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 comforts 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* (**E**arphone **A**coustics based **R**econstruction of 3D **F**ace), that shows the feasibility of reconstructing 3D facial motion using sensor embedded smart earphones, which are gaining in

*Co-primary authors

Authors' addresses: Shijia Zhang, scarlettzhang27@psu.edu, Penn State University, University Park, PA, USA; Taiting Lu, txl5518@psu.edu, Penn State University, University Park, PA, USA; Hao Zhou, hfz5190@psu.edu, Penn State University, University Park, PA, USA; Yilin Liu, yzl470@psu.edu, Penn State University, University Park, PA, USA; Runze Liu, rml6043@psu.edu, Penn State University, University Park, PA, USA; Mahanth Gowda, mahanth.gowda@psu.edu, Penn State University, University Park, PA, USA.

2577-6207/2023/8-ART $15.00
https://doi.org/10.1145/3614438

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

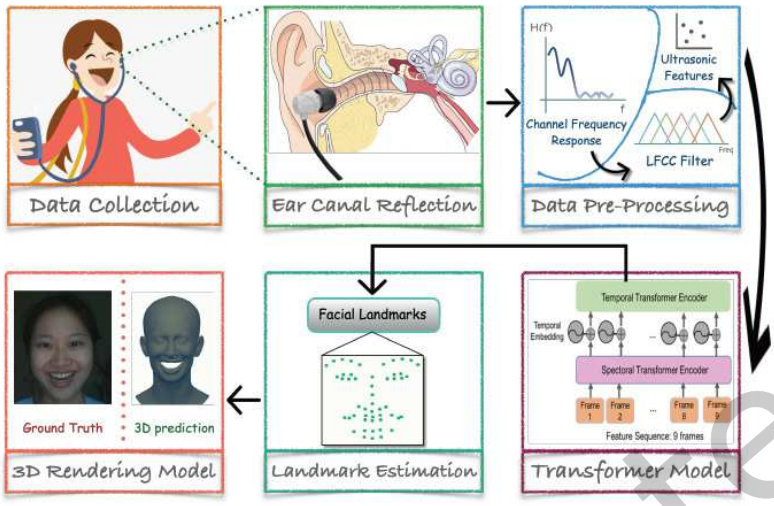

Fig. 1. Overview of *EARFace*: The ultrasonic reflections from ear canals is used for extracting the CFR. Features are extracted from the CFR and processed by a 2D transformer based ML model that captures spectral and temporal dependencies in the ear canal CFR to predict facial landmarks, which are ultimately used for reconstructing the 3D face model of the user.

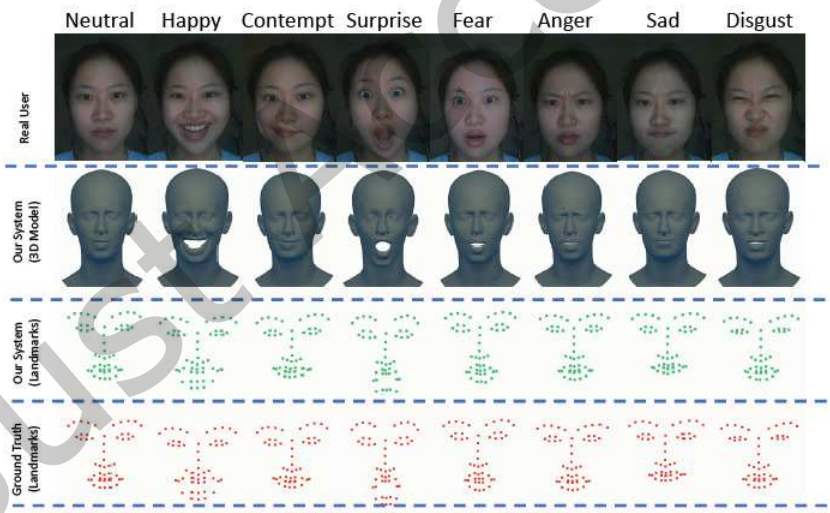

Fig. 2. 3D facial reconstruction and landmark estimation in *EARFace*.

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 upto 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 upto 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.

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

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 ultrasnoic 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, EarIO 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, noise 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

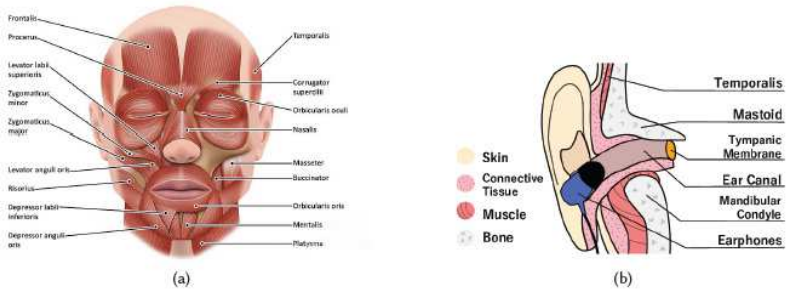

Fig. 3. (a) Facial muscle anatomy[3] (b) Ear canal anatomy[20]

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 ($\vec{\beta} \in \mathcal{S} = \mathbb{R}^{|\vec{\beta}|}$): Defines the deformations in the face due to the unique identity of the subject. (ii) Pose Parameters ($\vec{\theta} \in \mathcal{P} = \mathbb{R}^{|\vec{\theta}|}$): Defines the deformation in the face because of the rotation of the head around the neck or motion of the jaw. (iii) Expression Parameters ($\vec{\psi} \in \mathcal{E} = \mathbb{R}^{|\vec{\psi}|}$): 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.

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

Each of the three sets of parameters above (shape $\vec{\beta}$, pose $\vec{\theta}$, and expression $\vec{\phi}$) are represented in a highly compressed space, the bases (i.e., $\mathcal{S}$, $\mathcal{P}$, and $\mathcal{E}$) 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 $\mathbf{T} \in \mathbb{R}^{3N}$. To take the shape variation of different identities into consideration, a linear shape blendshape function $B_S(\vec{\beta}; \mathcal{S}) = \sum_{n=1}^{|\vec{\beta}|} \beta_n \mathbf{S}_n$ adds the offsets to the template $\mathbf{T}$, where $\vec{\beta} = [\beta_1, \beta_2, ..., \beta_{|\vec{\beta}|}]^T$ denotes

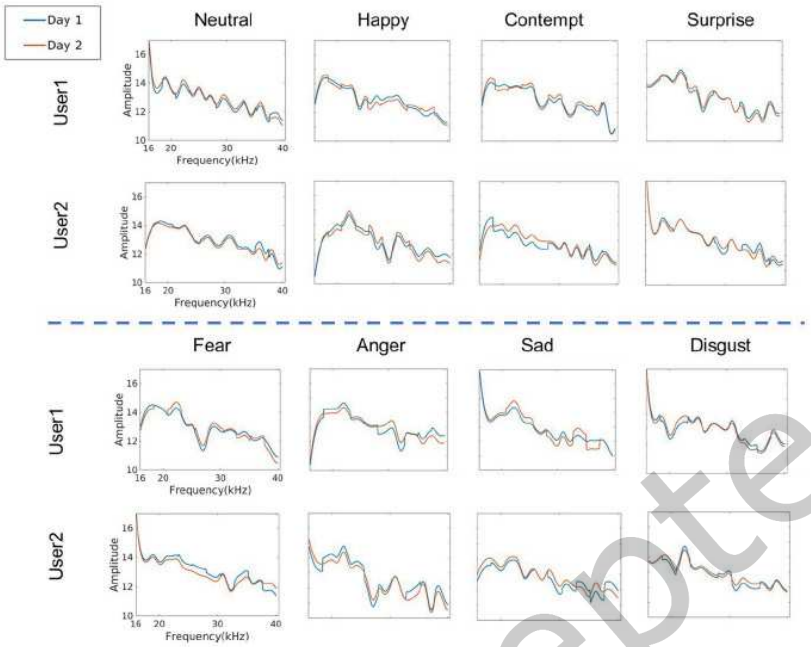

Fig. 4. Feasibility of Facial Landmark Tracking Using CFR. CFR is consistent acorss multiple days for same facial expressions of a user and it varies with change in facial expressions. Although there are some similarities across different users for same facial expressions, there are also differences due to individual differences in ear canal shape which can be handled by domain adaptation. Detailed facial reconstruction results are presented in Section 6.

the coefficients for shape variations and $\mathcal{S} = [\mathrm{S}_1, \mathrm{S}_2, ..., \mathrm{S}_{|\vec{\beta}|}]$ denotes the orthogonal shape basis. Similarly, the offsets calculated from a pose blendshape function $B_P(\vec{\theta}; \mathcal{P})$ and an expression blendshape function $B_E(\vec{\beta}; \mathcal{E})$ 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 $(\vec{\beta}, \vec{\theta}, \vec{\phi})$ 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,

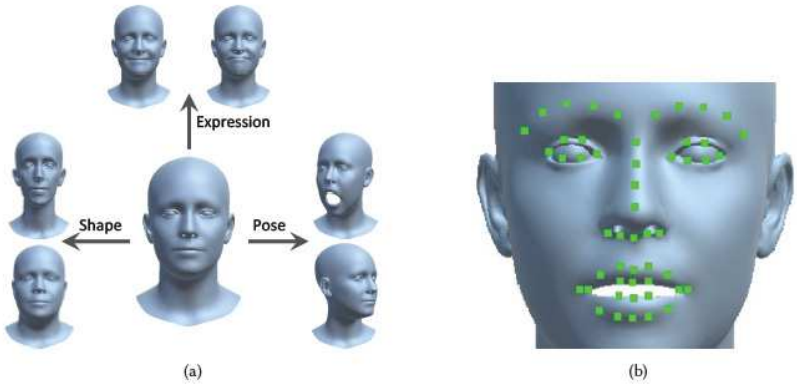

Fig. 5. (a) Parameterization in FLAME [70]. Starting from a neutral face with no expression, deformations of shape, pose and expression can be captured by the corresponding parameters (b) Key Facial Landmarks used in *EARFace* for 3D facial reconstruction

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

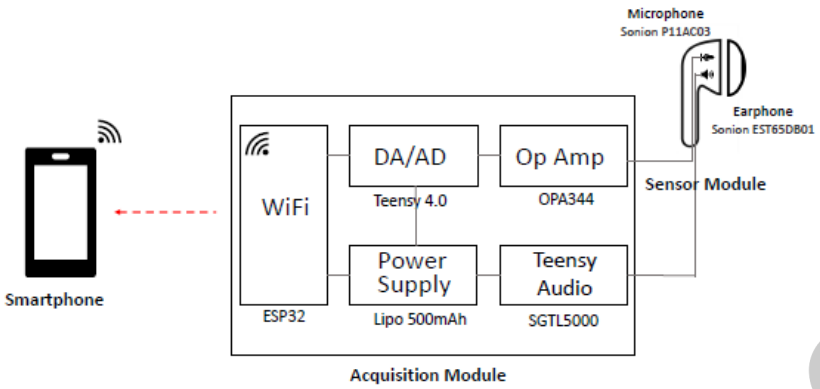

Fig. 7. Design of the earphone platform in *EARFace*

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.

$$H(k) = \frac{Y(k)}{X(k)} \; \forall \; k \in \{1, 2, ..4000\}, \tag{1}$$

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 100×2). 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

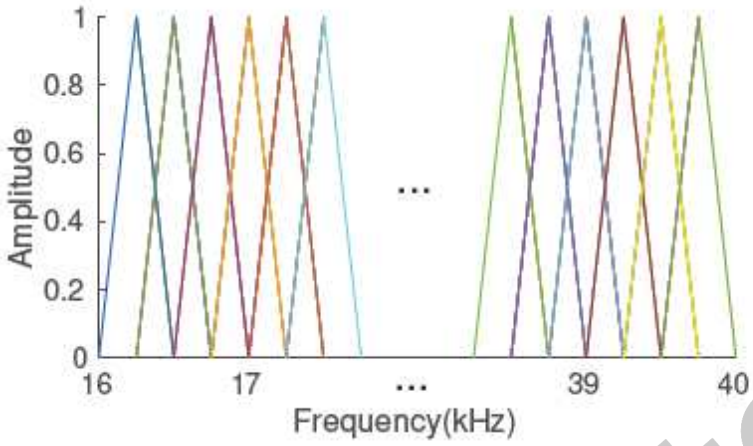

Fig. 8. LFCC filters for extracting features from the CFR

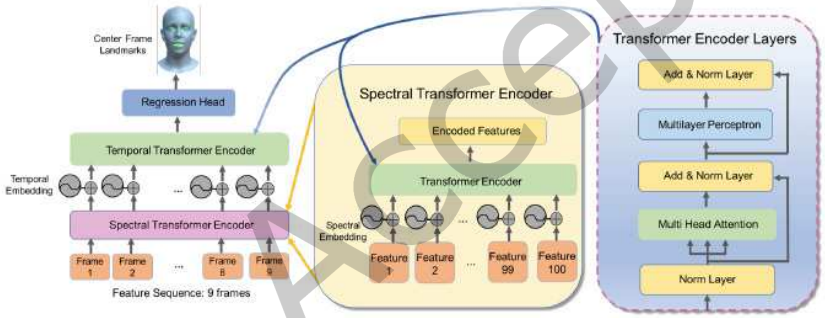

Fig. 9. ML model combines spectral and temporal transformer for predicting facial landmarks from earphone acoustic sensor data

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 100×2 are passed through a linear projection layer that outputs a tensor that is of size 100×1. 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 ($\sum_{j=1}^{j=N}(q_i \cdot k_j)v_j$). 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 transfromers 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 *pretrain*ed 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 *pretrain*ed 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[1]. (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

[1]We 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. $MAE = \frac{1}{N} \times \sum \|v - e\| \times \frac{d_{or}}{d_{on}}$ 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 compute 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}. $NME = \frac{1}{N} \times \sum \frac{\|v-e\|}{d_{on}} \times 100\%$ 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

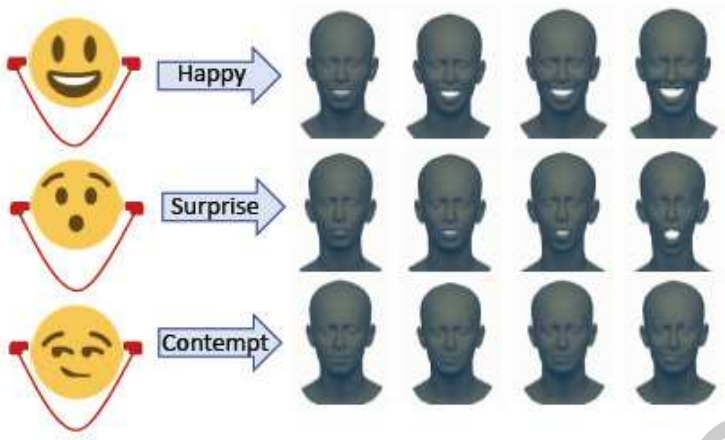

Fig. 10. Progression of facial feature changes as predicted by *EARFace*

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.

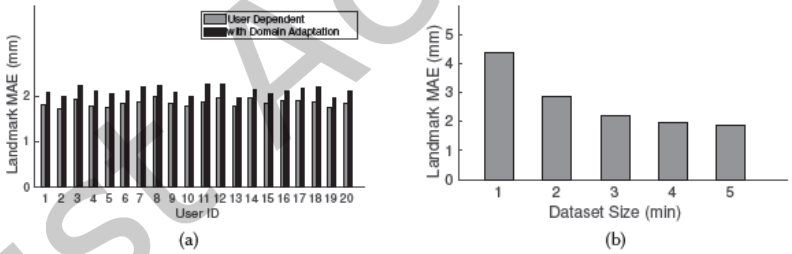

Fig. 11. (a) Accuracy vs Users. *EARFace* achieves consistent accuracy across all users (b) Accuracy as a function of size of training data. *EARFace*'s ML models converge with relatively low overhead in training

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

performance.

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.

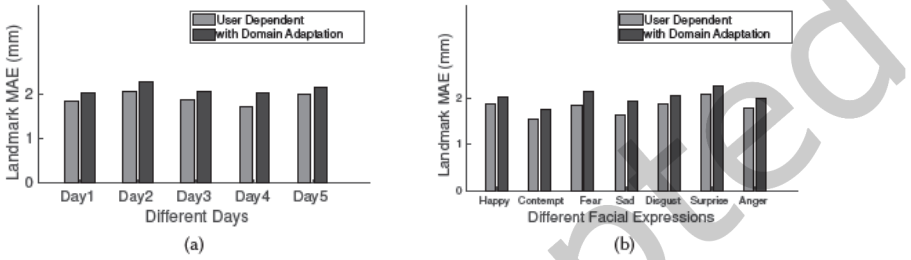

Fig. 12. (a) Accuracy is consistent across different days of experimenting, with robustness to natural variation in sensor position and orientation (b) *EARFace* can track commonly occurring facial expressions

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 2mm 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 ($\approx$ 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

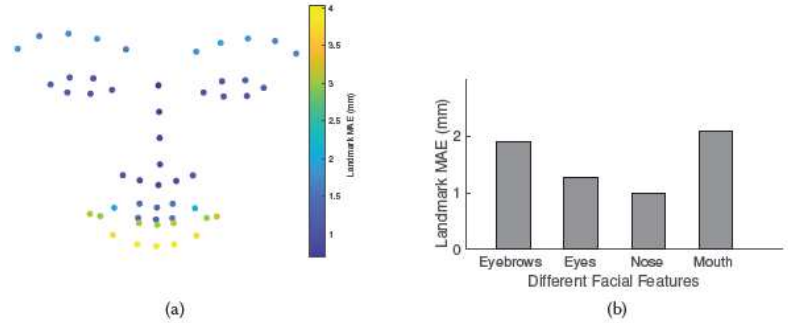

Fig. 13. Accuracy variation as a function of Facial Landmarks (a) Variation across individual landmarks (b) Region specific variation

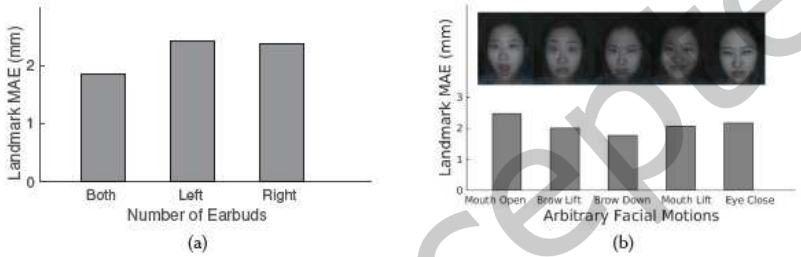

Fig. 14. (a) Accuracy vs Number of Earphones. Both earphones together achieve the best accuracy whereas individual earphones also provide reasonable levels of tracking (b) Tracking in *EARFace* generalizes to arbitrary facial motions

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.

Table 2. Ablation Study

Method	MAE (mm)
Encoder Decoder (CNN)	2.25
Transformer (Temporal)	1.93
Transformer (Temporal + Spectral)	**1.83**

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

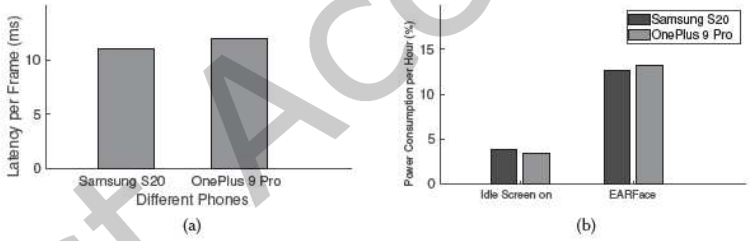

Fig. 15. (a) Latency of ML model execution on smartphone devices (b) Power Consumption on Smartphone Devices

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

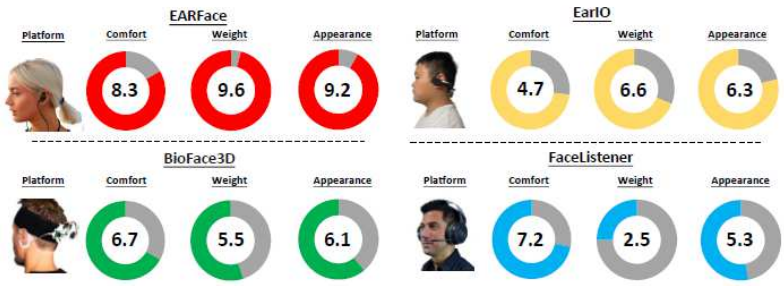

Fig. 16. User experience survey on *EARFace* compared with three alternative sensing platforms

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 wireless 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

[1] [n. d.]. Airpods - Apple. "https://www.apple.com/airpods/".
[2] [n. d.]. ESP32. "https://blog.taotronics.com/headphones/tws-headphones/".
[3] [n. d.]. Facial muscle anatomy. "https://fineartamerica.com/featured/face-muscle-anatomy-maurizio-de-angelisscience-photo-library.html".
[4] [n. d.]. Muscles of Facial Expression. "https://geekymedics.com/muscles-of-facial-expression/".
[5] [n. d.]. Profile battery usage with Batterystats and Battery Historian. https://developer.android.com/topic/performance/power/setup-battery-historian.
[6] [n. d.]. Sine Sweep Test. "https://vru.vibrationresearch.com/lesson/sine-sweep-test/".
[7] [n. d.]. The Seven Universal Emotions We Wear On Our Face. "https://www.cbc.ca/natureofthings/features/the-seven-universal-emotions-we-wear-on-our-face#".
[8] [n. d.]. True Wireless Setero. "https://www.espressif.com/en/products/socs/esp32".
[9] 1998. Criteria for a recommended standard: occupational noise exposure. https://www.cdc.gov/niosh/docs/98-126/default.html.
[10] 2022. EST65DB01. https://www.sonion.com/product/est65da01/.
[11] 2022. P11AC03 datasheet. https://www.sonion.com/wp-content/uploads/ds-P11AC03_v3.pdf.
[12] 2022. scipy.signal.chirp. https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.chirp.html.
[13] 2022. Single-supply, rail-to-rail operational microamplifier series datasheet. https://www.ti.com/lit/ds/symlink/opa344.pdf.
[14] 2022. Teensy Audio Board. https://www.pjrc.com/store/teensy3_audio.html.
[15] 2022. Teensyduino. https://www.pjrc.com/teensy/teensyduino.html.
[16] 2022. Teensy&reg 4.1. https://www.pjrc.com/store/teensy41.html.
[17] 2022. WiFiNINA Library for Arduino. https://github.com/adafruit/WiFiNINA.
[18] Martín Abadi et al. 2016. Tensorflow: A system for large-scale machine learning. In *OSDI*. 265–283.
[19] Shideh Kabiri Ameri, Myungsoo Kim, Irene Agnes Kuang, Withanage K Perera, Mohammed Alshiekh, Hyoyoung Jeong, Ufuk Topcu, Deji Akinwande, and Nanshu Lu. 2018. Imperceptible electrooculography graphene sensor system for human–robot interface. *npj 2D Materials and Applications* 2, 1 (2018), 1–7.

[20] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. 2019. Facial expression recognition using ear canal transfer function. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 1–9.
[21] Sungtae An, Alessio Medda, Michael N Sawka, Clayton J Hutto, Mindy L Millard-Stafford, Scott Appling, Kristine LS Richardson, and Omer T Inan. 2021. AdaptNet: human activity recognition via bilateral domain adaptation using semi-supervised deep translation networks. *IEEE Sensors Journal* 21, 18 (2021), 20398–20411.
[22] 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 30th Annual ACM Symposium on User Interface Software and Technology*. 679–689.
[23] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2481–2495.
[24] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.
[25] Matteo Bodini. 2019. A review of facial landmark extraction in 2D images and videos using deep learning. *Big Data and Cognitive Computing* 3, 1 (2019), 14.
[26] Fady S Botros, Angkoon Phinyomark, and Erik J Scheme. 2022. Day-to-Day Stability of Wrist EMG for Wearable-Based Hand Gesture Recognition. *IEEE Access* 10 (2022), 125942–125954.
[27] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 2019. ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. In *The 25th annual international conference on mobile computing and networking*. 1–17.
[28] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2013), 413–425.
[29] Gaoshuai Cao, Kuang Yuan, Jie Xiong, Panlong Yang, Yubo Yan, Hao Zhou, and Xiang-Yang Li. 2020. Earphonetrack: involving earphones into the ecosystem of acoustic motion tracking. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 95–108.
[30] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.
[31] William Chan et al. [n. d.]. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *IEEE ICASSP 2016*. 4960–4964.
[32] Chien-Hsu Chen, I-Jui Lee, and Ling-Yi Lin. 2015. Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders. *Research in developmental disabilities* 36 (2015), 396–403.
[33] Chien-Hsu Chen, I-Jui Lee, and Ling-Yi Lin. 2016. Augmented reality-based video-modeling storybook of nonverbal facial cues for children with autism spectrum disorder to improve their perceptions and judgments of facial expressions and emotions. *Computers in Human Behavior* 55 (2016), 477–485.
[34] 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 33rd annual ACM symposium on user interface software and technology*. 112–125.
[35] Seokmin Choi, Yang Gao, Yincheng Jin, Se jun Kim, Jiyang Li, Wenyao Xu, and Zhanpeng Jin. 2022. PPGface: Like What You Are Watching? Earphones Can" Feel" Your Facial Expressions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–32.
[36] Romit Roy Choudhury. 2021. Earable computing: A new area to think about. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*. 147–153.
[37] Demo. 2022. https://streamable.com/t34w8l.
[38] Jia Deng et al. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*.
[39] Jacob Devlin et al. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[40] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. 2018. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 379–388.
[41] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
[42] Yang Gao, Yincheng Jin, Seokmin Choi, Jiyang Li, Junjie Pan, Lin Shu, Chi Zhou, and Zhanpeng Jin. 2021. SonicFace: Tracking Facial Expressions Using a Commodity Microphone Array. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–33.
[43] Google. 2019. Deploy machine learning models on mobile and IoT devices. "https://www.tensorflow.org/lite".

[44] Valentin Goverdovsky, Wilhelm Von Rosenberg, Takashi Nakamura, David Looney, David J Sharp, Christos Papavassiliou, Mary J Morrell, and Danilo P Mandic. 2017. Hearables: Multimodal physiological in-ear sensing. *Scientific reports* 7, 1 (2017), 1–10.
[45] Xiaobing Han et al. 2017. Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. *Remote Sensing* (2017).
[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[47] Sen He, Wentong Liao, Hamed R Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. 2020. Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision*.
[48] Marko Hiipakka, Miikka Tikander, and Matti Karjalainen. 2010. Modeling the External Ear Acoustics for Insert Headphone Usage. *Journal of The Audio Engineering Society* 58 (2010), 269–281.
[49] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).
[50] Rong Hu, Ling Chen, Shenghuan Miao, and Xing Tang. 2023. Swl-adapt: An unsupervised domain adaptation model with sample weight learning for cross-user wearable human activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 6012–6020.
[51] Shun Ishii, Anna Yokokubo, Mika Luimula, and Guillaume Lopez. 2020. Exersense: physical exercise recognition and counting algorithm from wearables robust to positioning. *Sensors* 21, 1 (2020), 91.
[52] Md Shafiqul Islam, Tahera Hossain, Md Atiqur Rahman Ahad, and Sozo Inoue. 2021. Exploring human activities using eSense earable device. In *Activity and Behavior Computing*. Springer, 169–185.
[53] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407* (2018).
[54] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: " Hearing" Your Silent Speech Commands In Ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.
[55] Stamos Katsigiannis and Naeem Ramzan. 2017. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics* 22, 1 (2017), 98–107.
[56] Fahim Kawsar, Chulhong Min, Akhil Mathur, Alessandro Montanari, Utku Günay Acer, and Marc Van den Broeck. 2018. esense: Open earable platform for human sensing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 371–372.
[57] Daniel P Kennedy and Ralph Adolphs. 2012. Perception of emotions from facial expressions in high-functioning adults with autism. *Neuropsychologia* 50, 14 (2012), 3313–3319.
[58] Md Abdullah Al Hafiz Khan, Nirmalya Roy, and Archan Misra. 2018. Scaling human activity recognition via deep learning-based domain adaptation. In *2018 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE, 1–9.
[59] Prerna Khanna, Tanmay Srivastava, Shijia Pan, Shubham Jain, and Phuc Nguyen. 2021. Jawsense: recognizing unvoiced sound using a low-cost ear-worn system. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*. 44–49.
[60] Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. 2020. T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6649–6653.
[61] Diederik P Kingma et al. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[62] Tomi Kinnunen et al. [n. d.]. Voice activity detection using MFCC features and support vector machine. In *SPECOM 2007*.
[63] Alex Krizhevsky et al. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.
[64] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.
[65] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. 2020. AvatarMe: Realistically Renderable 3D Facial Reconstruction" in-the-wild". In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 760–769.
[66] Seulki Lee, Bashima Islam, Yubo Luo, and Shahriar Nirjon. 2019. Intermittent learning: On-device machine learning on intermittently powered system. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–30.
[67] Won-Sook Lee, Prem Kalra, and Nadia Magnenat-Thalmann. 1997. Model based face reconstruction for animation. *Proc. Multimedia Modeling (MMM)* 97 (1997), 323–338.
[68] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8928–8937.
[69] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–24.

[70] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.
[71] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, and Shilin Zhou. 2022. Light field image super-resolution with transformers. *IEEE Signal Processing Letters* 29 (2022), 563–567.
[72] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. NeuroPose: 3D hand pose tracking using EMG wearables. In *Proceedings of the Web Conference 2021*. 1471–1482.
[73] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. When video meets inertial sensors: Zero-shot domain adaptation for finger motion analytics with inertial sensors. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 182–194.
[74] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2022. A Practical System for 3-D Hand Pose Tracking Using EMG Wearables With Applications to Prosthetics and User Interfaces. *IEEE Internet of Things Journal* 10, 4 (2022), 3407–3427.
[75] Roya Lotfi, George Tzanetakis, Rasit Eskicioglu, and Pourang Irani. 2020. A comparison between audio and IMU data to detect chewing events based on an earable device. In *Proceedings of the 11th Augmented Human International Conference*. 1–8.
[76] Yubo Luo and Yongfeng Huang. 2017. Text steganography with high embedding rate: Using recurrent neural networks to generate chinese classic poetry. In *Proceedings of the 5th ACM workshop on information hiding and multimedia security*. 99–104.
[77] Yubo Luo, Yongfeng Huang, Fufang Li, and Chinchen Chang. 2016. Text steganography based on ci-poetry generation using Markov chain model. *KSII Transactions on Internet and Information Systems (TIIS)* 10, 9 (2016), 4568–4584.
[78] Yubo Luo, Le Zhang, Zhenyu Wang, and Shahriar Nirjon. 2023. Efficient multitask learning on resource-constrained systems. *arXiv preprint arXiv:2302.13155* (2023).
[79] Balz Maag, Zimu Zhou, Olga Saukh, and Lothar Thiele. 2017. BARTON: Low power tongue movement sensing with in-ear barometers. In *2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 9–16.
[80] Katsutoshi Masai, Kai Kunze, Daisuke Sakamoto, Yuta Sugiura, and Maki Sugimoto. 2020. Face Commands-User-Defined Facial Gestures for Smart Glasses. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 374–386.
[81] Katsutoshi Masai, Yuta Sugiura, Masa Ogata, Kai Kunze, Masahiko Inami, and Maki Sugimoto. 2016. Facial expression recognition in daily life by embedded photo reflective sensors on smart eyewear. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 317–326.
[82] Denys JC Matthies, Bernhard A Strecker, and Bodo Urban. 2017. Earfieldsensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1911–1922.
[83] Denys JC Matthies, Bernhard A Strecker, and Bodo Urban. 2017. Earfieldsensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1911–1922.
[84] Denys JC Matthies, Chamod Weerasinghe, Bodo Urban, and Suranga Nanayakkara. 2021. CapGlasses: Untethered Capacitive Sensing with Smart Glasses. In *Augmented Humans Conference 2021*. 121–130.
[85] Abdelrahman Mohamed, Dmytro Okhonko, and Luke Zettlemoyer. 2019. Transformers with convolutional context for asr. *arXiv preprint arXiv:1904.11660* (2019).
[86] Aiko Murata, Hisamichi Saito, Joanna Schug, Kenji Ogawa, and Tatsuya Kameda. 2016. Spontaneous facial mimicry is enhanced by the goal of inferring emotional states: evidence for moderation of "automatic" mimicry by higher cognitive processes. *PloS one* 11, 4 (2016), e0153128.
[87] Wajahat Nawaz et al. 2018. Classification of breast cancer histology images using alexnet. In *International conference image analysis and recognition*. Springer.
[88] United States. Office of Noise Abatement. 1974. *Information on levels of environmental noise requisite to protect public health and welfare with an adequate margin of safety*. Number 2115. US Government Printing Office.
[89] OnePlus 9 Pro [n. d.]. OnePlus 9 Pro. "https://www.oneplus.com/us/9-pro".
[90] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10975–10985.
[91] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 296–301.
[92] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.
[93] JunRong Qi. 2016. *Cross-correlation between mandibular condylar movements and distortion of external auditory meatus*. Ph. D. Dissertation. https://ci.nii.ac.jp/naid/500000981228
[94] Chen Qu et al. 2019. BERT with history answer embedding for conversational question answering. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.
[95] Yassine Rabhi, Makrem Mrabet, and Farhat Fnaiech. 2018. A facial expression controlled wheelchair for people with disabilities. *Computer methods and programs in biomedicine* 165 (2018), 89–105.
[96] Samsung Galaxy Note 20 [n. d.]. Samsung Galaxy Note 20. "https://www.samsung.com/africa_en/smartphones/galaxy-note20/models/".

[97] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. 2019. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7763–7772.
[98] Fangyao Shen, Yong Peng, Wanzeng Kong, and Guojun Dai. 2021. Multi-scale frequency bands ensemble learning for EEG-based emotion recognition. *Sensors* 21, 4 (2021), 1262.
[99] Xingzhe Song, Kai Huang, and Wei Gao. [n. d.]. FaceListener: Recognizing Human Facial Expressions via Acoustic Sensing on Commodity Headphones. ([n. d.]).
[100] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–26.
[101] Paul Stoffregen. 2022. Teensy Audio implementation library. https://github.com/PaulStoffregen/Audio.
[102] Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. 2020. SEANet: A multi-modal speech enhancement network. *arXiv preprint arXiv:2009.02095* (2020).
[103] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.
[104] ARAKAWA Takayuki. 2019. Ear acoustic authentication technology: Using sound to identify the distinctive shape of the ear canal. *NEC Tech. J.-Special Issue Social Value Creation Using Biometrics* 13, 2 (2019), 87–90.
[105] Hadish Habte Tesfamikael, Adam Fray, Israel Mengsteab, Adonay Semere, and Zebib Amanuel. 2021. Simulation of Eye Tracking Control based Electric Wheelchair Construction by Image Segmentation Algorithm. *Journal of Innovative Image Processing (JIIP)* 3, 01 (2021), 21–35.
[106] Vishwas G Torvi, Aditya Bhattacharya, and Shayok Chakraborty. 2018. Deep domain adaptation to predict freezing of gait in patients with Parkinson's disease. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 1001–1006.
[107] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[108] Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. 2021. ExpressEar: Sensing Fine-Grained Facial Expressions with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–28.
[109] Pedro Villanueva. 2022. Teensy ADC implementation library. https://github.com/pedvide/ADC.
[110] Stefan Wager et al. 2013. Dropout training as adaptive regularization. In *Advances in neural information processing systems*.
[111] Susan W White, Carla A Mazefsky, Gabriel S Dichter, Pearl H Chiu, John A Richey, and Thomas H Ollendick. 2014. Social-cognitive, physiological, and neural mechanisms underlying emotion regulation impairments: Understanding anxiety in autism spectrum disorder. *International Journal of Developmental Neuroscience* 39 (2014), 22–36.
[112] 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 27th Annual International Conference on Mobile Computing and Networking*. 350–363.
[113] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 532–539.
[114] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. 2020. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5791–5800.
[115] Zhijian Yang, Yu-Lin Wei, Sheng Shen, and Romit Roy Choudhury. 2020. Ear-ar: indoor acoustic augmented reality on earphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.
[116] Weiwei Yu, Jian Zhou, HuaBin Wang, and Liang Tao. 2022. SETransformer: speech enhancement transformer. *Cognitive Computation* 14, 3 (2022), 1152–1158.
[117] Haoyu Zhang, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243* (2019).
[118] Ligang Zhang and Dian Tjondronegoro. 2011. Facial Expression Recognition Using Facial Movement Features. *IEEE Transactions on Affective Computing* 2, 4 (2011), 219–229. https://doi.org/10.1109/T-AFFC.2011.13
[119] Shijia Zhang, Yilin Liu, and Mahanth Gowda. 2022. Let's Grab a Drink: Teacher-Student Learning for Fluid Intake Monitoring using Smart Earphones. In *2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 55–66.
[120] Shijia Zhang, Yilin Liu, and Mahanth Gowda. 2023. I Spy You: Eavesdropping Continuous Speech on Smartphones via Motion Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–31.
[121] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2020. Revisiting few-sample BERT fine-tuning. *arXiv preprint arXiv:2006.05987* (2020).
[122] Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson, and Shihab Shamma. 2011. Linear versus mel frequency cepstral coefficients for speaker recognition. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 559–564.

[123] Zongwei Zhou et al. 2017. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *IEEE CVPR*.

[124] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. 2015. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4998–5006.