ExpressEar: Sensing Fine-Grained Facial Expressions with Earables

DHRUV VERMA* and SEJAL BHALLA*, Indraprastha Institute of Information Technology, Delhi, India
DHRUV SAHNAN, Indraprastha Institute of Information Technology, Delhi, India
JAINENDRA SHUKLA, Indraprastha Institute of Information Technology, Delhi, India
AMAN PARNAMI, Indraprastha Institute of Information Technology, Delhi, India

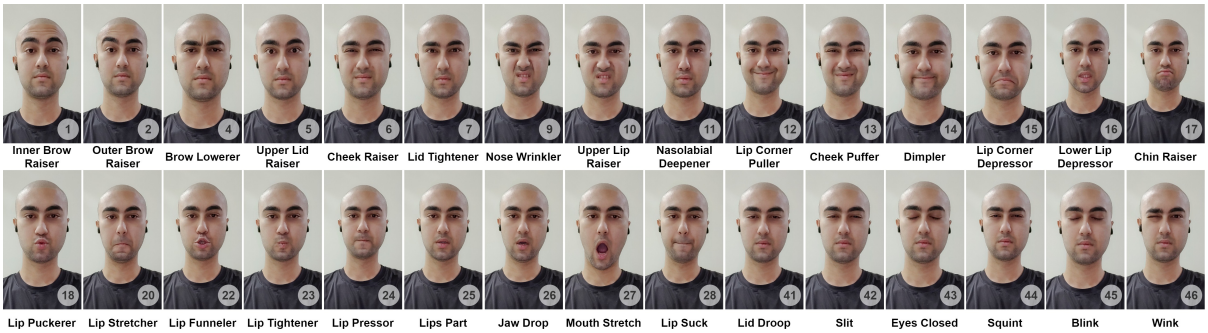


Fig. 1. 30 Facial Action Units (AUs) proposed by Facial Action Coding System (FACS) [29].
For asymmetrical AUs (2: *Outer Brow Raiser* and 46: *Wink*), the figure shows a single variant (left or right) only.

Continuous and unobtrusive monitoring of facial expressions holds tremendous potential to enable compelling applications in a multitude of domains ranging from healthcare and education to interactive systems. Traditional, vision-based facial expression recognition (FER) methods, however, are vulnerable to external factors like occlusion and lighting, while also raising privacy concerns coupled with the impractical requirement of positioning the camera in front of the user at all times. To bridge this gap, we propose *ExpressEar*, a novel FER system that repurposes commercial earables augmented with inertial sensors to capture fine-grained facial muscle movements. Following the Facial Action Coding System (FACS), which encodes every possible expression in terms of constituent facial movements called Action Units (AUs), ExpressEar identifies facial expressions at the atomic level. We conducted a user study (N=12) to evaluate the performance of our approach and found that ExpressEar can detect and distinguish between 32 Facial AUs (including 2 variants of asymmetric AUs), with an average accuracy of 89.9% for any given user. We further quantify the performance across different mobile scenarios in presence of additional face-related activities. Our results demonstrate ExpressEar's applicability in the real world and open up research opportunities to advance its practical adoption.

*Both authors contributed equally to this research.

Authors' addresses: Dhruv Verma, dhruv17046@iiitd.ac.in; Sejal Bhalla, sejal17100@iiitd.ac.in, Indraprastha Institute of Information Technology, Delhi, India; Dhruv Sahnan, Indraprastha Institute of Information Technology, Delhi, India, dhruv18230@iiitd.ac.in; Jainendra Shukla, Indraprastha Institute of Information Technology, Delhi, India, jainendra@iiitd.ac.in; Aman Parnami, Indraprastha Institute of Information Technology, Delhi, India, aman@iiitd.ac.in.

2474-9567/2021/9-ART129 $15.00
https://doi.org/10.1145/3478085

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile devices**; *Gestural input*.

Additional Key Words and Phrases: facial expressions, FACS, earable computing, IMU sensing

ACM Reference Format:
Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. 2021. ExpressEar: Sensing Fine-Grained Facial Expressions with Earables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 129 (September 2021), 28 pages. https://doi.org/10.1145/3478085

1 INTRODUCTION

Facial expressions have long been viewed as indicators of our mental state and emotions [26]. Their ability to communicate important non-verbal cues about human intent and affect has opened possibilities for a wide range of applications in affective computing, healthcare, education, entertainment and more. For example, continuously monitoring facial movements in a classroom environment can give instructors useful feedback about the engagement level of students [82]. In the context of healthcare, FER allows medical professionals to timely understand the mental state of children with autism and depression [5, 46]. Beyond the boundaries of real world, facial expressions have also proven to be useful for virtual (AR/VR) systems that need better ways of sensing user attention, intent, and context in order to create a more natural and immersive experience for the user [38]. Moreover, FER systems can enable users for multitasking by providing situational interactive controls without obstructing their normal course of action. For instance, in critical scenarios like driving, facial expressions can be used in the form of microinteractions "*... because they may minimize interruption; that is, they allow for a tiny burst of interaction with a device so that the user can quickly return to the task at hand.*" – Ashbrook [13].

Traditional FER methods based on vision-enabled recognition systems require a camera directed towards the user's face at all times. Accompanied with privacy concerns, these methods have limited tolerance to occlusion, position change, camera angle and lighting conditions. Thus, despite being highly accurate, vision-based approaches limit the continuous and unobtrusive detection of facial expressions. To address these challenges, alternate sensing techniques like wearable Electroencephalography (EEG) [1, 3], Electrodermal Activity (EDA) [4, 34], and respiration sensors [2] have been leveraged. However, most of these modalities either fail to capture fine-grained facial movements or can't be realised in the form of a practical system due to their uncomfortable form factor. This posits a clear need for a practical wearable sensing technology, preferably in the form-factor of a familiar ubiquitous device, that can continuously and unobtrusively monitor fine-grained facial expressions while preserving the privacy of the user.

Inertial Measurement Unit (IMU)-augmented Earable sensing presents itself as an intriguing alternative due to multiple reasons. First of all, ears represent an extremely good vantage point to sense facial activity owing to their anatomical connectivity with the facial muscles. Secondly, earables are discreet, privacy preserving, and already integrated into our daily lives. In addition to this, the growing adoption of IMUs in modern wireless earbuds opens up a new avenue for ubiquitous sensing of facial expressions without the need of any hardware modifications. To this end, we introduce *ExpressEar*, a novel FER system which exploits off-the-shelf IMU-augmented earables for sensing fine-grained facial muscle movements. The design of ExpressEar is based on the key observation that the characteristics of ear-mounted IMU signals are a rich source of information about the distinct movements associated with facial expressions. ExpressEar's ability to categorise a broad range of facial expressions allows us to envision a variety of applications which may involve the *voluntary* use of facial expressions for gestural interactions and hands-free control, or monitoring of *involuntary* facial expressions to infer the emotional state of a user. Furthermore, since ExpressEar relies on the capabilities of devices that are commercially available, it can easily be deployed on similar earable platforms merely through a software update, enabling new interactive experiences for users.

Instead of designing our own set of facial expressions, we turn our attention to the Facial Action Coding System (FACS) [29], specifically its constituent facial Action Units (AUs), for implementing ExpressEar. FACS defines a set of 30 AUs (see Figure 1) which descriptively encodes various facial expressions at the atomic level. For the purpose of this work, we consider both the left and right variants of AU2 and AU46, leading to a total of 32 distinct facial AUs. We conducted a user study (N=12) to evaluate the performance of our approach in detecting and distinguishing between the facial AUs as laid down by FACS. The collected data was used to train a Temporal Convolutional Network (TCN) to classify all the AUs. In a subject dependent setting, our best performing classifier achieved an average classification accuracy of 89.9% (Max: 94.5%, SD: 3.5). We also investigated the effect of plurality of sensors and found that ExpressEar was able to classify all the AUs with an average accuracy of 79.6% (Max: 87.0%, SD: 7.1) when trained on the data collected from the left IMU (using only left earbud) and 79.3% (Max: 88.7%, SD: 5.7) for right IMU (using only right earbud). This suggests that if required, we can rely on just one ear-mounted IMU to identify facial expressions with a reasonable confidence.

To inspect the effect on performance in free-living conditions, we evaluated the system under challenging conditions of noise and user mobility. While ExpressEar was able to differentiate facial AUs from other face-related movements with an accuracy of 99.2%, we witnessed a slight drop in performance along with a less extensive coverage of facial AUs for classification in mobile settings. We achieved an average per-user accuracy of 83.85% (SD: 1.25, Max: 85.61%, Chance: 7.143%) across 14 AUs performed in a moving rapid-transit train and 84.34% (SD: 0.63, Max: 84.89%, Chance: 25%) across 4 AUs performed while walking. Overall, the experimental results demonstrate the efficacy of our approach across various testing conditions.

To sum up, the main contributions of this work are:

- We demonstrate the ability of ear-mounted IMU sensors to detect atomic facial expressions, also known as facial AUs.
- We present *ExpressEar*, a practical and unobtrusive FER system supported by commodity wireless earbuds without the need of any hardware modification.
- We conducted a user study with 12 participants to evaluate the performance of our system across 32 distinct facial AUs, achieving an average per-user accuracy of 89.9%. Our experiments provide insights on subject variability, plurality of sensors used and suitability of specific facial AUs in different application domains.
- We assess the practical applicability of ExpressEar by conducting additional experiments with three participants, in the presence of other face-related movements (eating, speaking) and mobile settings (walking, rapid-transit systems).
- We envision a wide range of applications in context of voluntary and involuntary use of facial expressions, and further discuss the challenges, opportunities and limitations of our approach for future work.

2 RELATED WORK

In this paper, we evaluate the performance of ExpressEar in continuous FER using IMU-augmented earables. Therefore, we discuss the related work in three sections: 1) different approaches for FER, 2) advancements in the field of earable computing, and 3) state-of-the-art learning techniques for IMU-based activity recognition.

2.1 Face-Engaged Activity Recognition

In the vast amount of literature on facial action recognition, camera-based approaches using computer vision stand out as the most frequent contributor to the field [32, 70]. While this is a natural fit when the user is facing a computing system that can optimally capture the entire view of his/her face, it suffers from a range of issues as described in Section 1. A few related efforts have explored the possibilities of placing cameras in custom eyeglasses [49], VR headsets [38] and earphones [23]. However, the confined field of view, high susceptibility

to occlusion, ambient light effects and the inherent intrusiveness limit their application for facial sensing. This led to the increasing exploration of other sensing modalities such as proximity sensors, EMG, pressure sensors, EOG and audio. Eyeglass-embedded proximity sensors [52] can capture skin deformation by tracking the skin to sensor distance but in addition to being highly sensitive to the ambient lighting, they too need to point at the face and have a constrained field of view at close proximity. While EMG [36] and pressure sensor based methods [71] overcome these limitations, they require the sensor to be attached directly onto the users' faces or body, which is intrusive and may interfere with daily activities. Interferi [42], an intriguing on-body gesture sensing technique using acoustic interferometry, also poses similar challenges with a bulky form factor. EOG augmented eyeglasses [68] address most of the aforementioned concerns but are incapable of recognizing fine-grained facial expressions (the cited work could only recognise five distinct facial expressions). Moreover, most of these sensors still haven't found their place beyond research prototypes. In contrast, our approach focuses on IMU-augmented earables due to its highly promising sensor location, a form factor that is blended in our daily lives, commercial availability and insensitivity to a multitude of factors that adversely affect FER.

2.2 Earable Computing

Earables are in-ear wearables packed with sensors ranging from proximity sensors to heart rate monitors. These smart earpieces enable a plethora of compelling applications. They have been used to track physiological state of the user by calculating respiration rate [69]. Utilizing just microphones embedded in earables, Xu et al. devised Earbuddy [84], a novel system that could detect tapping and sliding gestures near the face and ears. The earables which host an IMU are capable of activity recognition including step counting [65], stay/walk detection, classification of speaking, eating, and head shaking episodes [40], drinking or chewing [55], and exercising [43]. Additionally, they enable passive monitoring of certain indicators of emotional state such as the presence of frown and smile [50] and head movement patterns [66].

In context of facial movements, there are several works using commercially available or custom-built earables to realise FER systems. CanalSense [10] uses barometers embedded in earphones to design an Outer Ear Interface (OEI) that recognizes face-related movements. The system responds to changes in the air pressure of the ear canal to differentiate between 11 facial movements with an accuracy of 87.6%. EarFieldSensing (EarFS) [53] is another system that customises an earphone to sense electric field changes in the ear canal through various electrodes. EarFS was able to recognize 5 facial expressions with an accuracy of 90%. Adding to this list, Amesaka et al. [9] installed a microphone next to the speaker of the earphone to efficiently record in-ear ultrasonic sounds in response to various facial movements. They reported an accuracy of 90% across 6 different facial expressions. Shifting attention to mouth-based movements, Taniguchi et al. [79] proposed a tongue movement recognition method using an infrared LED and a phototransistor. They showed the application of tongue movements as commands for starting and stopping music in a standard media player. On similar lines, Bedri et al. [17] developed an OEI to monitor jaw motion using non-contact proximity sensors.

As with the broader literature concerning FER, most of these works implement their own hardware prototypes. Additionally, the selection of facial movements for recognition is rather arbitrary and limited in number. Closest to our research is the system proposed by Lee et al. [50], which is the first to demonstrate the capability of IMU-augmented earables in detecting facial expressions. However, their work is significantly discounted by the limited number of detectable expressions (only two: smile and frown) which are rather coarse in nature (a combination of multiple facial AUs). In contrast, we propose the recognition of 32 distinct facial AUs based on the standard FACS [29]. To the best of our knowledge, ours is the first work to achieve such fine grained expression recognition by leveraging ear-mounted inertial sensors. In addition to this, our work extensively evaluates the performance of such a system in multiple scenarios of daily living, testing it further in presence of noisy events.

2.3 Learning Algorithms for IMU-Based Activity Recognition

Inertial sensors are by far the most commonly used sensors for capturing information related to movement of the human body, leading to extensive application in human activity recognition (HAR) and behaviour modelling. Since numerous human activities produce a characteristic pattern when observed in the these signals, it is trivial for modern pattern recognition and learning algorithms to classify them. The classic approach involves the extraction of handcrafted features including statistical (mean, variance, cross-correlation, etc), frequency-domain (spectral power, bandwidth etc.) and time-frequency domain (power-spectral density, wavelets etc.) features [16, 44]. Machine learning algorithms like Support Vector Machine (SVM) [60, 76], Tree-based algorithms (e.g.: Random Forest) [48, 51], etc, are then trained using these features, after appropriate dimensionality reduction procedure. These models perform well in HAR tasks with limited number of classes but lag behind in fine-grained action recognition. Newer approaches using deep learning methods are state-of-the-art algorithms for activity recognition. Andrey [41] used convolutional neural networks for real-time activity recognition from accelerometer data. Karantonis et al. [44] demonstrated the superior performance of a neural network based on convolutional and recurrent units as compared to traditional machine learning methods. Holmes et al. [39] extended the list of deep learning based approaches by deploying a deep residual bidirectional-LSTM which outperformed the state-of-the-art architectures specific to the classification task. Even though Bi-LSTM performed better than previously existing algorithms, it suffers from a long training time due to its non-parallelizable nature. This issue was addressed by the introduction of a new architecture called temporal convolutional network (TCN). Its ability to retain information from distant past as compared to previously existing LSTMs, RNNs and Gated Recurrent Units (GRUs), with a much simpler structure and better performance is what makes it so promising [15, 57].

This section informs our choice of model for implementing ExpressEar. However, the performance of a model highly depends on the task that it is modeling. Thus, in the absence of previous literature that uses IMU data to classify fine-grained facial expressions, we select a few models from the ones mentioned above and perform a comparative analysis to find the best model that suits our purpose (refer to Section 5.1).

3 EXPRESSEAR

We present ExpressEar, an ear-mounted IMU sensing technique that is capable of identifying fine-grained facial actions by leveraging the fact that different facial muscle movements produce distinct signal patterns in the IMU data stream. These subtle but perceptible differences in signal characteristics corresponding to different AUs can be effectively learned by state-of-the-art learning algorithms. Below, we describe our sensing principle and system design in detail.

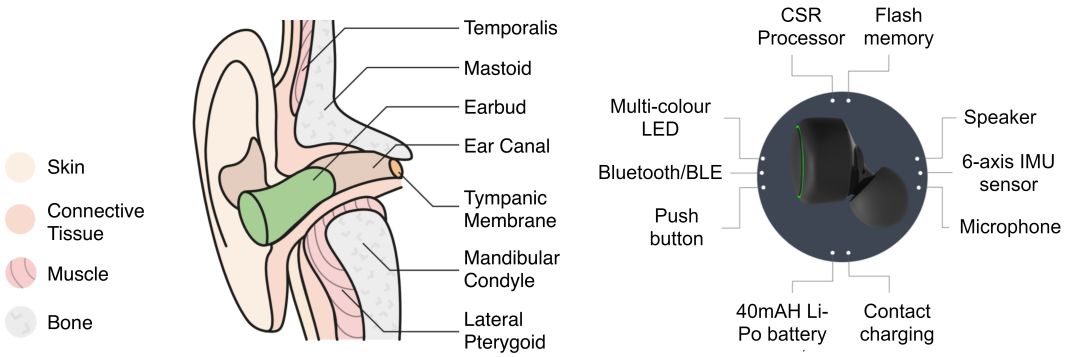
3.1 Sensing Principle

In order to effectively capture the richness and intricacy of facial expressions, behavioral scientists have developed objective coding standards to model them. The FACS developed by Friesen and Ekman in 1978 [29] is arguably the most comprehensive and influential of such coding standards. FACS is based on the anatomy of the human face and encodes expressions in terms of constituent AUs. In total, 46 AUs, including head and eye movements, are defined in the FACS. Among these, 30 AUs are anatomically related to the contractions of specific facial muscles: 12 from the upper face, and 18 from the lower face. These AUs represent visually discernible facial movements and can occur individually or in combination with each other.

Underneath our skin, a large number of muscles allow us to produce a variety of facial configurations. Conversely, when a user performs a face-related movement, the musculoskeletal system changes, affecting the shape of the ear canal [19]. As shown in Figure 2a, ear canal is the narrow tunnel between Mastoid and Mandibular Condyle. Facial AUs like Mouth Stretch (AU 27) or Jaw Drop (AU 26), cause an opening of the mouth which is triggered by a contraction of the Lateral Pterygoid. This process causes the Mandibular Condyle to slide forward,

thus altering the shape of the ear canal. Other facial movements like raising one or both eyebrows (AU 1 or 2) are triggered by muscle groups like Frontalis located on the forehead. However, these movements can also be tapped because most of the muscle groups are connected to Temporalis, the biggest muscle of the head, which transfers the movement to the ear canal. Therefore, when the ear canal is fitted with a material object such as a wireless earbud, the deformation of the ear canal induces a change in the position and orientation of the earbud.

The acceleration and rotation of the earbud show characteristic changes depending on the type and the degree of movement. To illustrate this, the waveforms of acceleration (as measured by an accelerometer) for all facial AUs recorded from the IMU placed in the left ear canal are shown in Figure 3. As shown in the figure, the waveforms differ from each other significantly. We gained additional insights about the richness of the signal by calculating the signal-to-noise ratio for each AU. Comparing the signal power of each AU by the signal power of no expression/baseline, we found the average SNR of all AUs to be 7.8 dB (SD: 1.9, Min: 4.9 dB, Max: 12.2 dB). Hence, by using an IMU-augmented earable inserted into the ear canal, facial muscle movements can be effectively captured. We further employ state-of-the-art learning techniques to learn the distinct patterns in the signal and predict the performed AU.



(a) Ear canal and its surrounding structure

(b) Nokia eSense

Fig. 2. Anatomy of the ear and the earbud's hardware

3.2 System Design

Our system consists of hardware, which includes a pair of wireless earbuds augmented with IMUs, and software that employs an algorithm for recognizing facial expressions from the changes in inertial values. Figure 4 illustrates the overall pipeline of the system, which we describe in detail below.

3.2.1 Hardware. For the purpose of this study, we operate on the *eSense* platform [6, 45] which consists of a pair of true wireless earbuds equipped with kinetic and acoustic sensors (see Figure 2b). Nowadays, these features are available in a range of commercial earbuds from popular brands like Bose, Apple, Amazon, etc. However, while the Bose SoundSport firmware is open, it lacks an application framework [8], and on the other hand, Apple's AirPods, Google's Pixel Buds or Amazon's Echo Buds have unofficial frameworks but no access to the IMU data [7]. eSense, therefore, fills a critical gap by making raw data from the available sensors accessible and offering complete flexibility in the configuration of its parameters. The left earbud of eSense has a 6-axis IMU with a triaxial accelerometer and gyroscope, and a dual-mode Bluetooth/Bluetooth low energy, powered by a CSR processor. To capture multi-stream IMU data from both left and right ear, we take 2 left earbuds and rotate the

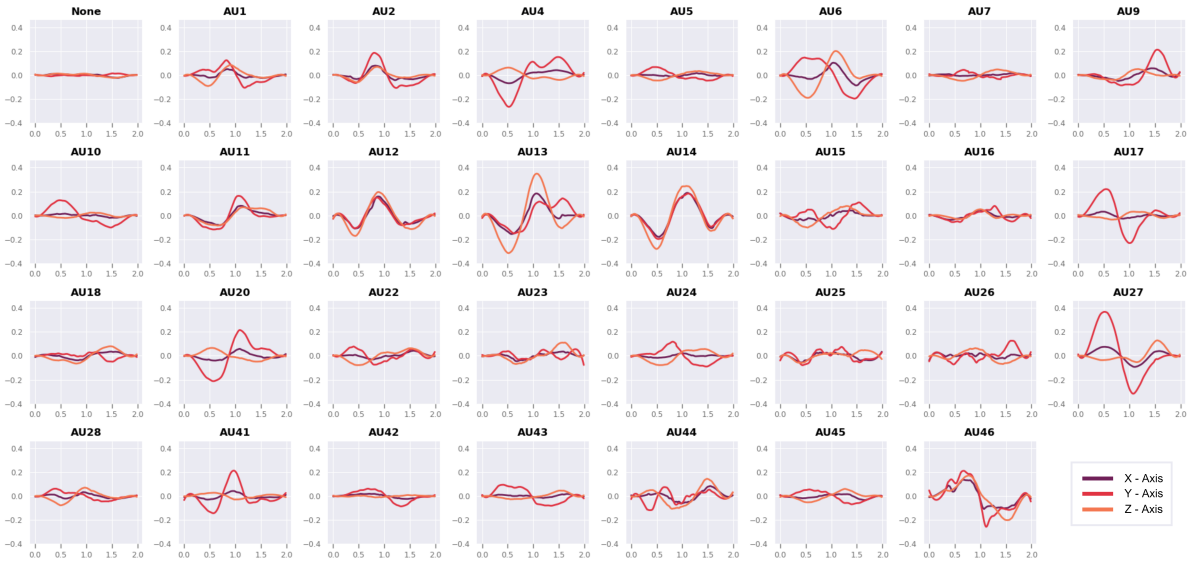


Fig. 3. Raw Accelerometer values (in m/s^2) for all AUs recorded from the left IMU. AU2 and AU46 represent AU2L and AU46L respectively. For ease of comparison, mean-shifted accelerometer values are presented.

earhook of one of them by 180 degrees so that a user can comfortably wear a modified left earbud in their right ear. Moreover, since it was crucial to ensure that the earbuds stay put in the ear canal, the user was made to choose suitably-fitting eartips so that the earbuds don't fall in case of significant movement of facial muscles. Other hardware specifications of eSense include a 40 mAh battery, a light-weight body weighing 20 grams, and dimension of $18 \times 18 \times 20$ mm (including the enclosure) [45]. With a reasonable energy profile, it offers 3 hours of inertial sensing at 50 Hz allowing continuous recording of multiple sessions without any disruption. We stream the 6-axis IMU data from both earbuds at a sampling rate of 50 Hz with a 16-bit resolution.

3.2.2 Software. The software of our system mainly consists of our recognition algorithm which is further divided into continuous extraction and processing of 2-second windows from two IMU data streams (Left and Right IMU), and a classification algorithm based on temporal convolutional networks to predict the performed action unit.

Calibration and Preprocessing. Since different users may have different neutral or natural head orientation, we calibrate the readings for each new user who uses ExpressEar. Once the hardware is set up and the user is ready with his/her head upright, we record 4 seconds of IMU data (200 samples) from all channels (axes) and take the sample mean of these values which we call offset. To account for the drift in gyroscope measurements, we adjust the gyroscope readings of each frame by the calculated offset.

After calibration, a sliding window procedure with a 2s window (chosen experimentally) and a step size of 500 ms is applied to extract frames for final prediction. Although our hardware is equipped with a built-in bandpass filter, filtering the data further can attenuate residual noise. Thus, to improve the signal-to-noise ratio, we apply both a low pass filter (cutoff frequency: 10Hz) and a high-pass filter (cutoff frequency: 0.1 Hz) on each axis of the segmented frame.

Classification. The processed triaxial accelerometer and gyroscope data from both the IMUs are stacked together to make a 12-channel window with a length of 100 samples (input size: 100×12) which is ready for classification. As described in Section 2.3, both traditional machine learning algorithms and complex neural

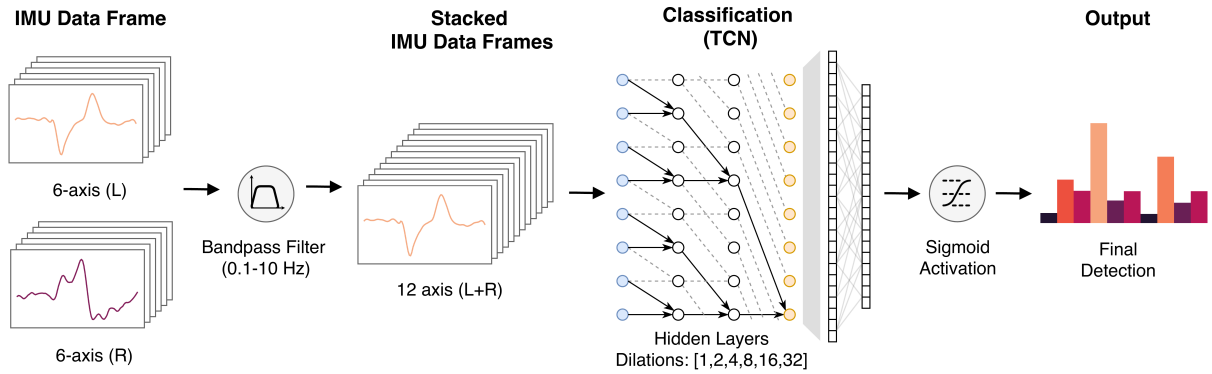


Fig. 4. ExpressEar's processing and classification pipeline overview. Our temporal convolutional network (TCN) architecture comprises several convolutional units (four shown here) and one fully connected layer with sigmoid activation.

networks show exceptional predictive power in a range of HAR tasks. Therefore, we evaluated the performance of standard ML models like XGBoost and SVM, and compared them with various deep learning techniques like 2D-CNN, Bi-LSTM and TCN as shown in Section 5.1. We found that a dilated TCN produced the best accuracy for our data, leveraging the superiority of temporal convolutional networks in terms of parallelizability, flexible receptive field, and most importantly a low number of trainable parameters despite having a deep network of convolutional layers. In context of TCNs, a network is made up of a series of blocks, each of which contains a sequence of convolutional layers. In this work, we use a 12-layer (single block) dilated TCN with a receptive field of 64 samples and a hidden size of 64 units for each layer. The structure of the TCN model is visualised in Figure 4. Each convolutional layer is followed by a batch normalization layer to ensure better generalization and a spatial dropout layer with a dropout rate of 0.3 to avoid overfitting. With dilations of 1, 2, 4, 8, 16 and 32, the number of trainable parameters for our TCN totalled up to 95,328.

The TCN was trained to identify AUs when they occur singly or in combination with each other. To achieve this, we apply the sigmoid activation, which facilitates multi-label classification, on the final fully connected layer:

$$Y_i = \frac{1}{1 + e^{-x_i}} \tag{1}$$

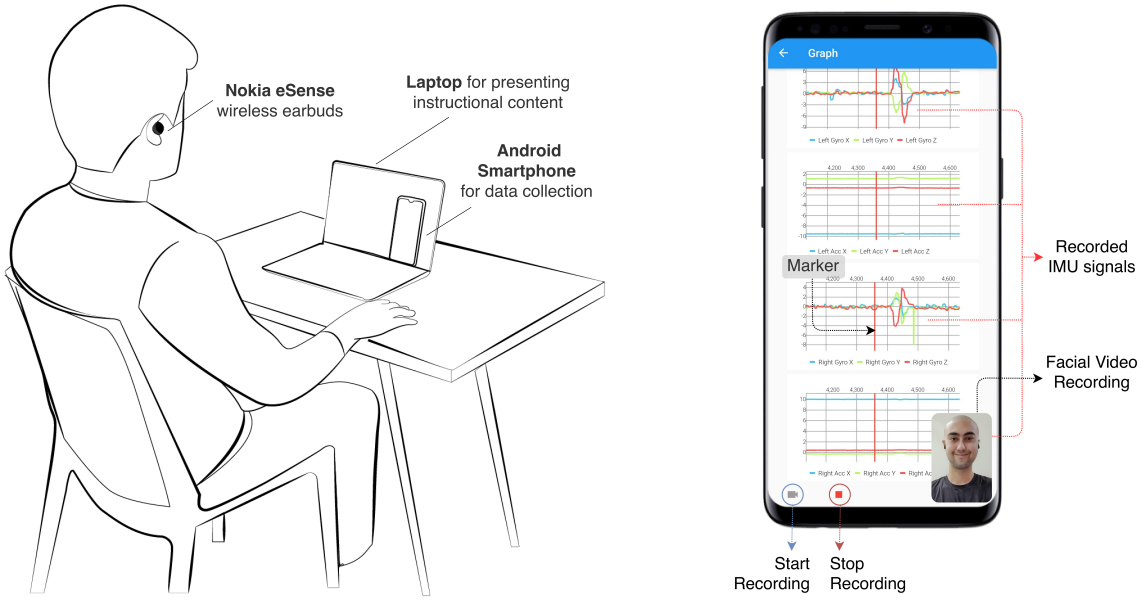
where x_i are the original outputs and Y_i are the transformed outputs. In this way, when AUs occur in combination, multiple output nodes could be activated and the model can simultaneously detect the presence of different AUs. We build and train the described network with an Adam optimizer (learning rate: 0.01) and binary cross-entropy loss function to produce the final classification model used by ExpressEar.

4 EVALUATION

We wanted to investigate several important questions related to sensing of facial AUs, described as follows:

(1) Is it feasible to accurately distinguish between a diverse set of facial AUs using ear-mounted IMUs?
(2) How well can modern machine learning or deep learning approaches perform on such a classification task?
(3) How well does the model perform when tested in a user-independent and dependent setting?
(4) Is it possible to detect facial AUs with just a single IMU sensor in either ear?

Therefore, we conducted a study with 12 participants, the details of which are described in the following subsections.



(a) Experiment Setup

(b) Custom android app for data collection

Fig. 5. Different components of the apparatus

4.1 Participants

Due to the ongoing COVID-19 pandemic, recruiting participants for an in-person study was exceptionally challenging. Moreover, after discussion with the Institutional Review Board (IRB) in our university, we only got approval for conducting the study remotely to avoid any risk. Therefore, we conducted the study with 12 participants (4 females), including three co-authors, ranging in age from 19-23 (Mean: 20.8, SD: 0.87). Since the study had to be conducted remotely, having access to the internet, availability of an android smartphone and a computer with webcam were set as inclusion criteria while recruiting participants. All of them participated voluntarily, after informed consent and were compensated with $5.4 for their time. All the participants reported a normal or corrected-to-normal vision, no prior history of injury in the ear canal or any kind of facial muscle/bone misalignment.

4.2 Setup

The study was conducted in a closed room at the participant's residence with a quiet and isolated environment. A generic setup (as shown in Figure 5a) consisting of a table, chair, laptop, an android smartphone and the Nokia eSense was used to ensure similarity in environmental conditions across participants. Each participant was connected to the experimenter via a video teleconferencing service (Google Meet), running on the laptop as a background process. The laptop (OS: Windows/Mac, Screen Resolution: 720p) was used to present instructional content for the study. After wearing the wireless earbuds in both ears, the IMU data was streamed (Sampling Rate: 50Hz) via Bluetooth and recorded on the smartphone using a custom-built android app. The android app captured IMU data (6-axis: *accel_x, accel_y, accel_z, gyro_x, gyro_y, gyro_z*) from both the earbuds (Left and Right) and also recorded facial videos (Resolution: 480p, Frames: 30fps) of the participant while performing the expressions. For

the same, the smartphone was positioned in a way that it could capture a clear view of the participant's face. Figure 5b illustrates the android app along with its functionalities.

4.3 Design and Procedure

In order to answer the aforementioned questions, we conducted an extensive within-subject study in which we recorded 2240 facial AUs (32 facial AUs $\times$ 70 repetitions) from each participant. To avoid fatigue, we split the data collection into seven rounds, where each round included ten repetitions of each facial AU performed in sequential order. This essentially led to a 7×32 factorial design with Round and Facial AU being factors. The order of 7 rounds and 32 facial AUs were counterbalanced to reduce the ordering effects.

At the beginning of the study, the researcher introduced the protocol and answered participant's questions, if any. Then the participant sat down on a chair along with the apparatus (as described in section 4.2). First, the researcher led each participant through a brief practice session to familiarise him/her with all the facial AUs. Then, the participants were asked to perform all 32 facial AUs in seven rounds of data collection, with each round containing ten repetitions per facial AU. An online form and the android app was used to facilitate data collection. The form presented all the facial AUs one-by-one with a label (e.g. AU 21) and an instruction GIF animation[1] for each of them. The participants were required to imitate the facial expression in the GIF and record the associated IMU signals (6-axis left and 6-axis right) and facial video using the app (as shown in Figure 5b). To label the ground truth, the participants were asked to enter the AU label (shown in the Google Form) in the Android app before starting a recording of that AU. In case of inability to perform specific facial AUs precisely, the participants had the option of skipping and reporting those facial AUs. For ease of segmentation, the participants were instructed to perform the facial AU only when they heard beeps from the app. In each recording, there were ten beeps (one for each repetition) which were scheduled at intervals of 3 seconds and acted as markers. After the end of each round, a 5-minute break was given, wherein the participants were asked to remove the earbuds and then put them back into their ears, allowing any possible variations in earbud positioning. Eventually, it took approximately 230 minutes (excluding breaks) per user to complete the study.

As mentioned earlier, the study was conducted remotely, and under the supervision of one of the researchers using video conferencing. Supervision was required for monitoring the precision of imitating the facial AUs and real-time support for troubleshooting. Since we wanted the participants to perform the shown AU in a natural and intuitive manner, we didn't control the intensity with which they performed it. This made our approach robust to varied intensity levels that could be observed across different rounds. While monitoring the participants, we tried to minimise the observer bias (*Hawthorne effect*) by running the video call as a background process so that they do not feel the presence of the experimenter.

4.4 Dataset and Annotation

After completing the study with all participants, three researchers examined the collected data (facial videos and IMU data) for any inconsistency in facial AUs performed or recording failure. They inspected facial videos to verify the ground truth and mark the start and end points of each facial AU to identify an appropriate window length for segmenting the IMU data. After analysis, the window-length was fixed to 2 seconds, resulting in clipping of 2 second-long segments from each recording, with their start marked by beeps. The data collection app allowed the users to enter facial AU number for each recording, which was used to label the clipped segments. In total, we accumulated an average of 2100 labelled segments per participant which sums up to 25200 segments (for 12 participants). Out of all, nearly 5% segments were discarded due to errors (imprecise facial AU or recording failures), leaving us with $\approx$ 24000 segments. In addition to this, we included time shifted windows ($\pm$200ms and $\pm$400ms) of these labelled segments for augmenting our dataset, resulting in a total of 120K labelled segments.

[1]Source: https://imotions.com/blog/facial-action-coding-system/

5 RESULTS

We applied learning-based predictive analysis on the collected data to support our investigation of the research questions. In this section, we describe the analysis methods adopted and report the results in different testing scenarios.

5.1 Model Selection

To assess whether ear-mounted IMUs provide sufficient information to distinguish between a diverse set of facial AUs, we trained various machine learning and deep learning models on our dataset. Before training, we filtered the dataset in the same way as described in the section 3.2.2.

For machine learning approaches, we featurised the filtered data into time domain (*mean, root-mean-square, zero-crossings,* etc.), frequency domain (*spectral power, entropy, DC component* etc.) and time-frequency domain (*wavelet coefficients*) features for each axis [78]. In total, we computed 71 features per axis resulting in a feature vector of 852 dimensions (71 features/axis $\times$ 6 axis/IMU $\times$ 2 IMU). We further applied a *Gini-importance* based feature selection to reduce the vector size to the most important 150 features. Using these features, we trained a SVM (C: 10, Gamma: 0.1, Kernel: RBF) and a XGBoost (Num Trees: 100) classifier (one-vs-rest). We used the *sklearn* [61] and *xgboost* [22] implementation of these models in python.

For deep learning approaches, we first stacked the filtered data from both IMUs to form a 12-dimension signal. Then, we rearranged this data according to the input size of the network used (refer Table 1). Since IMU data is multi-dimensional and inherently sequential in nature, we tried our hands on a wide range of architectures: CNN [74], Bi-LSTM (2 layers, Units: [64, 128]) [14] and TCN (Units: 64, Dilations: [1,2,4,8,16,32]) [15]. All these models were trained from scratch with Adam optimizer (Learning Rate: 0.01) coupled with learning rate decay (whenever the validation loss plateaued) of 0.1 and Sigmoid layer as a classifier with Binary Crossentropy as loss function. We used Keras [24, 67] and Python to implement and train these models.

We followed a 5-fold cross validation scheme to evaluate the models (96K train, 24K test instances). While creating the train-test split, we made sure that all the corresponding time-shifted segments (refer section 4.4) belong to the same set (train or test) as their central (non-shifted) segment. This was done to prevent any leakage of information. Table 1 provides the classification performance of all the models along with their input dimensions and the number of trainable parameters (wherever applicable). TCN, with its robust capability to model sequential data outperformed the rest with an average classification accuracy of 90.2% (SD: 0.82). Therefore, we chose TCN for implementing ExpressEar and consequently for further testing. Figure 6 presents the confusion matrix for TCN's performance across all 32 facial AUs. AU1: *Inner Brow Raiser* had the highest accuracy (98.1%), followed by AU2L: *Outer Brow Raiser-Left* (97.4%) and AU9: *Nose Wrinkler* (95.2%). AU42: *Slit* had the lowest accuracy (84.9%), which may be explained by the relatively subtle nature of the expression (see Figure 1). Overall, all the models performed well above chance (3.125%), which seems quite promising for a fine-grained FER system.

Table 1. Cross-validation results of different models across 32 facial AUs, along with input size and trainable parameters

Model	Input Size	Trainable Parameters	Accuracy (Mean ± SD)
XGBoost	$N \times 150$	-	82.3 ± 2.3
SVM	$N \times 150$	-	64.5 ± 4.5
Bi-LSTM	$N \times 100 \times 12$	450,952	85.6 ± 3.4
2D CNN	$N \times 100 \times 12 \times 1$	6,816,292	88.4 ± 1.8
TCN	$N \times 100 \times 12$	95,328	**90.2 ± 0.8**

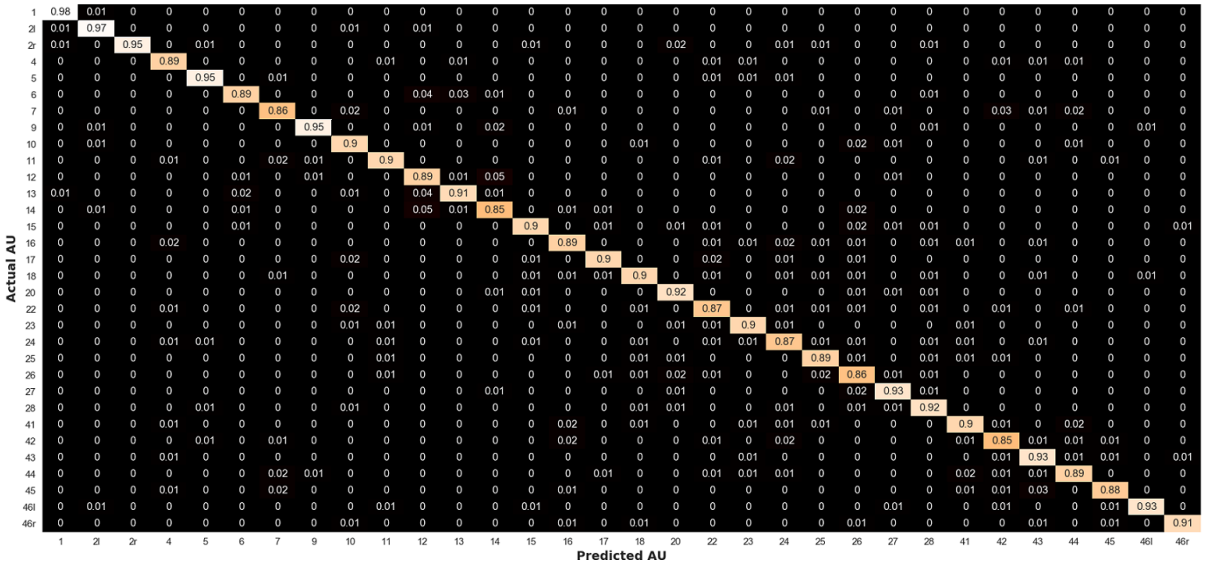


Fig. 6. Confusion matrix for the best model (TCN) on the best fold.

5.2 Leave-One-User-Out Results

We wanted to investigate the performance of our best model across users. Essentially, we wanted to test whether the selected model could work for new users. For the same, we conducted a leave-one-user-out evaluation using the data of one participant as test set and data from remaining participants as train set (110K train, 10K test instances). We repeated this procedure for all participants and then averaged the accuracy results. The mean leave-one-user-out accuracy was 42.1% (SD: 7.3) for our model, which is still well above chance (3.125%) for a 32 class classification. However, we observed a significant drop in accuracy compared to the previous result (90.2%). We speculate two main reasons for this drop: (1) Users may perform facial expressions in unique ways, and/or (2) users' unique facial structure can produce inertial movements in slightly different ways in response to a given facial AU. This motivated us to shift to a user-dependent analysis.

5.3 Per-User Results

In order to perform user-dependent analysis, we trained per-user classifiers for each participant. We trained the model using a single participant's data and performed a 5-fold cross validation using the same participant's data (8K train, 2K test instances). Across all participants and 32 facial AUs, we achieved a mean per-user accuracy of 89.9% (SD: 3.5, Max: 94.5%, Chance: 3.125%). It should be noted that we used data from both the IMUs (L+R) to train these models. Figure 7 shows the cross-validation accuracy for each per-user classifier.

5.3.1 Effect of Plurality of Sensors. As described in section 3.2.1, only the left earbud hosts the IMU sensor in eSense. Although we made suitable changes to support our exploration, we were intrigued to test fine-grained FER using a single IMU in either ear. For the same, we trained per user models similar to the previous experiment, using data for only one IMU at a time (6-axis data). The average per-user accuracy across all subjects was 79.6% (SD: 7.1, Max: 87.0%) and 79.3% (SD: 5.7, Max: 88.7%) for left (L) and right (R) respectively. As shown in Figure 7, there were significant differences (≈10%) between using both IMUs (i.e., L+R) and using a single IMU (i.e., L or R). We also validated the statistical significance of our result with a *Friedman test* (L and L+R: $p = 10^{-5} < 0.05$,

R and L+R: $p = 3 \times 10^{-6} < 0.05$). As anticipated, the opposite asymmetrical AUs: AU2R and AU46R (for left earbud), AU2L and AU46L (for right earbud) had the lowest classification accuracies of 60.1%, 54.5%, 62.7% and 57.2% respectively. These results verify that asymmetrical AUs produce asymmetrical facial movements, thus explaining why Lee et al. [50] could not observe a clear impulse response for AU2 (default: AU2R) in the inertial signals captured from left earbud. Nevertheless, a single IMU-augmented earbud can also be reliably leveraged for recognizing most of the facial AUs (excluding the asymmetrical ones) with a decent accuracy.

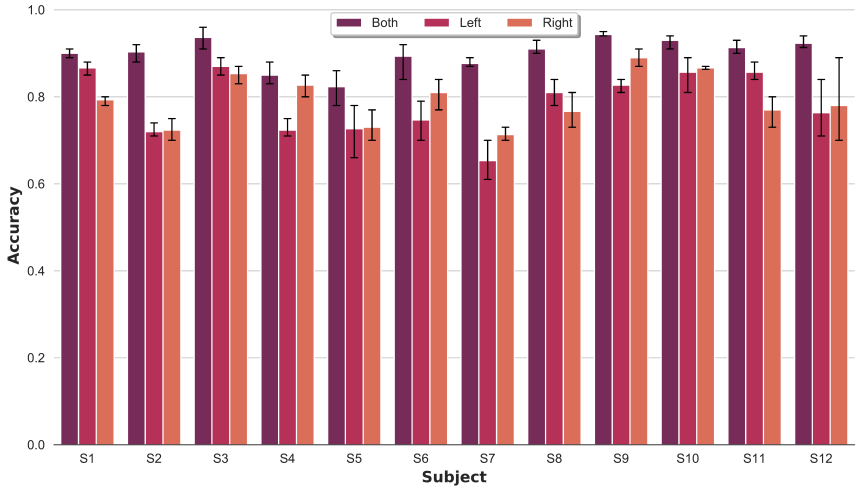


Fig. 7. Per-user Accuracy of all subjects across 3 IMU positions: mounted in left ear, right ear and both ears. The error bar indicates the variation across different folds.

6 BETTER ESTIMATING REAL WORLD ACCURACY

Although the aforementioned average of 89.9% per-user accuracy follows a standard evaluation procedure, it depicts the performance in a controlled laboratory environment. The absence of significant macroscopic movements of the jaw, head and the entire body allows the ear-mounted IMU sensors to capture the subtlest of facial expressions. However, in real-world scenarios, ExpressEar is likely to encounter “unknown” facial-related movements (e.g. movement of jaw caused by eating) and other motion artifacts (e.g. head movements caused by ambulation). Consequently, the absence of an unknown class might lead to undesirable false positives, whereas strong motion artifacts might suppress an expression altogether, again derailing the system’s performance in real-world settings. Thus, we conduct two additional experiments to estimate our system’s performance under challenging conditions of noise and user mobility. The objective, procedure and results of the two experiments are described in the following sections.

6.1 Distinction from "Unknown" Face-Related Movements

As described in Section 3.1, ExpressEar senses the change in position and orientation of an ear-mounted IMU in response to a characteristic ear canal deformation, arising from the movements of jaw, lips, nose, eyes, eyebrows and head. Apart from facial expressions, face-related movements mainly consist of mouth motions like speaking and chewing, and head motions. Ideally, our system should be able to reject these movements as noise or classify them as “unknown” and trigger the facial AU classifier only when the possibility of noise is eliminated. In line

with this idea, we conceptualise the recognition module of ExpressEar as a hierarchical classifier, wherein it first distinguishes between the presence and absence of facial AUs, followed by inter-AU classification conditioned on the presence of facial AUs. Since we have already evaluated the classification results across 32 AUs under different testing conditions, we now assess the discriminability of the complete set of facial AUs from the absence of any expression (which we call *baseline*) and other face-related movements.

6.1.1 Data Collection. To investigate the feasibility of noise detection, we recorded data for three classes, namely eating (chips), speaking and no action (baseline), from 3 participants. Instead of sampling at discrete time intervals, these activities were continuously segmented using a sliding window procedure (window size: 2s, step size: 100ms) over the entire duration of a data collection round. Hence, we collected a total of 8700 data samples (3 activities $\times$ 5 rounds/activity $\times$ 580 instances/round) from each participant, recorded across 5 rounds of 1 minute each. We used the same setup and apparatus as described in Section 4.2 and provided the participants with a bowl of chips for eating and an essay to read out loud for speaking. The participants were not restricted from moving their upper body or head while performing the instructed activities in an attempt to simulate a real-life situation.

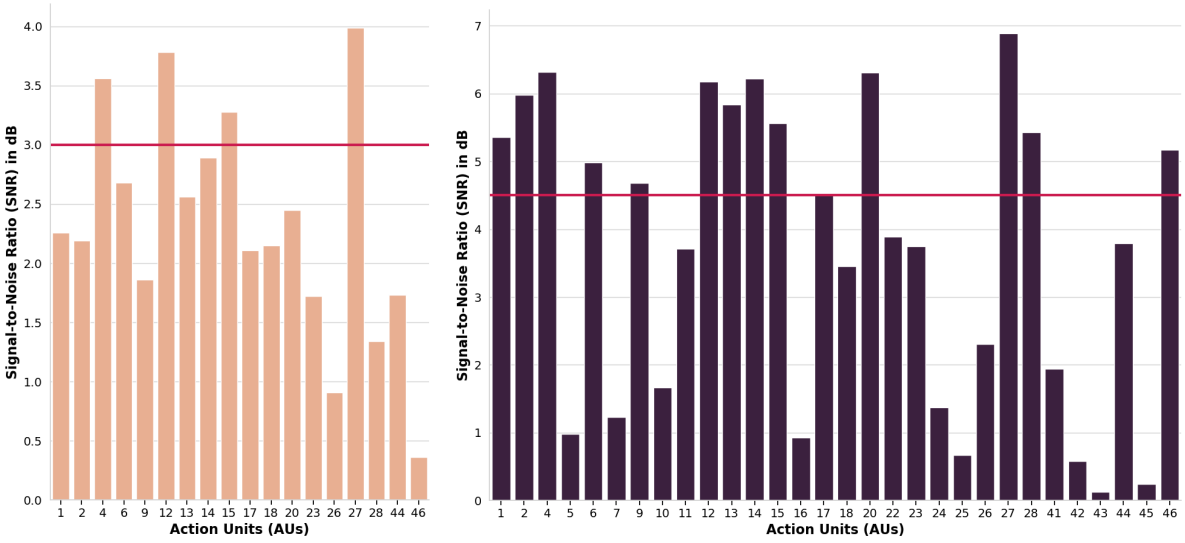
6.1.2 Results. As identified through multiple experiments in Section 5, ExpressEar achieves the best results with a TCN trained on multi-sensor (Left and Right IMU), per-user data. Thus, we trained per-user binary classifiers after labeling each of the additional classes (eating, speaking, none) as *unknown* and all facial AUs as *AU*. The model was evaluated using a 5-fold cross-validation scheme, resulting in 15K train and 3.74K test instances per fold. Across all participants, we obtained an average per-user accuracy of 99.2% (SD: 0.56, Max: 100%). The almost perfect result is unsurprising since there is a considerable difference in both the signal patterns and intensity (measured by SNR) of the two sets of classes. The mean SNR of the unknown class samples belonging to eating and speaking activities is 12 dB, which is 1.5 times the average SNR of AU samples (7.8 dB). On the other hand, the SNR of baseline samples is centered around 0 dB.

6.2 Impact of Body Motion

So far, we have studied the signal characteristics of ear canal deformations caused by various face-related movements. However, ear-mounted IMUs are susceptible to motion artifacts arising from other body motions of the user as well. While most of the arbitrary limb motions are naturally filtered out due to the placement of the IMU, whole-body motions comprising both passive movement in a vehicle or active locomotion, exert accelerative forces on the entire body and consequently, the worn sensor. These forces introduce significant noise in the inertial signals that is capable of altering, and sometimes even suppressing, the response of an expression. Thus, as representatives of passive and active motion classes, we evaluated the system's performance when subjected to a moving rapid transit system (Metro train in our city) and ambulation, respectively.

6.2.1 Data Collection. We collected a new dataset of all facial AUs performed by three participants in two scenarios: *walking* and *traveling in a metro train*. In order to maintain a stable, obstacle-free setup which doesn't distract the participant from the main task, we instructed the participants to walk on a treadmill operating at a speed of 4 km/hr for the first scenario. We also replaced the laptop shown in Fig 5a with a more portable smartphone which is held in the user's hand while traveling in a metro and placed against the display of the treadmill while walking on it. With an additional exception of omitting video conferencing since the participants were the researchers themselves, we follow the same procedure to record the data as mentioned in Section 4.3. The segmentation and ground truth labeling of the data windows was also carried out in a similar way by using beep markers in our Android app and a text field for the AU in our Google Form that is filled by the participant before recording an expression. In this manner, we were able to accumulate 3200 facial AU instances (2 scenarios $\times$ 32 AUs $\times$ 5 rounds $\times$ 10 instances/round). Each round of data collection was accompanied by a 1 minute baseline collection for both the scenarios. The baseline data, which refers to no expression performed while

walking or traveling in a metro, was segmented using a sliding window of 2s with a step size of 100ms, leading to a total of ≈11,800 (2 scenarios × 5 rounds × 1180 instances/round) per participant. For AU instances, we also include the time-shifted instances (±200ms, ±400ms) for classification, resulting in a total of 16K AU instances per participant.



(a) Walking

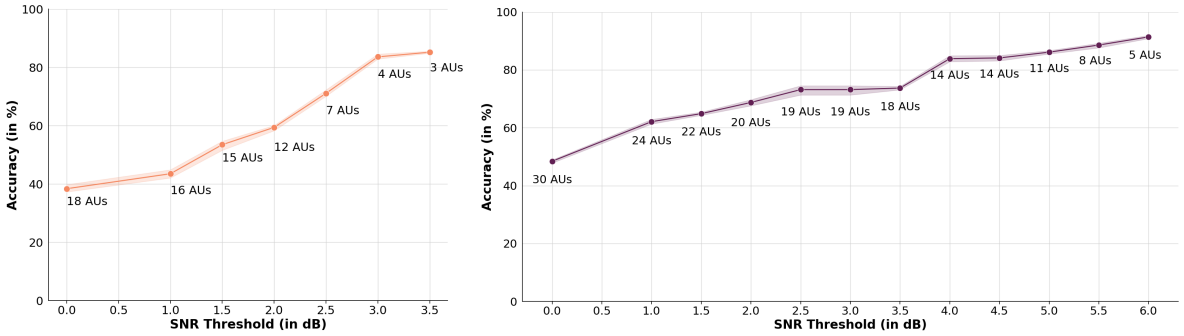
(b) Moving metro

Fig. 8. Average SNR for each AU performed in two different types of motion conditions. Only the AUs which had an SNR greater than 0 dB are displayed.

6.2.2 AU Selection. As described above, whole-body motions tend to make the ear-mounted inertial sensors insensitive to certain facial expressions which were effectively characterised by discriminable signal responses in a stationary setting. Thus, we determined the optimal set of potentially classifiable facial action units, for both walking and moving metro, in accordance with the signal-to-noise ratios (SNR) of the facial AUs. For calculating the SNR, we consider the baseline data in both cases to be the noise signal. The rationale for filtering out a set of AUs is to maximize the system's performance under varying conditions of user mobility. If the signal power of an expression does not even exceed, or is comparable to, the baseline (noise) signal's power, it is unlikely for that signal to be discernible by any learning algorithm.

We calculated each sample's SNR, averaged across all the 12 channels, and prepared a set of AUs for each scenario (walking and moving metro) based on the SNR threshold values, which were set to 3 dB and 4.5 dB for walking and metro respectively. These threshold values were computed by scaling the maximum baseline SNR (ratio of maximum baseline power to average baseline power) by a factor of 3. We also show the effect of the SNR threshold and number of selected classes on the average accuracy of the system, in Figure 9. It should be noted that each sample was compared to its corresponding baseline, i.e it was ensured that both the *signal* and *noise* belonged to the same participant and was collected in the same scenario (walking or moving metro). As shown in Figure 8, 14 AUs cross the SNR threshold for moving metro and just 4 AUs for walking. Although the number of AUs selected for walking is quite low, it is still an extension over prior work [50] where they distinguish between

two AUs in ambulatory conditions. We use these sets of AUs for final classification and report the performance of our system in the following section.



(a) Walking (b) Moving metro

Fig. 9. Average per-user accuracy vs the SNR threshold (to select the AUs) for each motion condition.

6.2.3 Results. We evaluated our system on detection and classification of facial AUs for both scenarios of user mobility. We trained individual classifiers for each scenario.

For AU detection, all the AUs which qualified (were greater than) the SNR threshold were labelled as class *AU*, and the remaining ones along with the baseline formed the *unknown* class. We believe that a low SNR indicates inconsistency or absence of a distinct signal pattern for an AU and therefore must be considered as unknown. On training per-user TCNs with 5-fold cross-validation, we achieved an average accuracy of 99.21% (SD: 0.02, Max: 99.48%) and 99.86% (SD: 0.04, Max: 99.92%) across 2 classes (AU and unknown) for walking and metro, respectively. Further, for inter-AU classification, we trained per-user TCNs on the selected set of AUs. With the same training parameters as before, it yielded an accuracy of 83.85% (14 classes; SD: 1.25, Max: 85.61%, Chance: 7.143%) for metro and 84.34% (4 classes; SD: 0.63, Max: 84.89%, Chance: 25%) for walking.

7 APPLICATIONS

Facial expressions occupy a central role in human social interaction. They are natural, intuitive, diverse and contextual, which make them critical to human behaviour as well. Because they are so deeply embedded in our daily lives, continuous monitoring of fine-grained facial expressions enable a wide range of applications that have been well motivated in previous research. We believe our work points towards a more discreet way of bringing these use cases closer to feasibility, that too with a commercially available wearable.

Human beings may perform facial expressions involuntarily (*e.g.*, a piece of good news induces a smile) or voluntarily (*e.g.*, purposeful use of 'eyebrow flash' to greet others). While monitoring the former could yield information about the user's affect and behaviour [31], the latter could be seen as an opportunity to create a novel input space leveraging these expressions for intuitive human-computer interaction [53]. Therefore, we segregate the potential applications of ExpressEar broadly based on the two categories mentioned above and briefly discuss them in the following subsections.

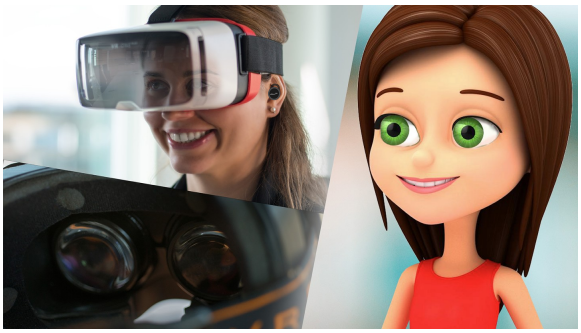
7.1 Monitoring Involuntary Facial Expressions

Involuntary facial expressions are believed to be innate in that they reflect the natural response and behaviour of a user. The ability of ExpressEar to continuously monitor these expressions presents interesting applications

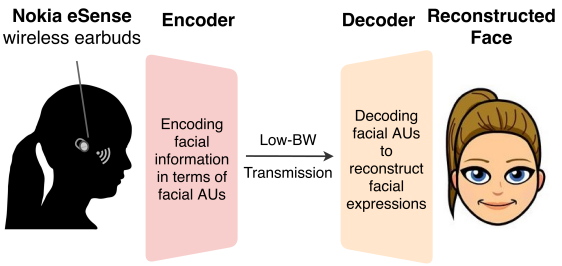
ranging from mirroring the expressions of a user on a virtual avatar to building intelligent agents that map expressions to emotions and respond accordingly.

7.1.1 Application Scenarios. We envision three different application scenarios demonstrating the use of monitoring involuntary facial expressions.

(1) **Facial Expressions in Virtual Reality:** Virtual social environments (e.g. collaborative VR games) can often lead to an inanimate experience due to the lack of liveliness in characters. In such a scenario, having real-time facial expressions mirrored on avatars can lead to personalised and more immersive social experiences. However, traditional camera-based FER systems suffer from occlusion due to head-mounted displays (HMD), and occlusion-invariant sensing mechanisms like eye-tracking (inside-HMD) [38] and facial EMG [56] fail to capture fine-grained facial expressions. With FACS being the "gold-standard" in facial graphics and animation [73], ExpressEar could suitably fill this void to perform FACS-based expression recognition for enhancing virtual social experiences.



(a) Facial expressions for virtual reality. User's expressions are monitored in real-time and imitated by a character in virtual environment leading to personalised and more immersive experiences.



(b) Low-bandwidth video conferencing using ExpressEar. (1) User's facial expressions are encoded into AUs, (2) the encoded information is transmitted to a decoder, and (3) the decoder reconstructs the facial expressions onto a virtual avatar.

Fig. 10. Application illustrations: involuntary facial motions

(2) **Low-Bandwidth and Hands-Free Video Conferencing:** With a global pandemic in place, whether it is to attend work meetings from home or to remain connected to near and dear ones, video conferencing technologies have impacted our lives deeply. Despite the rapid growth of internet communication technologies, there are still numerous places on earth where high-bandwidth internet is out of reach[2]. To realise very-low bandwidth video conferencing, a system could be designed in a way that it encodes the facial expression information at the transmitting side (using models like FACS), transmit the encoded face, and reconstruct the face on the receiving side (as described in Figure 10b). In contrast to the real-time transmission of pixel-based image frames, such a technique would consume significantly lower resources while delivering a near-to-similar experience. With the current capabilities of ExpressEar such a system could be implemented easily.

[2] https://www.fastmetrics.com/internet-connection-speed-by-country.php

(3) **Emotional Awareness for Intelligent Agents:** Numerous previous researches [28, 31] have tried to establish a stable relationship between the different emotional states and the FACS. For example, *happiness* could be detected from the combined presence of AU6 (Cheek Raiser) and AU12 (Lip Corner Puller). Monitoring emotional state could lead to a better understanding of the needs and expectations of a user. For instance, detecting sadness on the playback of a song could enable the music player to suggest a mood-lifting song, or monitoring user's frowning/surprised reaction to an intelligent assistant's (like Siri, Alexa) unexpected response could facilitate timely interruption of the assistant to initiate the correct response [85]. With the help of ExpressEar, all of this is even closer to feasibility without compromising robustness, comfort and privacy.

7.2 Voluntary Use of Facial Expressions

Humans have an exceptional capability of representing a diverse set of expressions through their face. Leveraging this capability, previous research [10, 53] has very well motivated the use of facial expressions as an input for enabling hands-free and eyes-free human-computer interaction. However, not all humans have the ability to perform certain facial expressions with the same level of control. These expressions may also differ in terms of their frequency of use in daily lives. Moreover, since they are quite closely related to social interaction, there may also be a distinction between expressions which could be performed in public or private. All these are important considerations to be taken into account while designing interactions. Intrigued by this, we conducted a short survey, described in the following subsection.

7.2.1 Survey: Facial expressions for input and interaction. We conducted a survey (N=18, 6 females, mean age: 20.8) to gauge user perception of the facial AUs based on the following factors:

- *Ability*: "I am able to perform the facial AU"
- *Ease*: "I can easily perform the facial AU with precision."
- *Social Acceptability*: "It is acceptable to perform the facial AU in public without social concern"
- *Fatigue*: "The facial AU makes me tired."
- *Frequency*: "I perform the facial AU often in my daily life."

Ability was a binary question ("yes" or "no") conditioned on which ("yes") the participants were allowed to answer further questions for a particular AU. For the other factors, participants rated each AU on a 5-point Likert scale (1: strongly disagree to 5: strongly agree). The order of presenting facial AUs was randomised for each participant to avoid any bias. It took approximately 20 minutes per participant to complete the survey

For analysing the survey responses, we first encoded the ordinal responses into numbers, for Ability ("yes": 1, "no": 0) and for others ("strongly disagree": 0, "disagree": 1, "can't say": 2, "agree": 3, "strongly agree": 4). Then we totalled the encoded ratings per factor for each AU. For every AU, we normalised the total (per factor) by the number of participants which were able to perform that AU. This was done for all factors except Ability. The overall mean ratings for *ability* across all facial AUs was above 16 (SD: 2.16, Max:18), indicating that a majority of facial AUs were considered 'able to perform' by atleast 88% of the participants. As speculated, AU2R and AU2L (Outer Brow Raiser - Right/Left) were rated lowest with a score of 9/18 and 12/18 respectively. In terms of *ease*, the average ratings across all AUs was above 3 ("Agree", SD: 0.33, Max: 4), which means that on an average, participants found it easy to perform most of the facial AUs, barring a few like AU6 (Cheek Raiser) and AU20 (Lip Stretcher). Talking about *social acceptance, fatigue* and *frequency*, all three of them were rated above 2 ("Can't Say", SD-Social Acceptance: 0.43, SD-Fatigue: 0.36, SD-Frequency: 0.64, Max: 4). This essentially means that most expressions were neither excessively tiring nor inappropriate to perform in social settings. The higher standard deviation value of *frequency* indicates a high variation among participant responses in terms of the use of AUs in their daily lives.

Finally, to rank these facial AUs according to their appropriateness for use as gestures, we computed a cumulative score for each AU based on the factors mentioned above. To calculate the cumulative score, we first

scaled all the ratings from 0-1. Then, we eliminated the bottom three AUs from each factor individually to remove extremely poor performing AUs. Finally, we combined all the factors (by taking their product) to yield a single score. It should be noted that before computing the cumulative score, we inverted the fatigue (gesture should not be tiring) and frequency scores (frequent AUs could lead to false positives viz. the system may confuse a frequent facial expression performed naturally, e.g. blink, with a gesture). Figure 11 shows the cumulative ratings for all the AUs, sorted in ascending order. The cumulative score weighs each factor equally and tries to optimise between them.

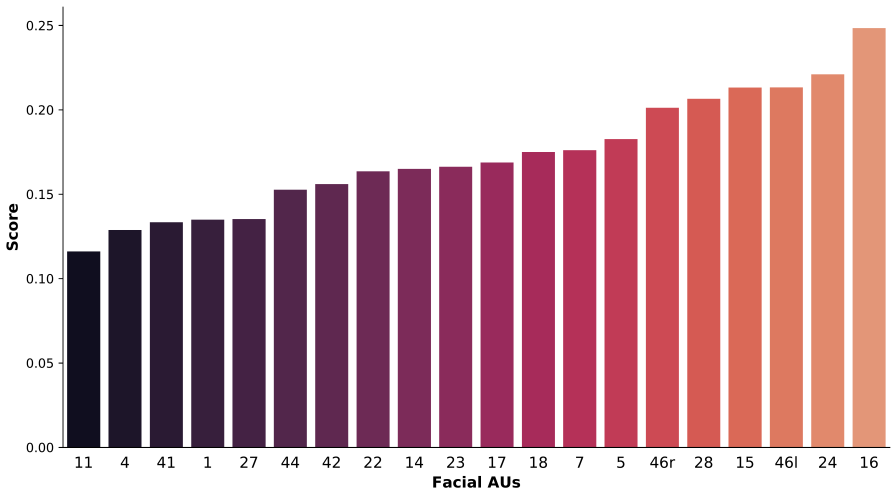
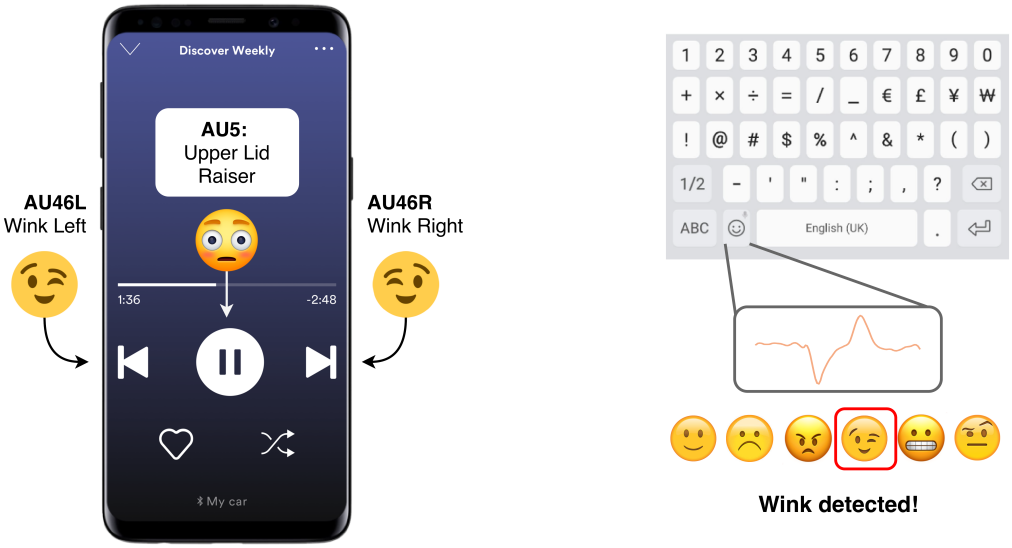


Fig. 11. Cumulative scores of facial AUs based on their ability to perform, ease, social acceptability, fatigue and frequency. AUs (2L, 2R, 6, 9, 10, 12, 13, 20, 25, 26, 43, 45) were discarded in filtering.

After discarding the less favourably rated AUs across different factors, most of the higher rated AUs, as shown in Figure 11, are simple expressions that are socially acceptable and hardly occur as a natural response to daily life situations. This analysis has useful implications for the design of facial expressions-based interactions. By taking various factors of user perception into account, the cumulative score indicates the suitability of an AU to act as a gesture or micro-interaction. We exemplify one such use-case of the same through an application illustrated in the following section.

7.2.2 Application Scenarios. We illustrate three distinct application scenarios leveraging voluntary control of facial expressions.

(1) **Subtle Gestures and Peripheral Microinteractions:** ExpressEar provides a hands-free and eyes-free interface to facilitate microinteractions in scenarios where interruptions may not be desirable. For instance, in an office meeting, it is suitable for the users to issue quick commands for rejecting a call or turning off the notifications. Such a system can be enabled by mapping the required action to the detection of a predefined AU. Similarly, the facial expressions can also be used as subtle gestures to switch between songs on a media player or scroll back and forth on a content reader. Informed by the survey results in Section 7.2.1, we show an illustration of the media player application which uses three of the top 10 AUs for interaction (see Figure 12a).



(a) Use of facial expressions as subtle gestures. User's facial expressions are used as a medium to interact with the music application. Commands: left wink, right wink and upper lid raiser corresponds to previous song, next song and play/pause respectively.

(b) Automatic emoji input. Instead of the conventional way of choosing emoji from the keyboard, the illustration reflects the use of ExpressEar to enable emoji input with just an imitation of facial expression.

Fig. 12. Application illustrations: voluntary facial expressions

(2) **Assistive Human-Machine Interaction for Physically Impaired:** Every year, quadriplegia (paralysis of all four limbs), neurodegenerative diseases like amyotrophic lateral sclerosis (ALS) and trauma puts several people into conditions which inhibit them from controlling joy-stick based wheelchairs [11, 12]. In such scenarios, ExpressEar could be utilised as an assistive interface by leveraging facial expressions for steering the wheelchair. Earable sensing could provide them with a precise, unintrusive and occlusion invariant alternative as opposed to the existing eye-tracking [83], EEG [27] and camera-based [64] solutions which may lack these qualities.

(3) **Automatic Emoji Input:** People often use emojis on social media and instant messaging platforms to express their sentiments. For a user to input emojis, he/she would often explore the long list of emojis present in the keyboard to select the most appropriate one. To make this experience more interactive and less annoying, ExpressEar can be used for emoji input, wherein the user can enter the desired emoji by imitating it using facial expressions (see Figure 12b).

8 DISCUSSION AND FUTURE WORK

Overcoming several impediments through our extensive evaluation, we demonstrate the potential of ear-mounted IMU sensors to detect and recognize a range of facial expressions in multiple scenarios of daily living. While technical challenges remain, the main goal of this work is to shift the community's attention to an unobtrusive,

privacy-preserving and ubiquitous sensing technique that holds the potential to continuously monitor rich non-verbal cues which encode all possible facial expressions at the atomic level. In the following subsections, we discuss the implications of our work, acknowledge its limitations and highlight the compelling research opportunities to advance the field further.

8.1 Motion Artifacts

Repetitive head bounces while walking produce extraneous noise which is prominently visible in the inertial signals [65]. After a thorough spectral analysis of these signals, we found that the noise (normal walking) spectrum overlaps with the spectrum of static (noise-free) expressions. Therefore, common signal processing techniques like bandpass filtering, Independent Component Analysis, and Spectral Gating are bound to fail in restoring the noise separated signal. Further exploring the possibility of noise reduction, we applied spectral subtraction, wherein an estimate of the average noise spectrum is subtracted from the noisy signal spectrum. However, the reconstructed time domain signal from the resultant spectrum neither showed resemblance to the plots shown in Figure 3 nor displayed a different discernible pattern. The failure of spectral subtraction proves that the noise is not additive. Said differently, walking and performing a facial AU simultaneously produces a distorted signal which is different from the individual inertial responses of both walking and the expression.

Similarly, while the motion of rapid-transit trains is close to uniform motion (between the stations), it is frequently interrupted by acceleration at the time of departure from each station, spurious stops and irregular tracks. Nevertheless, in comparison to walking, the intrinsic noise in this case is much lower, sparing a considerable number of AUs from significant distortion. However, due to the random nature of noise, the distribution of the data collected in a moving metro also differs from the static data distribution.

Owing to these challenges, we proposed three independent classifiers for each type of user motion in this work. However, motivated by previous work [18], we postulate that a high speed accelerometer with a much higher sampling rate and resolution would be able to register individual AU responses apart from the walking characteristics or other motion artifacts. An inherent constraint to such a system would be the limited battery life of commercial earables, which is tremendously affected by the power consumption of high speed sensors. Therefore, while the technological advancements attempt to navigate this tradeoff, ExpressEar would work excellently in a sufficiently stable setting and reasonably well under limited mobility conditions.

8.2 Privacy Concerns

Recent advancements in the inference of certain mobility characteristics like gait, which has shown promise to be used as a complementary biometric [80], pose some re-identification risks leading to privacy concerns. However, it is unlikely that an IMU placed at a significant distance from the lower body would be capable of providing a detailed analysis of one's walking pattern. Apart from this, there is growing evidence showing nefarious use of smartphone IMU sensors in location tracking and password sniffing by monitoring a user's application use and the corresponding movements/vibrations picked by the phone's inertial sensors [58, 59]. Although bleak, there are chances that the data collected from ExpressEar may also be exposed to a similar risk if transmitted to a smartphone for further processing and prediction. We acknowledge these risks and seek to design miniature learning modules for ExpressEar which could be easily deployed on the earable's microcontroller in the future. This would allow us to infer facial expressions on the fly without the need of transmitting or storing the raw data for a long time. Nevertheless, these are extreme cases which are less probable to take place in daily scenarios since ear-mounted inertial data has limited capabilities to divulge personal information that can be misused.

8.3 Co-occurrence of Facial AUs

Previous literature has identified more than 7,000 AU combinations in everyday life [37]. In fact, in spontaneous scenarios, it is unlikely for an AU to occur in isolation. Treating these combinations as new independent classes is impractical given the number of such combinations. While our model has the potential to support simultaneous detection of multiple AUs, an extensive study is required to fully understand the response of inertial signals in the presence of more than one AU at a time. Hence, for now, the best way forward is to model the semantics of facial behaviour, i.e., probabilistic modeling of the spatio-temporal interdependencies among groups of AUs. This solution leverages the fact that there exists a strong co-occurrence structure in AUs and there are certain combinations of AUs which often occur together in expressions of emotion [30]. Conversely, it is a well-known anatomical fact that certain AUs can't occur together. This domain knowledge enables us to consider the presence of one AU as a precursor to the presence or absence of other AUs. For example, AU6 is known to co-occur quite frequently with AU12 (in a "Duchenne smile"), so the presence of AU12 increases the chance of AU6 being activated. Utilising co-occurrence information has recently started to gain traction with the development of Bayesian Networks and its variants [77, 86]. Based on these state-of-the-art techniques for probabilistic modeling, we believe that with careful consideration of aforementioned factors of co-occurrence, we can adapt our model for conditional prediction with minor tweaks.

8.4 Individual Variability

The facial AUs defined in the FACS, though distinct, are not uniformly produced or perceived. FACS outlines the anatomic basis of facial expressions by mapping the corresponding facial movements to specific facial muscles. The quality of these movements, however, varies with differences in the facial structure [72]. Various facial muscles are found to be heterogeneous in their form, arrangement and innervation. For instance, Goodmurphy and Ovalle have shown that muscle fiber shapes, types, and sizes in *orbicularis oculi*, *pars palpebralis*, and *corrugator supercilii* are significantly different across people [35]. The presence of the muscles themselves is highly variable, with certain muscles appearing in some individuals and not in others [63]. A clear example of this is the *Risorious* muscle bundle which was observed to be absent in as many as 22 of 50 specimens in a study conducted by Pessa et al. [62]. Other facial features like furrows and facial skin deformations are also produced by variations in facial muscles, leading to individual differences in expression.

Besides the underlying physical variation in the face, empirically measured facial behavior differs in accordance with factors such as sex [20, 21], age [21], and cultural background [47]. Tzou et al. conducted a cross-cultural study with European and Asian subjects to show that in general, Europeans have larger facial movements in terms of the distance between opposite facial landmarks, as compared to Asians [81]. These characteristics affect the intensity of expression and the resultant movement caused by them. Looking at gender-based differences, women have been shown to have thicker *zygomaticus major* muscles [54] and there is also some evidence that while men specialise in performing expressions of anger, women are better performers of happy expressions [25].

Regardless of the degree of variation in facial expressions that can be detected empirically and experimentally, perceivers may not be able to notice these slight variations or may categorise them similarly, with high agreement [33, 75]. Therefore, while conducting the user study, what looked like a decent imitation of a facial expression, was actually responsible for producing different facial movements beneath the skin. Despite the possibility of significant differences in facial structure, we believe that our model performed well in a leave-one-user-out condition with an average accuracy of 42.1% (refer to section 5.2). Although these results are highly impractical for a real-time system, the user-independent models show potential for improvement if a larger training set capturing a variety of facial structures could be collected.

8.5 Influence of Environmental Factors and Audio Playback

Although ExpressEar has shown excellent performance in recognizing a large range of face-related movements, we can't overlook the factors affecting the ecological validity of our findings. These factors include the physical environment of a user such as an elevator which affects the gravity vector and hence the acceleration values. Further, inertial sensors are rated with a maximum and minimum temperature which denotes the acceptable temperature range beyond which the transducer's sensitivity can be permanently reduced. In our case, though, the InvenSense MPU-6500[3] integrated in the Nokia eSense[4] is built to perform seamlessly in a temperature range of -45°C to 80°C, suitable for daily living scenarios. While these environmental factors can have adverse effects on raw IMU signals in extreme conditions, the resultant changes in normal real-life settings would be rather negligible. Acoustic sensitivity, however, poses a cause of concern for further evaluation. The pressure waves from audio playback in the ear can excite the accelerometer and the earable itself. Although these induced vibrations are normally much less than the actual inherent structural vibrations, they remain something to consider. Since the primary purpose of earables is audio playback, we plan on investigating the effect of sound on the inertial data in order to incorporate required corrections that would make ExpressEar more robust.

8.6 Path to a Real-Time System

The challenges discussed so far, especially user variability and sensitivity to motion artifacts, constrain the practical applicability of our system. While user dependence can only be addressed through an extensive evaluation to accommodate substantial user variations or a calibration step in an adaptive model, ExpressEar shows success, albeit limited, in modeling facial AUs in mobile scenarios. However, the inconsistency in the noise introduced by different mobile settings eliminates the possibility of building a "one-fits-all" model. Thus, we propose to conceptualise ExpressEar as a combination of three independent hierarchical models for facial AU recognition in different mobile conditions (static, walking, moving metro). Each hierarchical classifier distinguishes between the presence and absence of an AU before proceeding with inter-AU classification. The user could either provide explicit input to switch between motion modes or rely on ubiquitous smartphone or smartwatch sensors, for the same. Finally, akin to most sensing systems, ExpressEar would encounter latency in processing and classification of facial AUs, along with the transmission time required to send and receive the data from another device that hosts the trained models. As discussed in Section 8.2, miniature learning modules could be easily deployed on the earable's microcontroller to tackle this challenge.

In conclusion, while hardware limitations including limited sampling rate and sensing potential affect the deployability of ExpressEar in mobile scenarios, our system can be easily realised in a stationary setting with trivial modifications.

9 CONCLUSION

In this work, we propose ExpressEar, a novel FER system that taps into the affordances of commercially available earables in order to effectively capture fine-grained facial muscle movements responsible for various expressions. ExpressEar utilises the IMU sensors, consisting of accelerometer and gyroscope, embedded in the earable to record signal characteristics in response to facial movements. While our results in subject independent settings were not as encouraging, we obtain an average per-user accuracy of 89.9% in classifying all the 32 Facial AUs. Beyond the controlled lab study, we also evaluated the performance of ExpressEar in less constrained and mobile settings. However, the greater promise of this work requires a collective effort by signal processing and hardware design communities to advance the practical applicability of such a system.

[3] https://invensense.tdk.com/products/motion-tracking/6-axis/mpu-6500/
[4] https://www.esense.io/share/eSense-User-Documentation.pdf

Finally, we demonstrate the utility of ExpressEar by illustrating applications based on both voluntary and involuntary use of facial expressions. In the course of our evaluation, we encountered certain limitations of our work, some of which can be addressed with minor tweaks. However, for major concerns like the effect of audio playback, we propose further evaluation in the form of appropriate studies in the future. Nevertheless, as a first step in the direction of continuous and non-intrusive sensing of fine-grained facial expressions, our results are promising and contribute significantly to the domains of earable computing and facial expression recognition.

ACKNOWLEDGMENTS

This research was supported by the Centre for Design and New Media (a TCS Foundation Initiative supported by Tata Consultancy Services) and the Infosys Centre for Artificial Intelligence at IIIT Delhi.

REFERENCES

[1] 2008. *Emotiv: Mobile EEG Brainwear.* https://www.emotiv.com/
[2] 2014. *MD2K: NIH Center of Excellence on Mobile Sensor Data-to-Knowledge.* https://md2k.org/
[3] 2015. *MUSE: The Brain Sensing Headband.* http://www.choosemuse.com/
[4] 2016. *PIP: The Stress Management Device.* https://thepip.com/
[5] 2018. *2018 IEEE International Conference on Advanced Manufacturing (ICAM).* IEEE. https://ieeexplore.ieee.org/servlet/opac?punumber=8606319 OCLC: 1125884622.
[6] 2018. *eSense: Open Earable Platform for Human Sensing.* https://www.esense.io/
[7] 2020. *AirPods Technical Description.* https://www.apple.com/airpods-2nd-generation/specs/
[8] 2020. *Bose SoundSport Firmware.* https://github.com/bosefirmware/ced
[9] 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 - ISWC '19*. ACM Press, London, United Kingdom, 1–9. https://doi.org/10.1145/3341163.3347747
[10] 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* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 679–689. https://doi.org/10.1145/3126594.3126649
[11] Brian S. Armour, Elizabeth A. Courtney-Long, Michael H. Fox, Heidi Fredine, and Anthony Cahill. 2016. Prevalence and Causes of Paralysis—United States, 2013. *American Journal of Public Health* 106, 10 (Oct. 2016), 1855–1857. https://doi.org/10.2105/AJPH.2016.303270
[12] Karissa C. Arthur, Andrea Calvo, T. Ryan Price, Joshua T. Geiger, Adriano Chiò, and Bryan J. Traynor. 2016. Projected increase in amyotrophic lateral sclerosis from 2015 to 2040. *Nature Communications* 7, 1 (Nov. 2016), 12408. https://doi.org/10.1038/ncomms12408
[13] Daniel L. Ashbrook. 2010. *Enabling Mobile Microinteractions.* Ph.D. Dissertation. USA. Advisor(s) Starner, Thad E. AAI3414437.
[14] Sara Ashry, Tetsuji Ogawa, and Walid Gomaa. 2020. CHARM-Deep: Continuous Human Activity Recognition Model Based on Deep Neural Network Using IMU Sensors of Smartwatch. *IEEE Sensors Journal* 20, 15 (Aug. 2020), 8757–8770. https://doi.org/10.1109/JSEN.2020.2985374
[15] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv:1803.01271 [cs.LG]
[16] Ling Bao and Stephen S. Intille. 2004. Activity Recognition from User-Annotated Acceleration Data. In *Pervasive Computing*, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Alois Ferscha, and Friedemann Mattern (Eds.). Vol. 3001. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–17. https://doi.org/10.1007/978-3-540-24646-6_1 Series Title: Lecture Notes in Computer Science.
[17] Abdelkareem Bedri, David Byrd, Peter Presti, Himanshu Sahni, Zehua Gue, and Thad Starner. 2015. Stick it in your ear: building an in-ear jaw movement sensor. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers - UbiComp '15*. ACM Press, Osaka, Japan, 1333–1338. https://doi.org/10.1145/2800835.2807933
[18] Abdelkareem Bedri, Diana Li, Rushil Khurana, Kunal Bhuwalka, and Mayank Goel. 2020. FitByte: Automatic Diet Monitoring in Unconstrained Situations Using Multimodal Sensing on Eyeglasses. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376869
[19] Henry S. Brenman, Robert C. Mackowiak, and M.H.F. Friedman. 1968. Condylar Displacement Recordings as an Analog of Mandibular Movements. *Journal of Dental Research* 47, 4 (July 1968), 599–602. https://doi.org/10.1177/00220345680470041501

[20] Nancy J. Briton and Judith A. Hall. 1995. Gender-based expectancies and observer judgments of smiling. *Journal of Nonverbal Behavior* 19, 1 (March 1995), 49–65. https://doi.org/10.1007/BF02173412

[21] Mark S. Chapell. 1997. Frequency of Public Smiling across the Life Span. *Perceptual and Motor Skills* 85, 3_suppl (Dec. 1997), 1326–1326. https://doi.org/10.2466/pms.1997.85.3f.1326

[22] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[23] 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*. ACM, Virtual Event USA, 112–125. https://doi.org/10.1145/3379337.3415879

[24] Francois Chollet et al. 2015. *Keras*. https://github.com/fchollet/keras

[25] Erik J. Coats and Robert S. Feldman. 1996. Gender Differences in Nonverbal Correlates of Social Status. *Personality and Social Psychology Bulletin* 22, 10 (Oct. 1996), 1014–1022. https://doi.org/10.1177/01461672962210004

[26] Charles Darwin. 2013. *The Expression of the Emotions in Man and Animals*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139833813

[27] J. del R. Millan, F. Renkens, J. Mourino, and W. Gerstner. 2004. Noninvasive Brain-Actuated Control of a Mobile Robot by Human EEG. *IEEE Transactions on Biomedical Engineering* 51, 6 (June 2004), 1026–1033. https://doi.org/10.1109/TBME.2004.827086

[28] S. Du, Y. Tao, and A. M. Martinez. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* 111, 15 (April 2014), E1454–E1462. https://doi.org/10.1073/pnas.1322355111

[29] Paul Ekman and Wallace V Friesen. 1978. *Manual for the facial action coding system*. Consulting Psychologists Press.

[30] Paul Ekman and Erika L. Rosenberg (Eds.). 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, New York, NY, US. Pages: xvi, 495.

[31] Paul Ekman and Erika L. Rosenberg. 2005. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

[32] B. Fasel and Juergen Luettin. 2003. Automatic facial expression analysis: a survey. *Pattern Recognition* 36, 1 (Jan. 2003), 259–275. https://doi.org/10.1016/S0031-3203(02)00052-3

[33] Alan J. Fridlund. 1997. The new ethology of human facial expressions. In *The Psychology of Facial Expression* (1 ed.), James A. Russell and José Miguel Fernández-Dols (Eds.). Cambridge University Press, 103–130. https://doi.org/10.1017/CBO9780511659911.007

[34] Maurizio Garbarino, Matteo Lai, Simone Tognetti, Rosalind Picard, and Daniel Bender. 2014. Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare - "Transforming healthcare through innovations in mobile and wireless technologies"*. ICST, Athens, Greece. https://doi.org/10.4108/icst.mobihealth.2014.257418

[35] Craig W. Goodmurphy and William K. Ovalle. 1999. Morphological study of two human facial muscles: Orbicularis oculi and corrugator supercilii. *Clinical Anatomy* 12, 1 (1999), 1–11. https://doi.org/10.1002/(SICI)1098-2353(1999)12:1<1::AID-CA1>3.0.CO;2-J

[36] Anna Gruebler and Kenji Suzuki. 2014. Design of a Wearable Device for Reading Positive Expressions from Facial EMG Signals. *IEEE Transactions on Affective Computing* 5, 3 (July 2014), 227–237. https://doi.org/10.1109/TAFFC.2014.2313557

[37] Jinni A. Harrigan, Robert Rosenthal, and Klaus R. Scherer (Eds.). 2005. *The new handbook of methods in nonverbal behavior research*. Oxford University Press, New York.

[38] Steven Hickson, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan Essa. 2017. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. *arXiv:1707.07204 [cs]* (July 2017). http://arxiv.org/abs/1707.07204 arXiv: 1707.07204.

[39] Susan P. Holmes and Dan Gusfield. 1999. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. *J. Amer. Statist. Assoc.* 94, 447 (Sept. 1999), 989. https://doi.org/10.2307/2670026

[40] Tahera Hossain, Md Shafiqul Islam, Md Atiqur Rahman Ahad, and Sozo Inoue. 2019. Human activity recognition using earable device. In *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers - UbiComp/ISWC '19*. ACM Press, London, United Kingdom, 81–84. https://doi.org/10.1145/3341162.3343822

[41] Andrey Ignatov. 2018. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Applied Soft Computing* 62 (Jan. 2018), 915–922. https://doi.org/10.1016/j.asoc.2017.09.027

[42] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019. Interferi: Gesture Sensing using On-Body Acoustic Interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300506

[43] Shun Ishii, Kizito Nkurikiyeyezu, Mika Luimula, Anna Yokokubo, and Guillaume Lopez. 2020. ExerSense: Real-Time Physical Exercise Segmentation, Classification, and Counting Algorithm Using an IMU Sensor. *Smart Innovation* (Aug. 2020).

[44] Dean Karantonis, Michael Narayanan, Merryn Mathie, Nigel Lovell, and B.G. Celler. 2006. Implementation of a Real-Time Human Movement Classifier Using a Triaxial Accelerometer for Ambulatory Monitoring. *Information Technology in Biomedicine, IEEE Transactions on* 10 (Feb. 2006), 156–167. https://doi.org/10.1109/TITB.2005.856864

[45] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Allesandro Montanari. 2018. Earables for Personal-Scale Behavior Analytics. *IEEE Pervasive Computing* 17, 3 (July 2018), 83–89. https://doi.org/10.1109/MPRV.2018.03367740

[46] Connor Tom Keating and Jennifer Louise Cook. 2020. Facial Expression Production and Recognition in Autism Spectrum Disorders. *Child and Adolescent Psychiatric Clinics of North America* 29, 3 (July 2020), 557–571. https://doi.org/10.1016/j.chc.2020.02.006

[47] Cinita Kupperbusch, David Matsumoto, Kristie Kooken, Sherry Loewinger, Hideko Uchida, Carinda Wilson-Cohn, and Nathan Yrizarry. 1999. Cultural influences on nonverbal expressions of emotion. In *The social context of nonverbal behavior*. Editions de la Maison des Sciences de l'Homme, Paris, France, 17–44.

[48] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter* 12, 2 (March 2011), 74–82. https://doi.org/10.1145/1964897.1964918

[49] Jangho Kwon, Da-Hye Kim, Wanjoo Park, and Laehyun Kim. 2016. A wearable device for emotional recognition using facial expression and physiological response. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Orlando, FL, USA, 5765–5768. https://doi.org/10.1109/EMBC.2016.7592037

[50] Seungchul Lee, Chulhong Min, Alessandro Montanari, Akhil Mathur, Youngjae Chang, Junehwa Song, and Fahim Kawsar. 2019. Automatic Smile and Frown Recognition with Kinetic Earables. In *Proceedings of the 10th Augmented Human International Conference 2019 on - AH2019*. ACM Press, Reims, France, 1–4. https://doi.org/10.1145/3311823.3311869

[51] Hong Lu, Jun Yang, Zhigang Liu, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. 2010. The Jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems - SenSys '10*. ACM Press, Zürich, Switzerland, 71. https://doi.org/10.1145/1869983.1869992

[52] 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 - IUI '16*. ACM Press, Sonoma, California, USA, 317–326. https://doi.org/10.1145/2856767.2856770

[53] Denys J. C. 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*. ACM, Denver Colorado USA, 1911–1922. https://doi.org/10.1145/3025453.3025692

[54] R. McAlister. 1998. An ultrasound investigation of the lip levator musculature. *The European Journal of Orthodontics* 20, 6 (Dec. 1998), 713–720. https://doi.org/10.1093/ejo/20.6.713

[55] Chulhong Min, Akhil Mathur, and Fahim Kawsar. 2018. Audio-Kinetic Model for Automatic Dietary Monitoring with Earable Devices. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, Munich Germany, 517–517. https://doi.org/10.1145/3210240.3210810

[56] MindMaze. 2017. *Mask: Synchronizing facial expression across the physical and virtual worlds*. https://www.mindmaze.com/labs/mask

[57] Nitin Nair, Chinchu Thomas, and Dinesh Babu Jayagopi. 2018. Human Activity Recognition Using Temporal Convolutional Network. In *Proceedings of the 5th international Workshop on Sensor-based Activity Recognition and Interaction*. ACM, Berlin Germany, 1–8. https://doi.org/10.1145/3266157.3266221

[58] Sashank Narain, Triet D. Vo-Huu, Kenneth Block, and Guevara Noubir. 2016. Inferring User Routes and Locations Using Zero-Permission Mobile Sensors. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Jose, CA, 397–413. https://doi.org/10.1109/SP.2016.31

[59] Rui Ning, Cong Wang, ChunSheng Xin, Jiang Li, and Hongyi Wu. 2018. DeepMag: Sniffing Mobile Apps in Magnetic Field through Deep Convolutional Neural Networks. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, Athens, 1–10. https://doi.org/10.1109/PERCOM.2018.8444573

[60] Jun-geun Park, Ami Patel, Dorothy Curtis, Seth Teller, and Jonathan Ledlie. 2012. Online pose classification and walking speed estimation using handheld devices. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*. ACM Press, Pittsburgh, Pennsylvania, 113. https://doi.org/10.1145/2370216.2370235

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[62] Joel E. Pessa, Vikram P. Zadoo, Earle K. Adrian, Cheng H. Yuan, Jason Aydelotte, and Jaime R. Garza. 1998. Variability of the Midfacial Muscles: Analysis of 50 Hemifacial Cadaver Dissections:. *Plastic and Reconstructive Surgery* 102, 6 (Nov. 1998), 1888–1893. https://doi.org/10.1097/00006534-199811000-00013

[63] Joel E. Pessa, Vikram P. Zadoo, Peter A. Garza, Erle K. Adrian, Adriane I. Dewitt, and Jaime R. Garza. 1998. Double or bifid zygomaticus major muscle: Anatomy, incidence, and clinical correlation. *Clinical Anatomy* 11, 5 (1998), 310–313. https://doi.org/10.1002/(SICI)1098-2353(1998)11:5<310::AID-CA3>3.0.CO;2-T

[64] Paulo Gurgel Pinheiro, Claudio Gurgel Pinheiro, and Eleri Cardozo. 2017. The Wheelie — A facial expression controlled wheelchair using 3D technology. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE,

Lisbon, 271–276. https://doi.org/10.1109/ROMAN.2017.8172313

[65] Jay Prakash, Zhijian Yang, Yu-Lin Wei, and Romit Roy Choudhury. 2019. STEAR: Robust Step Counting from Earables. In *Proceedings of the 1st International Workshop on Earable Computing*. ACM, London United Kingdom, 36–41. https://doi.org/10.1145/3345615.3361133

[66] Sharmin Akther Purabi, Rayhan Rashed, Mirajul Islam, Nahiyan Uddin, Mahmuda Naznin, and A. B. M. Alim Al Islam. 2019. As you are, so shall you move your head: a system-level analysis between head movements and corresponding traits and emotions. In *Proceedings of the 6th International Conference on Networking, Systems and Security - NSysS '19*. ACM Press, Dhaka, Bangladesh, 3–11. https://doi.org/10.1145/3362966.3362985

[67] Philippe Remy. 2020. Temporal Convolutional Networks for Keras. https://github.com/philipperemy/keras-tcn.

[68] Soha Rostaminia, Alexander Lamson, Subhransu Maji, Tauhidur Rahman, and Deepak Ganesan. 2019. W!NCE: Unobtrusive Sensing of Upper Facial Action Units with EOG-based Eyewear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (March 2019), 1–26. https://doi.org/10.1145/3314410

[69] Tobias Röddiger, Daniel Wolffram, David Laubenstein, Matthias Budde, and Michael Beigl. 2019. Towards Respiration Rate Monitoring Using an In-Ear Headphone Inertial Measurement Unit. In *Proceedings of the 1st International Workshop on Earable Computing*. ACM, London United Kingdom, 48–53. https://doi.org/10.1145/3345615.3361130

[70] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2015. Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 6 (June 2015), 1113–1133. https://doi.org/10.1109/TPAMI.2014.2366127

[71] Jocelyn Scheirer, Raul Fernandez, and Rosalind W. Picard. 1999. Expression glasses: a wearable device for facial expression recognition. In *CHI '99 extended abstracts on Human factors in computing systems - CHI '99*. ACM Press, Pittsburgh, Pennsylvania, 262. https://doi.org/10.1145/632716.632878

[72] Karen L. Schmidt and Jeffrey F. Cohn. 2001. Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American Journal of Physical Anthropology* 116, S33 (2001), 3–24. https://doi.org/10.1002/ajpa.20001

[73] Mike Seymour. 2019. FACS at 40: facial action coding system panel. In *ACM SIGGRAPH 2019 Panels on - SIGGRAPH '19*. ACM Press, Los Angeles, California, 1–2. https://doi.org/10.1145/3306212.3328132

[74] Yilei Shi, Haimo Zhang, Kaixing Zhao, Jiashuo Cao, Mengmeng Sun, and Suranga Nanayakkara. 2020. Ready, Steady, Touch!: Sensing Physical Contact with a Finger-Mounted IMU. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (June 2020), 1–25. https://doi.org/10.1145/3397309

[75] Ronald E. Shor. 1978. The Production and Judgment of Smile Magnitude. *The Journal of General Psychology* 98, 1 (Jan. 1978), 79–96. https://doi.org/10.1080/00221309.1978.9920859

[76] Timothy Sohn, Alex Varshavsky, Anthony LaMarca, Mike Y. Chen, Tanzeem Choudhury, Ian Smith, Sunny Consolvo, Jeffrey Hightower, William G. Griswold, and Eyal de Lara. 2006. Mobility Detection Using Everyday GSM Traces. In *UbiComp 2006: Ubiquitous Computing*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Paul Dourish, and Adrian Friday (Eds.). Vol. 4206. Springer Berlin Heidelberg, Berlin, Heidelberg, 212–224. https://doi.org/10.1007/11853565_13 Series Title: Lecture Notes in Computer Science.

[77] Yale Song, Daniel McDuff, Deepak Vasisht, and Ashish Kapoor. 2015. Exploiting sparsity and co-occurrence structure for action unit recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, Ljubljana, 1–8. https://doi.org/10.1109/FG.2015.7163081

[78] Weslllen Sousa Lima, Eduardo Souto, Khalil El-Khatib, Roozbeh Jalali, and Joao Gama. 2019. Human Activity Recognition Using Inertial Sensors in a Smartphone: An Overview. *Sensors* 19, 14 (July 2019), 3213. https://doi.org/10.3390/s19143213

[79] Kazuhiro Taniguchi, Hisashi Kondo, Mami Kurosawa, and Atsushi Nishikawa. 2018. Earable TEMPO: A Novel, Hands-Free Input Device that Uses the Movement of the Tongue Measured with a Wearable Ear Sensor. *Sensors* 18, 3 (March 2018), 733. https://doi.org/10.3390/s18030733

[80] Hoang Minh Thang, Vo Quang Viet, Nguyen Dinh Thuc, and Deokjai Choi. 2012. Gait identification using accelerometer on mobile phone. In *2012 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, Saigon, Vietnam, 344–348. https://doi.org/10.1109/ICCAIS.2012.6466615

[81] C.H.J. Tzou, P. Giovanoli, M. Ploner, and M. Frey. 2005. Are there ethnic differences of facial movements between Europeans and Asians? *British Journal of Plastic Surgery* 58, 2 (March 2005), 183–195. https://doi.org/10.1016/j.bjps.2004.10.014

[82] Jacob Whitehill, Marian Bartlett, and Javier Movellan. 2008. Automatic facial expression recognition for intelligent tutoring systems. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, 1–6. Backup Publisher: IEEE.

[83] Erik Wästlund, Kay Sponseller, and Ola Pettersson. 2010. What you see is where you go: testing a gaze-driven power wheelchair for individuals with severe multiple disabilities. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10*. ACM Press, Austin, Texas, 133. https://doi.org/10.1145/1743666.1743699

[84] Xuhai Xu, Haitian Shi, Xin Yi, WenJia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K. Dey. 2020. EarBuddy: Enabling On-Face Interaction via Wireless Earbuds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376836

[85] Yukang Yan, Chun Yu, Wengrui Zheng, Ruining Tang, Xuhai Xu, and Yuanchun Shi. 2020. FrownOnError: Interrupting Responses from Smart Speakers by Facial Expressions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376810

[86] Yan Tong, Jixu Chen, and Qiang Ji. 2010. A Unified Probabilistic Framework for Spontaneous Facial Action Modeling and Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 2 (Feb. 2010), 258–273. https://doi.org/10.1109/TPAMI.2008.293