks extracted are rotated such that the line joining the two eyes is aligned with the x-axis, and translated such that the tip of the nose is at the center. (iv) Finally, the landmark locations are scaled uniformly about the origin such that the interocular distance is set to 1. The interocular distance is defined as the distance between the pupils of the left and right eyes when the eyes are in normal fixation, which can be captured by the camera.

5.3 Decreasing Training Overhead via Domain Adaptation
For the machine learning model proposed above, training separate models for each user will be burdensome. Therefore, we explore domain adaptation strategies to pretrain a model with one (source) user and fine-tune it to adapt to new users with low training overhead.
The main steps in the domain adaptation process are as follows: (i) We generate a model for one user (source) by extensively training the model with labeled data from that user known as the pretrained model. (ii) We collect small training data with only few labels from the new (target) user. Instead of developing the model for the target user from scratch, we initialize the model weights to be same as the pretrained model. (iii) We make various layers in the model trainable at different learning rates so as to increase the efficiency of domain operation with limited training data (elaborated next). Using the few labels from the target user, we update the trainable layers to minimize the loss function. This is called fine tuning. The model thus generated will be used for making inferences on the target user. We explore the below complementary approaches for performing the domain adaptation.

(i) **Grouped Layer-wise Learning Rate:** During domain adaptation, we apply discriminative learning rates for different layers instead of using the same learning rate for all layers of the transformer model. This is expected to improve the efficiency of domain adaptation for deep neural networks as shown for tasks such as text classification [49, 76, 77, 121]. Specifically, we break all the layers into two groups: hidden layers and self-attention layers. The learning rates are set differently for these two groups. The learning rate for hidden layers is 2e-3 and the decay rate is 0.99 while the learning rate for self-attention layers is 1e-3 and the decay rate is 0.9. The intuition is that the hidden layers encode user specific information whereas the self-attention layers encode more generic information useful across multiple users. Therefore adapting the hidden layers with a larger learning rate will likely generate a model that fits the new user with high accuracy. Moreover, the hidden layers have smaller number of parameters size than self-attention layers thus making the domain adaptation easier with small amounts of training data.

(ii) **Stochastic Weight Averaging:** While deep neural networks are typically optimized using Stochastic Gradient Decent (SGD), we incorporate latest advances into our design based on stochastic averaging for better stability [53]. Specifically, we take the average weights during the last 25% of the total 300 epochs during training time. This is expected to further enhance the accuracy and generalization during domain adaptation with group-wise learning rates as explained above.

5.4 3D Rendering

We render a 3D model of the user’s face that provides a realistic visualization of the facial expression and articulation. Towards this end, we perform an optimization that finds the best set of FLAME model parameters described in Section 3.4 (shape, pose, expression), that minimizes the error between the set of 2D landmarks predicted by EARFace and the ones projected into 2D from the FLAME model parameters. Before optimization, the 3D head model is initialized with a default template. Then, the optimization is performed in two steps: (i) In the first step, camera calibration is achieved by optimizing the parameters of scale, rotation, and translation, thus minimizing the $L_2$ error between the 2D landmarks predicted by EARFace and corresponding 2D projection from the current 3D head model in FLAME. (ii) In the second step, the FLAME model parameters (shape, pose, expression) are adjusted to further minimize the $L_2$ error between the projected 2D points from the FLAME model and the ones predicted by EARFace. The above two steps are iterated multiple times until the $L_2$ error converges. This ultimately renders a 3D face shape with a realistic appearance, facial expression, and articulation that best fits the facial landmarks predicted by the ML model. Note that 3D rendering of EARFace does not require camera parameters because 3D faces are rendered at relative scales from a predefined template mesh. We found this 3D rendering is sufficient for qualitative purposes because as shown in Fig. 2, by iteratively optimizing using the

---

1We adopted code from FLAME’s official codebase: https://github.com/TimoBolkart/TF_FLAME/
above steps, any deformation that accounts for the user’s pose, facial expression, etc., can be reflected. Therefore, EARFace is free of camera settings, which makes the system more ubiquitous.

6 EVALUATION
In this section, we evaluate the EARFace system. Below, first we summarize our findings and later describe the data collection and present the detailed results.

- The overall accuracy of facial landmark tracking is 1.83 mm which is consistent across diverse users and comparable to prior works including vision based systems.
- The accuracy is consistent across days and multiple sessions with the sensor removed and remounted thus indicating robustness to natural variation in sensor position and orientation.
- The accuracy generalizes to various facial motion and is consistent across different facial regions such as mouth, nose, and eyes, as well as robust to body and head motion.
- The ML models are implemented on smartphone devices with a latency of 13 ms and a low power consumption profile.
- Usability studies depicts the benefits of earphone based sensing platform in EARFace over alternative platforms due to EARFace’s design with a small form-factor, in-ear sensing with appearance of a typical earphone, and comfort levels particularly with long duration of wearability.

6.1 User Study

Data Collection Methodology: Our study was approved by the IRB committee. The users wear the smart earphones (Section 4) with embedded acoustic sensors as shown in Fig. 6 on both ears. 20 users participated in the study with 12 males and 8 females, with their ages ranging from 21 to 55 and body weight ranging from 56 to 94 kgs. The users were then instructed to perform a sequence of facial motion cycling through following universal expressions of emotion [7] in a random order: happy, sad, anger, contempt, fear, disgust, and neutral. These expressions are depicted in Fig 2, and the users were shown pictures of these expressions before and during the study. While the facial expressions are provided as a guideline to capture diverse motion, EARFace is designed to track not only the expressions but also the continuous transition of facial landmarks between them. Elaborated later in this section, we also validate EARFace for other facial motion without the above facial expressions which indicates the generalizability of the system to arbitrary facial motion. In addition, to assess the robustness of the earphone placement, we conducted a test where we removed and remounted the earphones to simulate natural variations in position. We followed up with all the users over multiple days to evaluate the robustness over time and natural variation in the placement of the earphone. Finally, we also conduct experiments where the users also performed other activities such as walking or moving the head randomly so as to evaluate the robustness due to body motion.

Labels for Training and Testing: While the users perform the facial motion, acoustic data from both earphones were collected at a sampling rate of 80 kHz. Towards, validating the accuracy of EARFace, videos of the user’s face were captured. We use the front-facing camera of a smartphone [96] to collect the ground truth such that the full view of the user’s face is included in the camera. Facial landmarks are extracted from the video using techniques in [24], which serve as the ground truth. The facial landmarks predicted by EARFace are compared with the above ground truth using the MAE and NME error metrics as elaborated further in this section.

Training Data: Each session of data collection lasted for 60 seconds. A total of 24 sessions were conducted for each user with sufficient rest in between sessions. The earphone sensor was removed and remounted across sessions to validate the robustness to natural changes in sensor positions during daily usage. 5 mins of data (5
sessions) was used for training (accuracy saturates after that based on Fig. 11b) and the rest was used for testing in a randomized cross-validation fashion. With domain adaptation, a pretrained model from one user was fine-tuned to a new user with only 2 minutes (2 sessions) of user-specific training data thus decreasing the training overhead further. Users can conveniently use their smartphone front-facing cameras for collecting the small training data that is needed for domain adaptation. This is a one-time task. The users need to expose their complete faces to the camera and perform a set of basic facial expressions. However careful placement or alignment of the camera is not needed because the extracted landmarks from the camera will be normalized to a standard framework – more details as discussed in Section 5.2. Therefore, we believe this training data is not a big overhead.

**Evaluation Metric:** We evaluate EARFace using the following two metrics. (i) We use the Mean Absolute Error (MAE) metric which is defined as follows. 

\[
MAE = \frac{1}{N} \times \sum \| e - v \| \times \frac{d_{on}}{d_{or}}
\]

where \(v\) and \(e\) denote the landmarks extracted from vision ground truth and EARFace in a normalized coordinate frame (discussed in Section 5.2), \(d_{or}\) denotes the real interocular distance in mm (we measure \(d_{or}\) for each user who participates in the study) and \(d_{on}\) denotes the interocular distance in the normalized coordinate frame \(d_{on} = 1\) from Section 5.2. This provides the MAE in units of mm which is the primary metric used for evaluating EARFace. (ii) Normalized Mean Error (NME) is a metric popular in the computer vision community for validating the accuracy of facial landmarks. To compare with computer vision and other prior works, we also compute NME. This is the mean error between the ground truth and reconstructed landmarks, normalized by a factor of \(d_{on}.\)

\[
NME = \frac{1}{N} \times \sum \frac{\| e - v \|}{d_{on}} \times 100\%
\]

where \(d_{on}\) is the interocular distance in the normalized coordinate plane as defined in Section 5.2.

6.2 Implementation

EARFace is implemented on a combination of desktop and smartphone devices. The ML model is implemented with TensorFlow [18] packages and the training is performed on a desktop with Intel i7-8700K CPU, 16GB RAM memory, and Nvidia GTX 1080 GPU. We use the Adam optimizer [61] with a learning rate of 0.001. To avoid over-fitting issues that may happen in the training process, we add dropouts [110] with a parameter of 0.5 following each RELU activations. Once a model is generated from training, the inference is done entirely on smartphone devices using TensorFlowLite [43] on Samsung S20, and Oneplus 9 Pro smartphones [89, 96].

6.3 Performance Results

**Qualitative Results:** Fig. 2 depicts the qualitative rendering results in EARFace across various facial expressions. Evidently, the 3D rendering closely follows the real expression of the user. For example, the contempt expression indicates how the 3D rendered model accurately follows the asymmetric motion of the mouth. The surprise expression shows the ability of the model to capture the raising eyebrows and mouth wide open. The smiling expression captures the motion of the cheek. Fig. 10 depicts how the facial motion varies as the expression changes from a neutral expression. Overall, the 3D model captures all facial features such as mouth, eyes, cheek, nose, etc. Furthermore, a demo video is included in the url [37] where even the blinking of the eye is evident. We believe these results are encouraging.

**Accuracy vs Users:** Fig. 11a depicts the accuracy across users who participated in the study. The graph displays both user-dependent and domain adaptation results. For user-dependent results, we use 5 minutes of training data, while for domain adaptation results, we use 2 minutes of training data from the target user. Therefore, this is a reduced dataset than the full dataset. The pretrained model for domain adaptation comes from a random user because our experiments suggest that the accuracy will only vary by 0.1mm when we vary the user from which we obtain the pre-trained model. Evidently, the accuracy is consistent across users with diversity in facial shapes, natural posing behaviour, gender, etc. Regardless of the characteristic of the user, the ML models

ACM Trans. Internet Things
in EARFace can learn the mapping between facial motion and the ear canal shape, and learn to recognize them through ultrasonic sensing. Therefore, we believe the sensing and ML techniques in EARFace generalizes across multiple users. The overall mean accuracy is 1.83mm (MAE) and 3.17 (NME), which is comparable to vision based systems (elaborated later) and suitable for several applications including facial animation generation, emotion recognition.

Fig. 10. Progression of facial feature changes as predicted by EARFace.

Fig. 11. (a) Accuracy vs Users. EARFace achieves consistent accuracy across all users (b) Accuracy as a function of size of training data. EARFace’s ML models converge with relatively low overhead in training.

Accuracy vs Size of Training Data: Fig. 11b depicts the accuracy as a function of the size of training data. Evidently, even with 1 minute of training data, the accuracy is already at 4.36 mm whereas it saturates at 5 minutes of training data to about 1.83 mm. Given the training overhead is only a one-time cost for a few minutes, we believe this is not an overhead.

Decreasing Training Overhead by Domain Adaptation: To further decrease the training overhead, a pre-trained model from a different user was taken and fine-tuned using techniques in Section 5.3 such that only a small fraction (120 seconds) of user-specific training data is used for developing a model for the user. Fig. 12a and Fig. 12b compares the difference between a user-dependent model and the model with domain adaptation. The requirement of training data can be significantly reduced by domain adaptation without much degradation in
Accuracy vs Days: Fig. 12a depicts the accuracy over different days of the user study. Although the earphones can fit snugly, there might be small variations in earphone position across days. The training and test data sets were sampled across completely different days to validate robustness to variation in earphone positions. The accuracy is consistent across days because the training data incorporates such diversity thus enhancing the robustness of the models.

Accuracy vs Facial Expressions: Fig. 12b depicts the accuracy as a function of various facial expressions. Minor variation in accuracy occurs across facial expressions depending on the range of motion of various facial features. EARFace is able to track various expressions with accuracy close to 2 mm. This indicates the viability of EARFace in modeling complex facial motion occurring in daily life.

Accuracy vs Facial Features: Fig. 13 provides a breakup of the accuracy as a function of different facial features: eyebrows, eyes, nose, and mouth. Evidently, the accuracy is consistent across all features on the face. The eyes and nose have a slightly lower error because of their smaller range of motion in comparison to eyebrows and mouth. Nevertheless, the overall accuracy is close to 2 mm for all the facial features, which indicates reliable tracking.

Accuracy vs Number of Earphones: Fig. 14a depicts the accuracy as a function of number of earphones. While usage of both earphones can provide the best accuracy by integrating sensor data from both ears, the accuracy with individual earphones (left or right) is also close (≈ 2.40 mm, 2.36 mm). This provides the opportunity to integrate the acoustic earphones in only one of the earphones while leaving the other one open for potentially integrating other sensors.

Generalizability of the Model: Our machine learning models are trained with universal facial expressions as discussed earlier. These universal expression are known to include most movements of the facial parts in daily life. Therefore, we expect the model to generalize to any facial motion that might not be exactly identical to these universal expressions. To validate this, we conduct new experiments involving arbitrary motions as shown in Fig. 14b, which shows a consistent accuracy for capturing various facial motion. We believe this indicates the generalizability of EARFace to any facial motion. We want to emphasize that we have considered and integrated all of the unseen facial expressions of our users. Although the graph only shows one user’s photo to represent
Fig. 13. Accuracy variation as a function of Facial Landmarks (a) Variation across individual landmarks (b) Region specific variation

Fig. 14. (a) Accuracy vs Number of Earphones. Both earphones together achieve the best accuracy whereas individual earphones also provide reasonable levels of tracking (b) Tracking in EARFace generalizes to arbitrary facial motions

the unseen facial expression, the results comprise data from all 20 users. According to our findings, our system can accurately track unseen facial expressions across different users.

Robustness to Body and Head Motion: We conducted additional experiments where users walk naturally, as well as shake the head and change the head pose to emulate normal usage scenarios of body and head motion. During the process, we also followed the user with a camera but we had to manually change the camera angle to clearly capture the user’s face for ground truth purposes. The MAE under this setting was 1.90 mm whereas the MAE when the user is sitting on a desk is 1.83 mm. Body and head motion can cause minor variations in earphone position and orientation. However, EARFace is robust to natural variation in sensor position as evaluated in Fig. 12a. Moreover, the earphone is snugly fit to the extent possible. Thus, we did not notice any significant changes in accuracy due to body and head motion.

Comparison with Vision and Other Systems: The last column in Table 1 directly contrasts the accuracy in EARFace with state-of-the-art vision systems and wearable systems. Since vision based systems typically only provide a relative error based on NME metric (because camera parameters may not be available) instead of the MAE metric which is absolute, we compare all systems using the NME metric. Note that EarIO [69] computes neither NME nor MAE (in mm) but only a relative MAE with respect to the camera used for ground truth. We did not find sufficient details in the EarIO paper about camera processing to convert their metric into an NME or absolute MAE in mm. However, we were able to compare EARFace using the NME metric with other systems.
The vision based works are evaluated on datasets with manually labeled ground truth that can be highly accurate. Because EARFace's dataset is self-collected and ground truth is based on a ML model that automatically extracts landmarks from cameras [24], the comparison may not be entirely fair. Nevertheless, we believe a comparable accuracy to vision based systems is encouraging. Overall, the accuracy in EARFace is comparable to prior works while offering extra benefits of robustness to occlusion/lighting/head-motion, ubiquitous sensing anywhere (without needing a camera in front of the face always or other external infrastructure), and higher levels of comfort owing to in-ear sensing with an earphone form-factor.

Ablation Studies: Table 2 depicts the accuracy as a function of different configurations of the ML model. A basic CNN architecture in an encoder-decoder form [23] achieves an MAE of 2.25 mm. However, the transformer architecture can capture richer temporal dependencies, which decreases the MAE to 1.93 mm. Finally, the design in EARFace that incorporates both spectral and temporal transformers achieves the best MAE of 1.83 mm.

Latency and Power Consumption: The power consumption of the sensor device itself was discussed in Section 4. Here, we analyze the latency and power consumption of the ML model of EARFace as implemented on smartphones. Fig. 15a depicts the latency of executing ML models in EARFace. Fig. 15b depicts the power consumption of executing ML models in EARFace on smartphone devices. Evidently, the latency for executing the ML model for computing facial landmarks is only 13 ms. For profiling the energy of the TensorflowLite model, we use Batterystats and Battery Historian [5, 66] tools. We compare the difference in power between two states: (i) The device is idle with the screen on. (ii) The device is making inferences using the TensorflowLite model. Fig. 15b depicts a low power consumption profile, with the discharge rate supporting upto 8-9 hours of continuous operation. The device can be put under sleep when not in active use to extend the battery life further.

Usability Study: We conducted a user experience survey among the study participants who wore the sensors continuously under free-living conditions for six hours at their apartments. Each user wore all of the different platform so that they can compare the platforms with each other. The users conducted normal daily life activities which included working on their laptops (typing, browsing, etc), eating, drinking, watching movie, etc while they continue to wear the sensor. We compare EARFace and three other alternatives: (i) OpenBCI based sensing as employed in BioFace3D [112] where the electrodes surround the ear and the controller device sticks to the

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder Decoder (CNN)</td>
<td>2.25</td>
</tr>
<tr>
<td>Transformer (Temporal)</td>
<td>1.93</td>
</tr>
<tr>
<td>Transformer (Temporal + Spectral)</td>
<td>1.83</td>
</tr>
</tbody>
</table>
back of the head. (ii) Headphone based sensing as used in FaceListener [99]. (iii) EarIO [69] based platform where sensors are placed outside the ear to capture facial skin deformation due to facial expressions. Note that we only create a dummy prototype of EarIO based on size and weight specifications in the paper because full details of the electronics is not publicly available. Participants rated the four devices anonymously based on comfort, weight and appearance from 0 to 10. Fig. 16 depicts the results. Higher the rating, better the usability of the device. BioFace3D, FaceListener and EarIO platforms were perceived bulky, whereas earphone devices designed in EARFace are miniature enough to get a high rating on the weight feature. Also, some of the users complained about the discomfort in wearing the BioFace3D platform because it surrounds the ear and the user needs to have a device at the back of the head. Similarly, the EarIO platform protruding outwards and behind the ear was a matter of concern for many users, more so for users with longer hairs. While FaceListener platform rated better than the BioFace3D and EarIO form-factors on comfort, some users expressed feeling pain in the ear, especially for longer duration of wearability. On the otherhand, the earphone platform in EARFace scored higher across all three dimensions owing to its small form-factor that fits naturally within the ears. Because of higher levels of usability ratings, earphones are gaining in popularity with many applications in mobile health [36, 51], user interfaces [29], speech enhancement [59, 102], etc.

7 DISCUSSION, LIMITATIONS, AND FUTURE WORK

Safety of Ultrasound Sensing: We have enough evidence that our device is safe for daily wearing, since: (i) Based on Centers for Disease Control and Prevention(CDC)[9] recommendations, a person can continuously be exposed to 85 dB over 8 hours in the work space. Additionally, U.S. Environmental Protection Agency (EPA) recommends an average exposure level of less than 70 dB over a 24 hour time period [88]. Our device only has a power level of 60-65 dB when emitting ultrasound signals, lesser than both of the above limits. The actual average exposure is likely to be much lower since the device is used only when tracking is needed; (ii) Accordingly, speaker components (EST65DB01) of our hardware manufactured by Sonion have passed safety legislation around ultrasound to confirm that the power levels are under safety limits. Thus, we believe the usage of ultrasound for sensing does not cause adverse effects.

Impact of Head and Body Motion: The data was collected while users perform natural head and body motion. In addition, as identified in Section 6 we conducted experiments while having the users move their heads in a more pronounced manner and walk naturally. Because the earphone platform is snugly fit to the ears and otherwise robust to minor variation in sensor positions (Fig. 12a), the accuracy is not impacted by body and head motion.
Wireless Earphone: In the current form, the earphone is connected to a smartphone-like controller with wire-
less streaming of sensor data that lets the user move freely while using the device for facial motion tracking. In the
future, we plan to develop a fully wireless earphone like Apple AirPods [1] by exploiting advances in True Wireless Stereo (TWS) [8], in which two earbuds could communicate with smartphone simultaneously using Bluetooth.

Applications: EARFace designs a generic pipeline for facial motion tracking using earphone platforms that
are gaining in popularity. Without needing a camera to always face or follow the user or dependency on external
infrastructure, EARFace can provide ubiquitous sensing anywhere and anytime. We believe several applications
can be built on the top of EARFace such as facial expression recognition for sensing emotional well being, driver
behavior monitoring, fatigue and stress detection, AR/VR applications, user interfaces including accessibility
applications, etc. We leave a thorough investigation of the application space for future research.

8 CONCLUSION

This paper designs EARFace, a system that showed the feasibility of reconstructing the 3D face of a user with
in-ear sensing. To enable EARFace, an earphone platform is designed with microphones and speakers that can
track the shape of the ear canal by sensing high bandwidth ultrasonic reflections. By tracking changes in ear
canal shape, EARFace can track the facial motion. A transformer architecture is designed to exploit the spectral
and temporal dependencies of ear canal CFR, ultimately leading to highly accurate and continuous tracking of
facial motion. An extensive study with 20 users provides an accuracy of 1.83 mm in tracking facial landmarks.
The ML models run on smartphones with a latency of 13 ms and low power consumption. Despite progress, we
believe we only scratched the surface. A number of applications in the areas of affective computing, emotional
well being monitoring, facial expression recognition, AR/VR, accessibility and user interfaces can be explored.
We leave this to future work.

REFERENCES


ACM Trans. Internet Things


ACM Trans. Internet Things


I am an Earphone and I can Hear my Users Face: Facial Landmark Tracking using Smart Earphones • 27


[99] Xingze Song, Kai Huang, and Wei Gao. [n. d.]. FaceListener: Recognizing Human Facial Expressions via Acoustic Sensing on Smart Headphones. [n. d.].


