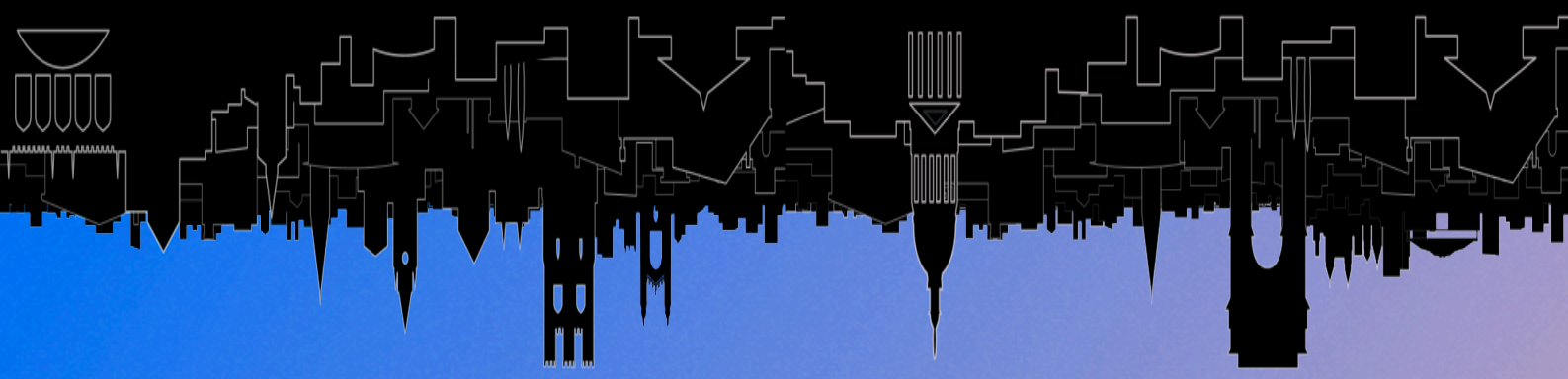


EarComp 2022

3rd International Workshop on Earable Computing

In conjunction with UbiComp 2022

September 15th, 2022, Cambridge, United Kingdom

Excerpt of PPGface: Like What You Are Watching? Earphones Can "Feel" Your Facial Expressions

Seokmin Choi
University at Buffalo, SUNY
Buffalo, NY, USA

Yang Gao
Northwestern University
Evanston, IL, USA

Yincheng Jin
University at Buffalo, SUNY
Buffalo, NY, USA

Se jun Kim
University at Buffalo, SUNY
Buffalo, NY, USA

Jiyang Li
University at Buffalo, SUNY
Buffalo, NY, USA

Wenyao Xu
University at Buffalo, SUNY
Buffalo, NY, USA

Zhanpeng Jin*
University at Buffalo, SUNY
Buffalo, NY, USA

ABSTRACT

Facial expression recognition has been widely explored to demonstrate people's emotional states. However, existing systems primarily rely on external devices which seems less accessible and efficient. To this end, we propose PPGface, a ubiquitous facial expression recognition platform that leverages earable devices with built-in PPG sensor. PPGface understands the facial expressions through the dynamic PPG patterns resulting from facial muscle movements. Through several comprehensive studies, this work validates a great potential to be employed in future commodity earable devices.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile devices**; **Human computer interaction (HCI)**.

KEYWORDS

Photoplethysmogram, PPG, Facial Expression, Ear Canal, Blood Vessel Deformation

ACM Reference Format:
Seokmin Choi, Yang Gao, Yincheng Jin, Se jun Kim, Jiyang Li, Wenyao Xu, and Zhanpeng Jin. 2022. Excerpt of PPGface: Like What You Are Watching? Earphones Can "Feel" Your Facial Expressions. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2022 ACM International Symposium on Wearable Computers (UbiComp-ISWC'22 Adjunct).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3544793.3563418

1 INTRODUCTION

As more viewers are moving from traditional pay-TV to streaming market trend, many media and service providers have explored different approaches to capture users' emotional reactions on the fly through video watching habits [2]. Facial expressions have been deemed as the universal non-verbal language to express internal emotional states, which are also accompanied by unconscious body postures. Although there are few approaches to understand facial expressions, they either raise privacy issues (computer vision-based) or require external hardware setup close to the target user (WiFi-based). To enable a more privacy-preserving and ubiquitous system, we propose PPGface, which leverages in-ear PPG signals with complementary accelerometer data. Specifically, PPGface understands different facial expressions through dynamic PPG patterns which are altered by the facial muscle motions where the signals are collected inside the ear.

2 RELATED WORK

Computer vision (CV) based Facial expression recognition (FER) system is one of the most intuitive way to recognize different facial expressions. Zeng *et al.* [5] proposed the deep sparse autoencoders (DSAE) which extracts meaningful features in an unsupervised manner.

Besides the CV-based approach, researchers employed different sensing modalities to build FER system. For example, Chen *et al.*[1] introduced a WiFi based facial expression recognition system to classify six different emotions by using a laptop and three antennas. In addition, Gruebler *et al.* [3] proposes a wearable device embedded with an EMG sensor and shows that the device can continuously track smiling and frowning. Although some perform relatively good, it is unrealistic to setup extra hardware devices while tracking facial expressions or impractical to attach multiple sensors to the face in daily lives.

Researchers have employed PPG signals to measure cardiac activities such as heart rate or blood pressure. Due to the fact that PPG signals are also vulnerable to motion artifacts, previous studies also focused on removing those motion artifacts. Recently, unlike previous research objectives, studies have started to re-purpose the PPG motion artifacts instead of regarding them as noises. Zhao *et al.* [6] proposed a gesture recognition system using PPG signals and motion sensors embedded in a wrist worn device that can discriminate finger movements.

*This is the corresponding author (zjin@buffalo.edu).

UbiComp-ISWC'22 Adjunct, September 11-15, 2022, Atlanta, USA and Cambridge, UK,

ACM ISBN 978-1-4503-8461-2/22/09…$15.00
https://doi.org/10.1145/3544793.3563418

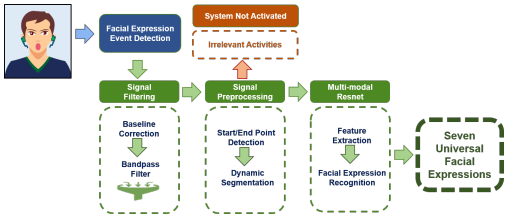


Figure 1: PPGface system overview

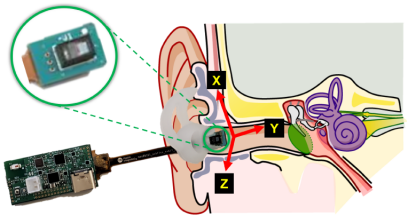


Figure 2: The depiction of the sensor

3 PPGFACE SYSTEM DESIGN

As shown in Fig. 1, our proposed FER system, PPGface, is to understand the user's different facial expressions by fusing both the behavioral (IMU signals from spontaneous body posture) and physiological (PPG signals from the facial muscle movement) patterns.

First, the device measures the in-ear PPG and accelerometer signals when user makes an expression. The measured signals show distinctive patterns between the resting stage and when the user makes a facial expression. In addition, if the user performs other physical activities which are irrelevant to facial expressions, such as yawning or swallowing saliva, these unrelated signals are detected and removed for the further steps.

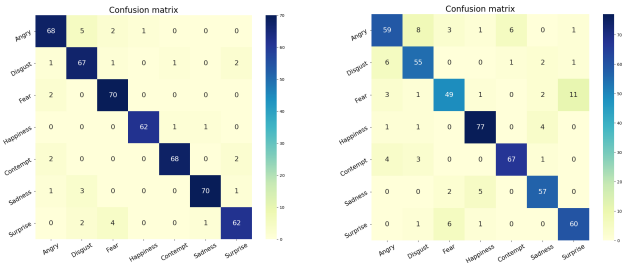
After the event detection step, we apply detrend technique to remove the DC components caused by respiration followed by the bandpass filter, which filters out both the high and low frequency components to preserve the characteristics of both the facial expressions and related spontaneous body postures.

The dynamic signal preprocessing step involves two sub-stages: segmentation and detection. First, the peak detection algorithm and continuous wavelet transform (CWT) technique are used to select possible candidate sets of start and end points of the PPGface signals. By analyzing these signals, the system is able to detect the facial expression activity events, given the final start and end points of each segment corresponding to the facial expression activities.

At last, preprocessed signals will be the inputs to the classification module. To overcome the limited amount of data, we apply the following data augmentation methods: Rotation, time-warping, and scaling techniques. Then, multimodal ResNet is used to extract the representative features and classify seven different facial expressions. The user's profile is generated to classify seven facial expressions during the training stage. During the testing, PPGface goes through feature profiling to classify any unseen data based on the pre-train model.

4 EXPERIMENTAL SETUP AND RESULTS

In this study, we deployed MAXIM86161EVSYS evaluation kit [4] as shown in Fig. 2. The sampling rate of three LDEs (i.e., green, infrared (IR), and red) was set to 128 Hz. Along with the LEDs, the sensor is also equipped with 3-axis accelerometer sensors synchronized with PPG sensor.



(a) User-dependent case **(b) User-independent case**

Figure 3: Confusion matrix of seven universal facial expressions. (a) user-dependent, (b) user-independent case.

We evaluate the user-dependent performance of the PPGface in terms of classification accuracy. PPGface achieves 0.935 accuracy (STD = 0.086), 0.951 precision (STD = 0.086), 0.934 recall (STD = 0.074), and 0.934 f1-score (STD = 0.094) (STD: Standard Deviation). It is observed from Fig. 3a that accuracy of fear and surprise is lower than other expressions, which is because facial muscles to express those expressions overlap a lot.

We wanted to explore more whether PPGface is also able to distinguish well under the user-independent case. In this case, we combined each facial expression without considering the user label. Performance of the PPGface's accuracy is 0.848, precision is 0.854 (STD = 0.048), recall is 0.854 (STD = 0.074), and f1-score is 0.842 (STD = 0.054). Performance shows lower accuracy than the user-dependent case due to a difference among each user when making facial expressions which can be shown from Fig. 3b.

5 CONCLUSION

We introduce PPGface, a novel approach to understand different facial expressions by leveraging unique in-ear PPG variations with the aid of an accelerometer. It is believed that this work has a potential to provide insight into non-intrusive in-ear sensing research. In the future, we plan to design customized prototype, enhance the versatility of PPGface, and examine the permanence and robustness from a larger and more diverse subject pool.

REFERENCES

[1] Yanjiao Chen, Runmin Ou, Zhiyang Li, and Kaishun Wu. 2020. WiFace: facial expression recognition using Wi-Fi signals. *IEEE Transactions on Mobile Computing* (2020).
[2] Michael Goodman. 2021. *U.S. SVOD Forecast (2010 - 2026).* Technical Report. Strategy Analytics, Newton, Massachusetts USA. https://www.strategyanalytics.com/ Last accessed: 2021-08-25.
[3] Anna Gruebler and Kenji Suzuki. 2010. Measurement of distal EMG signals using a wearable device for reading facial expressions. In *Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 4594–4597.
[4] Maxim Integrated. 2021. MAXM86161 Single-Supply Integrated Optical Module for HR and SpO2 Measurement. Data Sheet.
[5] Nianyin Zeng, Hong Zhang, Baoye Song, Weibo Liu, Yurong Li, and Abdullah M Dobaie. 2018. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing* 273 (2018), 643–649.
[6] Tianming Zhao, Jian Liu, Yan Wang, Hongbo Liu, and Yingying Chen. 2019. Towards Low-cost Sign Language Gesture Recognition Leveraging Wearables. *IEEE Transactions on Mobile Computing* 20, 4 (2019), 1685–1701.

A Taxonomy of Noise in Voice Self-reports while Running

Tao Bi
University College London
London, United Kingdom
t.bi@ucl.ac.uk

Temitayo Olugbade
University College London
London, United Kingdom
temitayo.olugbade.13@ucl.ac.uk

Akhil Mathur
Nokia Bell Labs
Cambridge, United Kingdom
akhil.mathur@nokia-bell-labs.com

Catherine Holloway
University College London
London, United Kingdom
c.holloway@ucl.ac.uk

Aneesha Singh
University College London
London, United Kingdom
aneesha.singh@ucl.ac.uk

Enrico Costanza
University College London
London, United Kingdom
e.costanza@ucl.ac.uk

Nadia Berthouze
University College London
London, United Kingdom
nadia.berthouze@ucl.ac.uk

ABSTRACT

Smart earables offer great opportunities for conducting ubiquitous computing research. This paper shares its reflection on collecting self-reports from runners using the microphone on the smart eSense earbud device. Despite the advantages of the eSense in allowing researchers to collect continuous voice self-reports anytime anywhere, it also captured noise signals from various sources and created challenges in data processing and analysis. The paper presents an initial taxonomy of noise in runners' voice self-reports data via eSense. This is based on a qualitative analysis of voice recordings based on eSense's microphone with 11 runners across 14 in-the-wild running sessions. The paper discusses the details and characteristics of the observed noise, the challenges in achieving good-quality self-reports, and opportunities for extracting useful contextual information. The paper further suggests a noise-categorization API for the eSense or other similar platforms, not only for the purpose of noise-cancellation but also incorporating the mining of contextual information.

CCS CONCEPTS

• Human-centered computing • Human computer interaction (HCI) • Empirical studies in HCI

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09…$15.00
https://doi.org/10.1145/3544793.3563421

KEYWORDS

Taxonomy; Voice; Self-reports; Runners; Running; In-the-wild Experience Sampling; Smart earbuds; Earables; eSense;

ACM Reference format:

Tao Bi, Temitayo Olugbade, Akhil Mathur, Catherine Holloway, Aneesha Singh, Enrico Costanza, Nadia Berthouze. 2022. A Taxonomy of Noise in Voice Self-reports while Running. *In Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11--15, 2022, Cambridge, United Kingdom, 5 pages.* https://doi.org/10.1145/3544793.3563421

1 INTRODUCTION

Smart earables offer great opportunities for experience sampling and collection of user self-reports [1] in ubiquitous contexts. Traditional experience sampling methods (ESM) that use mobile phone based applications require people to physically interact with a phone multiple times [2]. This is not very practical when collecting self-reports from runners while running because runners use their hands and arms to maintain balance and movement flow. Smart earbuds can offer a hands-free experience for people to self-report via voice recordings [3]. Runners receive ESM prompts via earbud speakers and use speech to self-report back to the earbud microphone, which creates less physical or biomechanical interference with the runner's body movement [4]. Also, earbuds are lightweight and widely accepted by many runners to consume music during running.

In our study, eSense was used to deliver an ESM schedule and to record verbal self-reports of feelings of runners at run time

in the wild. eSense is a multi-sensory earable platform that is widely used in the HCI research community [5, 6] for collecting audio recordings via its embedded microphone. Overall, 11 runners (five males, and six females) had a total of 14 running sessions. Nine of the runners completed only one session, one runner completed two sessions, and another runner completed three sessions. The duration of running sessions (and eSense data recording) averaged 34.7 ± 15.1 minutes. The researcher guided participants during the 5-min trail session. The study did not control the running environment, runner ability, or runner performance. Participants were given with flexibility to run anywhere at any self-selected pace. For more details on the data collections and other aspects of the work see [3, 4].

Despite the earbuds offering a more ubiquitous and less intrusive ESM, we faced challenges in data processing and analysis due to the noise in the captured audio data. To characterize the noise sources and their information value, we used Nvivo [7] to label all potential noise information and then qualitatively analyzed the characteristics of all potential noise information. The noise coding process is mainly based on the first author's listening and interpretation of audio. It also used the author's observation and contextual notes taken during data collection process, i.e.., observation of the runner's apparel and accessories through the study session when participants run indoor or at reachable distance outdoor. The following sections will present an initial taxonomy of the noise observed in the runner voice self-reports.

2 A TAXONOMY OF NOISE IN RUNNERS' VOICE SELF-REPORTS

As shown in Table 1 (see next page), this taxonomy consists of 14 noise categories: Car noise, Traffic noise, Sound of foot strike, Outdoor terrain noise, Earbuds rubbing noise, Breathing sound, Clothes rubbing noise, Wind noise, Personal item vibration noise, Treadmill machine noise, Foot strike and treadmill impact sound, Animal sound, Passenger talking, and Shop's speaker noise.

Each noise category is associated with its relevant running context (e.g., road run, gym treadmill run). For each category, the taxonomy also includes the characteristics and factors that influence the impact of the noise (e.g., pitch, volume, wind direction). It also classifies whether the noise is synchronized with the run, i.e. whether or not the noise is concomitant with the run. In addition, each category is evaluated with respect to its impact on the quality of voice self-reports. For example, a noise category is labelled as high impact if it severely renders the voice recording difficult to be understood and hence transcribed.

The noise category itself as well as the characteristics of the noise can be useful contextual info for understanding the noise experience, e.g., outdoor surface vs treadmill surface enables detection of running context, animal sounds might explain references to animals in self-report, frequency of foot strike can be useful in capturing fatigue or pacing, etc.

Table 1: Taxonomy of noise in runners' voice self-reports

Noise type	Characteristics & Factors	Synchronized with running?	Running context	Impact on speech recognition
Car noise	Scrapping or chirping sound when a car passes; Increasing pitch when the car is approaching; Decreasing pitch when the car is driving away; Horning sound;	No	Road run	High
Traffic noise	Motorbike engine noise; Car engine noise; Police car or emergency car alarm sound	No	Road run Park run	High
Sound of foot strike	High-rhythm sound during a fast run; Low-rhythm sound during a slow run; High volume sound on heavy landing; Low volume sound on light landing; Heel-to-front transition sound	Yes	Road run Gym run	High
Outdoor terrain noise	Road; Running Track; Grass; Trail; Treadmill	Yes	All run	High
Earbuds rubbing noise	sSense earbuds are unstable in intense physical activities. The rubbing noise has a very high frequency.	Yes	All run	High

Table 1: Taxonomy of noise in runners' voice self-reports (cont.)

Noise type	Characteristics & Factors	Synchronized with running?	Running context	Impact on speech recognition
Breathing sound	Inhale and exhale sound different.	Yes	All run	High
Clothes rubbing nose	Same frequency & rhythm as food strike. Chuffing noise from arm, hands, legs	Yes	All run	High
Wind noise	Headwind; downwind; crosswind Wind direction causes different noise levels. Wind direction interacts with voice volume.	No	Outdoor run	High
Personal item vibration noise	Keys; Earrings; Necklaces	Yes	All run	High
Treadmill machine noise	Treadmill belt noise: squeaking, screeching and whining noises; Treadmill motor noise; High frequency & High volume; Other gym machine noise	Yes	Gym run	High
Foot strike and treadmill impact sound	Landing sound volume; Soft or heavy landing; High speed creates high pitched noise from treadmill machine;	Yes	Gym run	High
Animal sound	Birds sound, e.g., seagulls	No	Outdoor run	Low
Passenger talking	Group talk vs individual talk lower frequency voice vs higher frequency voice	No	Road run Gym run	High
Shop's noise	Music; Ads	No	Road run	High

3 DEMONSTRATIONS OF AUDIO AND NOISE SIGNALS IN A RUNNER'S VOICE SELF-REPORTS

The taxonomy above shows that the noise observed had different characteristics and were typically high impact. Here, we further provide visual representations of the raw audio highlighting the noise signals, using the software Audacity [8], based on the data captured from one of the running sessions introduced in Section 2 as a case study. These visual representations aim to 1) illustrate not only how different noise categories and their characteristics impact the raw audio signals, and 2) illustrate how raw audio and noise signals could infer meaningful information.

Figure 1 shows the full voice recording for the running session. In this recording, all sounds (voice and noise signals) overlapped, which makes it difficult to recognize a runner's speech. Figure 2 shows an excerpt (approximately 3 minutes) of the full voice recording in Figure 1. In Figure 2, we distinguished between noise signals and speech. Voice signals when the runner was talking (segment C in the figure) occur throughout the excerpt. Meanwhile, there are aural occlusions at several points, e.g., from noise of one foot (segments A in the figure) or other (segments B in the figure) landing. The feet strike noise (segments A and B) overlaps with the voice signal (segment C). Further, the noise (segments A and B) are of higher volume than the voice (segment C), which makes it challenging to extract the voice content from the overlapped area.

Beyond the negative implications of the noise, we deduced potentially valuable information from them. For instance, as can be seen in Figure 2, segment A always has higher volume than segment B, which suggests that this runner may have a higher landing impact in one foot than in the other.

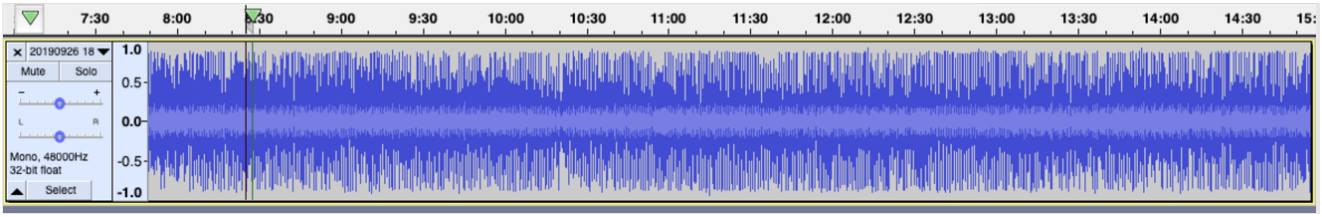


Figure 1: An overview of a runner's voice self-report while running.

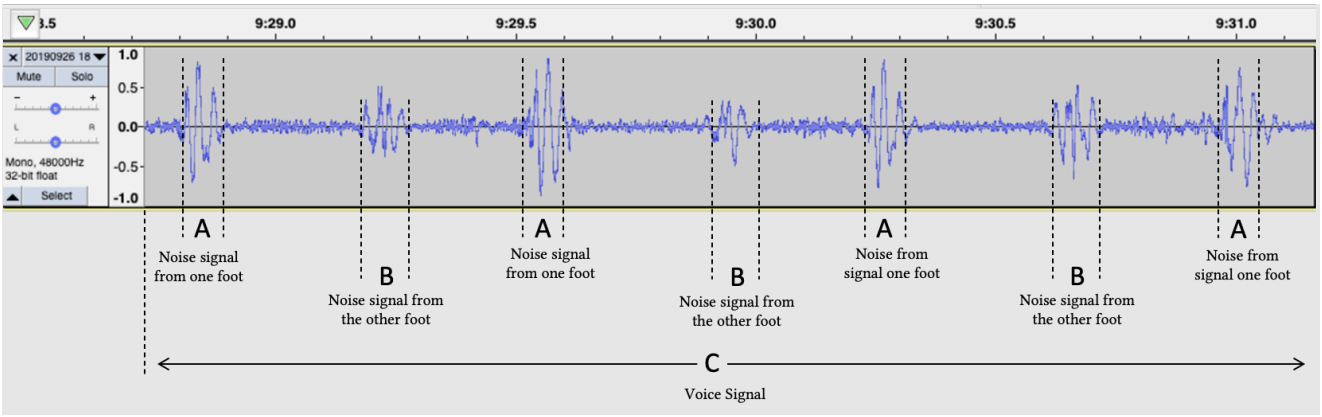


Figure 2: An excerpt of a runner's voice self-report.

3 DISCUSSIONS

As shown in Table 1, noise in runner voice self-report is complex but rich in information. Although we applied simple noise reduction techniques offered by Audacity [9], the results were not satisfactory enough to produce a good-quality voice recording. This is due to the irregular pattern of some of the noise such as wind noise, traffic noise, and noise from road surfaces with variations. In addition, the volumes of step and strike noise were much higher than the runner's voice. The noise from the earbuds' rubbing was even stronger as it was the closest to the microphone. Such factors and characteristics make it challenging to apply existing noise reduction techniques. While our analysis and discussion is limited in that we did not review how existing noise cancellation technologies may address the challenges that the observed noise and their characteristics pose, our findings outline critical issues that arise in running scenarios, some of these issues may be unique to ESM in running and may not be observed in other use cases (e.g. listening to music during outdoor walking, private phone call in a public space) that noise cancellation technologies will typically target.

The proposed taxonomy sheds a light on what type of noise existed in the running context, and the summary of features of each noise category can be useful to guide development of noise reduction API (or real-time noise cancellation software) that address the specific scenario of running ESM or ESM in the wild in general. For instance, as highlighted in Table 1, headwind, downwind, or crosswind, not only due to the natural wind direction but also the running direction of a runner themselves, degrades the quality of voice signal as wind vibrates air particles that the voice vibrates [10]. Further, while crosswind direction could additionally alter the direction of voice transmission to one side; headwind could collide with the voice, more than tailwind. Noise cancellation and voice augmentation techniques could better accommodate such factors, to help researchers generate a better-quality voice recording via eSense or other similar platforms.

However, noise can have value, particularly in providing contextual information and personal affective experience. For instance, wind noise can provide information useful for contextualizing the runner's experience self-report. Headwind noise signals can be different from downwind noise, whereas headwind can contribute to making a run extremely difficult. In the transcribed voice self-reports, a runner for example said "fight the headwind" which could indicate that they were struggling or putting much higher effort into the moment. On the contrary, a runner "feels a bit easier now, with the downwind". Here is another example, a runner said, "I am cold,

I'm going downhill so the wind catches me, I'm sure it will get better shortly". Such wind direction can be a factor that makes a runner perceive the run as difficult. In such a situation, runners might benefit from some sort of digital cheering that can be delivered via earbuds. On another hand, headwind can also be a feel-good factor in the run. For instance, a runner referred to wind as "nice to run into a breeze on a hot day". When the temperature is high, but wind is gentle and cool, this could be a moment in which a runner mentally enjoys the run. Therefore, wind noise can not only serve as a contextual measurement but also a measurement for personal affective experience. Similarly, car noise can infer traffic and be used as a safety measurement. Digital reminders could be sent to runners to be aware of traffic if they are running on the road. Foot strike noise even has more information that can be used to detect a runner's running performance such as cadence and stride type (heel stride, middle stride, front stride). It could also be used to infer a runner's mental state again. For example, fatigue could make a runner switch to more heel strikes, which is associated with a higher impact on the ground [11].

5 CONCLUSIONS

This paper reports a preliminary taxonomy of noise observed in voice recordings from runners during verbal self-report, based on a qualitative analysis of the audio recordings. The taxonomy includes the noise types, their characteristics and factors, whether the noise type is synchronized with running or not, and how bad the noise affects the speech recognition quality. Our findings highlight need for noise investigation before audio data collection in the wild. Our taxonomy further highlights that noise can be both negative and positive factors. While noise clearly undermines extraction of the desired self-report of experience, it has the potential to provide rich contextual information for deeper understanding of the experience. eSense or other similar smart earbuds platforms will be more valuable when equipped with modules (e.g., API) capable of noise reduction, noise categorization, and extraction of contextual information from non-verbal aural signals. The research in this paper is still in early stages and the aim of this paper is to prompt discussion within/across relevant areas (human-computer interaction, audio engineering, machine learning).

ACKNOWLEDGMENTS

Tao Bi was supported by a studentship deriving from a grant awarded by the London Legacy Development Corporation (C02955) to the Global Disability Innovation Hub, University College London. The work was also supported by the EU Future and Emerging Technologies Proactive Programme H2020 (Grant No. 824160: EnTimeMent - https://entimement.dibris.unige.it/). The eSense earbuds used in this work were supported by Nokia Bell Lab, Cambridge, UK.

REFERENCES

[1] K. Doherty and G. Doherty, "The construal of experience in HCI: Understanding self-reports," *International Journal of Human-Computer Studies,* vol. 110, pp. 63-74, 2018.

[2] N. Van Berkel, D. Ferreira, and V. Kostakos, "The experience sampling method on mobile devices," *ACM Computing Surveys (CSUR),* vol. 50, no. 6, pp. 1-40, 2017.

[3] N. B.-B. Tao Bi, Enrico Costanza, Aneesha Singh, "Designing Voice-based Runner Experience Sampling Methods to Build Wearable Runners' Experience Recognition System.," 2022.

[4] T. Bi *et al.*, "Towards Chatbot-Supported Self-Reporting for Increased Reliability and Richness of Ground Truth for Automatic Pain Recognition: Reflections on Long-Distance Runners and People with Chronic Pain," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 43-53.

[5] F. Kawsar, C. Min, A. Mathur, A. Montanari, U. G. Acer, and M. Van den Broeck, "eSense: Open Earable Platform for Human Sensing," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018: ACM, pp. 371-372.

[6] eSense. "eSense." https://www.esense.io (accessed 13 Sep, 2022).

[7] Nvivo. "Nvivo." https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home (accessed 13 Sep, 2022).

[8] Audacity. "Audacity." https://www.audacityteam.org (accessed 13 Sep, 2022).

[9] Audacity. "Noise Reduction in Audacity." https://manual.audacityteam.org/man/noise_reduction.html (accessed 13 Sep, 2022).

[10] M. Stephen, "Effect of Wind on Sound Transmission," ed. sciencing.com, 2018.

[11] R. Mann *et al.*, "The effect of shoe type and fatigue on strike index and spatiotemporal parameters of running," vol. 42, no. 1, pp. 91-95, 2015.

enVolve: Are You Listening? Inertial Sensing to Monitor the Involvement of Silent Listeners during an Online Interaction

Garvit Chugh*
chugh.2@iitj.ac.in
Indian Institute of Technology, Jodhpur
India

Suchetana Chakraborty†
suchetana@iitj.ac.in
Indian Institute of Technology, Jodhpur
India

Ravi Bhandari‡
ravibhandari2006@gmail.com
Jodhpur City Knowledge and Innovation Foundation,
Indian Institute of Technology, Jodhpur
India

Sandip Chakraborty§
sandipc@cse.iitkgp.ac.in
Indian Institute of Technology, Kharagpur
India

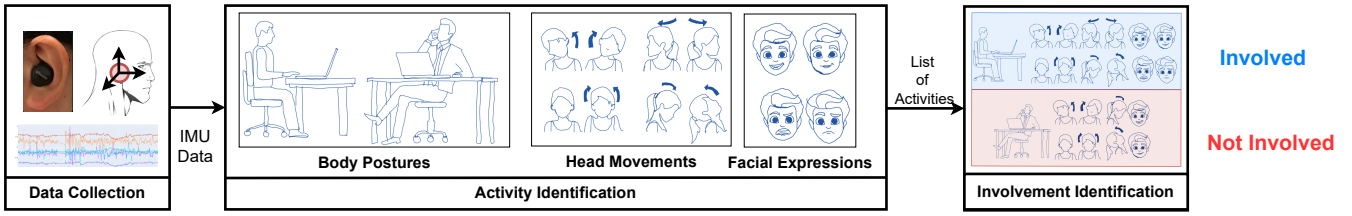


Figure 1: *enVolve* (proposed framework) gathers IMU data using eSense from the listener and recognizes activities during an online class. Then, *enVolve* calculates the level of interest and shows the speaker if the listener is involved or not.

ABSTRACT

Understanding the level of participation for a remote attendee in an online meeting setup could significantly improve the quality of experience for virtual interaction. However, gauging audience involvement over an online meeting becomes particularly challenging when the attendees prefer to turn off the cameras. IMU data have shown promising results in the past to pervasively monitor users' body language, including the determination of various bodily gestures, postures, and facial expressions. This paper demonstrates how earables could help address the stated problem. We provide a motivational study to assess earables for detecting body language corresponding to involved listeners. We further compare it with other sensing modalities like smartwatches and smartphones and accordingly develop a platform called *enVolve*. A lab-scale study with 17 participants demonstrates the efficacy of the proposed system with an average F1 score of more than 80%.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Software and its engineering** → **Development frameworks and environments**; • **Computing methodologies** → **Model development and analysis**.

KEYWORDS

online meeting, earables, inertial sensing

ACM Reference Format:
Garvit Chugh, Suchetana Chakraborty, Ravi Bhandari, and Sandip Chakraborty. 2022. *enVolve*: Are You Listening? Inertial Sensing to Monitor the Involvement of Silent Listeners during an Online Interaction. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3544793.3563419

1 INTRODUCTION

Several verbal and non-verbal cues, such as gaze, gestures, and posture, in a physical meeting setup help the speaker gauge the audience's involvement continuously. However, the recent trends of online teaching-learning, virtual meetings, and webinars have significantly reshaped the core of human interactions, with the participants not sharing a similar spatio-temporal context on a typical online platform. This results in a significant communication gap regarding low cognitive perception between the speaker and the listeners [3, 15]. This gap is often aggravated due to limited bandwidth and poor network connectivity, thus compelling the users to turn off their webcams while attending an online meeting/lecture.

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09…$15.00
https://doi.org/10.1145/3544793.3563419

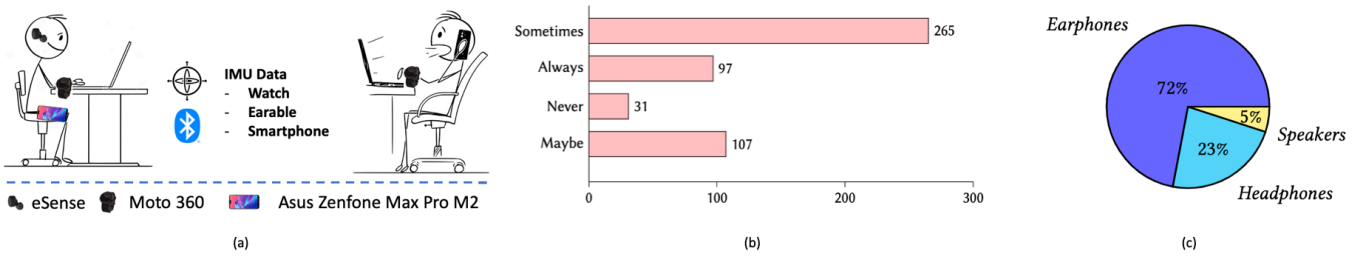


Figure 2: (a) IMU data collection using earable, smartwatch, and smartphone. (b) Survey results of listeners performing movements silently in an online conversation. (c) Survey results of listeners in terms of usage of ear-wears

Moreover, in an online multi-party setup, the audience generally mutes themselves to avoid ambient noise, echo, and interference with the speaker. This limits the usability of audio and video-based approaches for estimating user involvement simply because of the absence of conventional cues such as human gaze, hand gestures, or body posture of the audience. Thus, assessing the listeners' involvement (whether the listener is following or able to follow the talk) in an online meeting setup is particularly challenging for the speaker. While conventionally, based on the visual cues, the speaker could alter the content delivery style, it is difficult to do so in an online setup where the cues are absent. Consequently, the conversation may turn down to be dull and monotonous. Hence, this demands the development of a mechanism that provides real-time feedback to the speaker regarding listeners' involvement in a visually intuitive fashion, thus making online-based content delivery more fruitful to both the speaker and the listeners. In this context, our paper explores using IMU (Inertial Measurement Unit) data obtained from earables to detect different cues generated by the human head and body manifestations. In recent times, earables have drawn considerable interest amongst the research community for analyzing user behavior and detecting visual cues such as human gaze, hand gestures, and other verbal or nonverbal cues. The work of [1, 6, 9, 12] has demonstrated that *eSense* [8] is sensitive enough to identify not just head movements but minute changes in facial muscles as well, which we believe are essential cues for gauging listener involvement. Also, the signals emerging from the earable are less prone to noise due to restrictions on the degrees of freedom of the human head, making the signal much cleaner and usable compared to that obtained from other wearables.

Selecting and identifying features that will help mimic verbal and non-verbal cues is crucial to quantify listener involvement. In this regard, as shown in Fig. 1, our system *enVolve* uses IMU data obtained from eSense and detects facial expressions, head movements, and body postures of the listener during online interaction and uses these gestures to understand how well the person is involved in the conversation and synced up with the speaker. The main idea behind this implementation is that in virtual interaction, the silent gestures of the listener act as a reflection of acknowledgment; this acknowledgment could be presented to the speaker as a visual aid to understand the listeners' involvement in the interaction. However, developing a model to infer such acknowledgments is challenging because the IMU data captured from the earables need to be mapped to different gesticulations at different body parts (like facial muscles, heads, upper body parts, etc.). The inference is needed to be drawn from a collective analysis of such gesticulations. Interestingly, the IMUs from the earables return a complex amalgamation of signals from different body parts. For example, a listener can simultaneously smile and nod their head, and the IMU produces a compound signal from both gesticulations.

In this paper, we develop and implement *enVolve*, a framework for monitoring the involvement of silent listeners in online meetings, on *eSense* IMU-based earbuds. Here, eSense is chosen to collect the accelerometer and gyroscope data for the listeners moving their heads in different orientations and paces. For the preliminary study, we used 17 volunteers in the age of 20 – 25 (Female: 12, Male: 5) and recorded 50 – 55 minutes of IMU data for each volunteer while they attended online meetings (online classrooms). From this data, we first explore whether earables are the better choice for inferring the involvement level of silent listeners compared to other sensing modalities like smartwatches and smartphones, and then develop an initial model for the same by exploiting different associated features from the upper body part and facial movements of the meeting participants.

In contrast to the existing works, the contributions of this paper are as follows. (1) We identify features from gestures, such as head movements, facial expressions, and body postures, that mimic verbal and non-verbal cues and demonstrate the efficacy of earables. (2) Our preliminary evaluation reveals that *enVolve* identifies overlapping features with reasonable accuracy (around 80%), typically a complex problem for hand-worn wearables [16, 17]. (3) We present our ongoing attempts toward developing a Google Meet extension that provides the speaker with a visually intuitive interface to gauge user involvement.

2 BIFURCATING COGNITIVE INVOLVEMENTS: DOES EARABLE HELP?

It is not uncommon to find an involved and engaged listener in an online meeting or lecture, to perform affirmative body gestures following the speaker's progress. To mimic the offline classroom process, the listeners' spatio-temporal context must be shared with the speakers. But, *which features of the body movements are helpful to capture the involvement of a listener?* And, *What is the best mode of collecting/capturing these features?*

To answer the first question, we surveyed around 500 students of the *Indian Institute of Technology, Jodhpur* and the *Indian Institute of Technology, Kharagpur* to understand the various features required

for identifying the involvement of a silent listener in an audio-only online interaction. The students who participated in the study belong to undergraduate, postgraduate, and doctorate programs. The participants' ages lie between 18 to 40 years (mean: 20.77, std: 7.49). Based on the survey results, it was observed that (a) the listeners sit in an upright position when they are involved or trying to get involved, (b) there are gesticulations in their facial expressions, and (c) the listeners may acknowledge by nodding their head, or may not nod if they are uninvolved. Thus, head movements, body posture, and facial expressions become important features to capture the spatio-temporal context of the listener. It is to be noted that all of these movements are performed involuntary by most of the participants (Fig. 2(b)). The survey also observed that just a tiny percentage (6%) of the participants believe that they have never indulged in gestures during an online meeting or lecture.

To answer the second question, we observe another outcome of the survey, which reveals that 72% of the participants use earphones while attending online lectures (Fig. 2(c)). This makes IMU-based earables a natural instrumentation choice to understand listener involvement levels. Other methods of collecting IMU data, such as smartphones and smartwatches, have found widespread use for human activity recognition. Notably, most of the features that could be linked with the involvement levels of the listeners are upper body features, which would be difficult to capture using smartphones or smartwatches simply because these are placed in the user's pockets or worn in hand. To test this hypothesis, we instrumented some volunteers with an earable, smartwatch, and smartphone each while attending an online lecture, as depicted in Fig. 2(a). A cursory examination of the IMU data generated from these devices (Fig. 3) shows that earables are more sensitive in capturing the movements emanating in the upper body and hence are ideal to be used in online meetings. We present a further detailed analysis of this in Section 4.

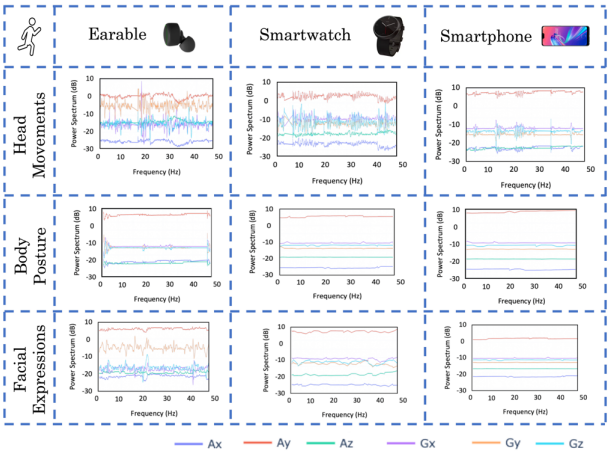


Figure 3: Different IMU Signatures captured from earable, smartwatch, and smartphone for different body movements from one participant at the same timestamp for all devices.

Our findings and prior investigations [6, 8, 9, 12] have indicated that visual clues of human body language, such as an affirmative head shake (horizontal head movement), perplexed facial expressions (inward cheek movement), or correct body postures (upright), convey evidence of conversational involvement. Combinations of these cues might be interpreted as elements of listener involvement during a conversation. This paper proposes using *head, body, and facial expressions* to measure listener engagement. Table 1 explains these qualities.

Table 1: List of features considered for identifying involvement in *enVolve*

Type of movement	Feature Class
Facial	Inward movement in eyebrows and cheeks
	No movement in eyebrows and cheeks
	Outward movement in eyebrows and cheeks
Body	LaidBack Body Posture
	UprightBody Posture
Head	Head movement in Horizontal Axis
	Head movement in Vertical Axis

3 *ENVOLVE*: SYSTEM DESIGN

enVolve comprises two modules: the listener engagement analyzer and the speaker visualizer. First, we present the process for analyzing listener involvement, then how this data is shared with the speaker via a visualizer. Both activities are basic, lightweight, and take very little bandwidth, allowing them to be used online without difficulty, even on a slow machine. Fig. 4 depicts the proposed listener involvement analyzer's processing pipeline.

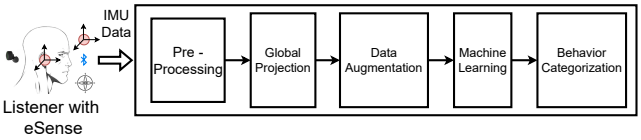


Figure 4: *enVolve* pipeline for involvement identification.

Pre-Processing: *enVolve* gathers 6-axis IMU data from the listener *eSense* [10] and stores it in windows of 10 seconds. Noise and bias must be removed from this data. This phase includes user identification. Integration magnifies noise and bias readings, which impacts the Global Projection (addressed next). Noise is spikes that can be observed without listener movement. This discovery increases the processing window, which improves estimates. Low-frequency signals from head motions are eliminated.

Global Projection: IMU data is saved in its local reference frame; however, movements can happen in such a way that they can cause the default frame of reference to change. As shown in Fig. 1, the head orientation varies according to the users' movement. In order to keep the axes global for all readings, the orientation of the IMU device needs constant rotation. Thus, the data needs to be projected in the global frame where constant changes in the axes can be seen. We use an existing approach like [18] for this purpose.

Data Augmentation: Listeners' features overlap heavily. When a listener acknowledges by nodding, facial expressions help establish/represent participation in the dialogue. Machine learning should separate these actions from the IMU signal. Different data

augmentations/views are created and fed to the models for better learning. We use uniformly random 3D rotation, signal inversion, and order 4 random scrambling for various views. Models are trained with this augmented data. We found that data augmentation improved the models' capacity to discover similarities between unknown and learned data with highly distinguishable feature overlap.

Machine Learning: This phase focuses on the system's machine learning model, which learns to detect the numerous activity features in Table 1 performed by the listener in real-time and shares the prediction with the next step, *Behavior categorization*. A few well-known machine learning models ((a) Convolutional Neural Network (CNN), (b) Long short-term memory (LSTM), (c) Support Vector Machine (SVM), and (d) Decision Tree (DT)) have been picked for proof of concept. Table 2 lists hyperparameters. After *preprocessing*, we generate a three-dimensional array based on the training set's windows, window size, and channels. We use raw signal data for CNN and LSTM (after preprocessing and data augmentation). Others use feature engineering to get statistical and frequency domain feature vectors. Fast Fourier Transform (FFT) is applied to all frequency domain feature vectors. Mean, median, standard deviation, energy measure, skewness, and kurtosis are used to obtain new features. Six FFTs on original 6-axis features and 36 statistical operations on all features create 42 features. Table 1 lists 7 labels for head motions, body postures, and face movements.

Table 2: Hyperparameters of ML Models used

Model	Layers	Filters/Kernel	Filter Shapes	Batch Size	Depth
CNN	3	32	30, 30, 32	16	-
LSTM	4	64	28, 28, 64	16	-
SVM	-	Linear	12,12	-	-
DT	-	-	-	-	7

Behaviour Categorization: The pipeline's behavior categorization module gets Table 1's collection of listener actions. This module employs a basic heuristic-based technique to link body postures, head motions, and facial expressions, then apply a rule-based approach to determine if a subject is engaging in a meeting's talks during a time window Δ. *enVolve* uses upright body posture as a marker of involvement in its rule-based modeling. We associate body position with head gestures and facial expressions to avoid confusion. This module assumes active listeners have an upright posture, shake their heads to acknowledge, and have involuntary facial emotions, as observed from the survey presented in §2. Further, this module works on the principles of sliding window over Δ, as it processes the model's outcome for every instance of that window and labels either involved or not-involved based on the activities detected; it manages a window of 10 seconds cumulative of all the cases.

The data from the listener involvement analyzer is communicated with the speaker via a visualizer (Fig. 5), which gives feedback on listener engagement. This visualizer works with Chrome and Google Meet. The visualizer has three parts. *First*, it extracts the list of google meet listeners and adds their involvement level. *Second*, it assesses each listener's level and adds a description under their name. *Third*, it calculates the audience's state as a pie chart and delivers analysis to the speaker under the chart. For instance, as shown in Fig. 5, a couple of listeners are lacking behind because

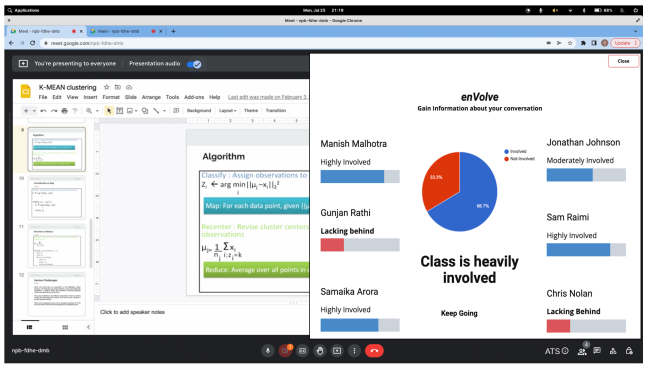


Figure 5: Visualiser providing information about the listeners' involvement level along with overall class status

their engagement level is low; however, at the same time, other students have high engagement levels, creating an overall heavily engaged audience. This visualizer only takes half of the display size to convey the information and does not cause any disruption in the flow of the conversation from the speaker's side.

4 IMPLEMENTATION AND EVALUATION

enVolve is built with compatibility with eSense, an IMU-based earable device. The IMU data, collected at 50 Hz, is recorded through an android application via Bluetooth. The proposed framework *enVolve*'s core is implemented using Tensorflow. The smartphone IMU data is obtained from Asus Zenfone Max Pro M2, and the smartwatch used for the same purpose is Moto 360. Since the data collected is heavily imbalanced, the measuring scale used for this study is the F1 score, which is a better alternative to judge a system's work when dealing with imbalanced data.

4.1 Data Collection

A lab-scale study was conducted at the campus of *Indian Institute of Technology, Jodhpur* to collect the data for the experiment. 17 volunteers in the age range $20 - 25$ were recruited to generate the dataset for this study (mean age: 23.52, std age: 1.538, female: 12, male: 5, ethnicity: Indian). Fig. 2(a) shows the setup for volunteers with device placement and the model of the device used for collecting data; these volunteers did not participate in the survey conducted in section 2. For each volunteer, $50 - 55$ minutes of accelerometer and gyroscope readings are recorded. To boost the dataset's variance, five volunteers were allowed to surf the Internet and view videos while listening to the speaker; the rest were required to simply keep the browser open to listen to the speaker. These volunteers were videotaped from two angles: (a) front-facing using a laptop webcam and (b) mobile-mounted to capture the whole body. These recordings helped annotators mark the dataset's ground truth. Three non-participating annotators were chosen. They observed participants' body posture, facial expressions, and head movements and marked involved or not involved every second, correlating to accelerometer and gyroscope data in the CSV file. All three annotators marked the final ground truth, but the majority voted. We computed *Cohen's kappa* [14] and obtained

Table 3: F1 Score, Precision, Recall scores of testing algorithms for detecting engagement activities

Algorithm	F1 Score	Precision	Recall
Conv-NN	79.64	72.24	80.84
LSTM	77.68	73.65	89.60
SVM	68.82	73.84	84.43
DT	76.67	72.47	79.21

Table 4: Class wise F1 score of CNN for all participants while detecting engagement activities

Algorithm	Class Name	F1 Score	Precision	Recall
Conv-NN	Inward Facial Movement	78.13	71.77	80.34
	Outward Facial Movement	80.67	77.15	79.27
	No Facial Movement	76.25	68.87	77.49
	Laid Back Body Posture	77.18	78.35	76.11
	Upright Body Posture	72.97	60.16	70.87
	Horizontal Head Movement	88.89	76.88	91.81
	Vertical Head Movement	83.40	72.54	89.99
	Average	79.64	72.24	80.84

Table 5: Class wise F1 score of CNN for all disengaged participants

Algorithm	Class Name	F1 Score	Precision	Recall
Conv-NN	Surfing the internet	70.17	68.49	77.33
	Engaged in other activities	81.23	81.05	82.44
	Average	75.65	74.07	78.44

a *kappa* value of 0.8 which indicates a good/substantial agreement among the annotators.

4.2 Device Type Feasibility

Earables, smartphones, and smartwatches were tested in two ways. *First*, find the gadget with the cleanest raw data. *Second,* Determine the dataset's usability. In these investigations, machine learning algorithms (as mentioned in Section 3) were utilized to identify features and total involvement scores based on IMU data. Our evaluation yielded the following results.

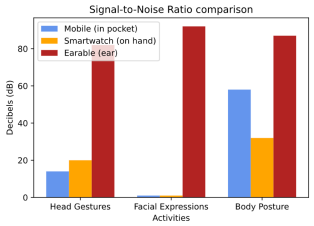


Figure 6: SNR over data captured over three devices for identifying gestures

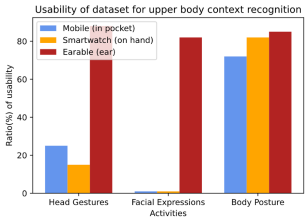


Figure 7: Dataset usability for three devices, for identifying gestures

4.2.1 Signal-to-Noise Ratio (SNR). Here we study the power signals observed and recorded from the three devices. Power spectral density can be generated from an FFT of the IMU data [13]. For the context of activity recognition, we compute the SNR from the power spectral density as $SNR = 20 \log \frac{S}{N}$, where S is the power of the signal for a time window t, and N is the noise power during the previous time window $t-1$. We divide the captured IMU signals in 10sec windows for this computation. The constant 20 is used in the above equation, as the signal strength measurements are in watts (power). Fig. 6 shows the results favoring the earable with the highest SNR value amongst all three devices for detecting target activity features like the change in body posture, head gestures, and facial expressions.

4.2.2 Usability of dataset recorded from multiple devices. Here we study the usability of the dataset for upper body context recognition recorded from the three devices. We collected data from the three modalities from 5 different participants and ran the classifiers for detecting various body movements shown in Table 1. The ratio was taken from the collected dataset to ascertain how many times we could identify the different activities correctly and shown in terms of % in Fig. 7. With the same parameters as shown in Table 2, CNN and SVM were used to classify different activities from the data collected from all three devices. It can be seen from Fig. 7 that smartphones and smartwatches are not quite sensitive to recording the minute movements of facial muscles and seem to lose signal power while reaching head movements. A similar observation can be made from Fig. 3, that earables provide much cleaner IMU signatures even for sensitive muscle movements. One interesting observation can be seen that all three devices are more than capable of identifying the body postures with high accuracy.

4.3 Performance Benchmarking of *enVolve*

enVolve was tested on two parameters: correctly identifying listeners' engagement level during online interaction using CNN, SVM, LSTM, and DT (described in Section 3) and measuring the framework's runtime resource usage. The benchmarking compares the system's memory and bandwidth usage with and without the suggested framework. Benchmarking shows the system's real-world performance and usefulness.

4.3.1 Model Performance. Table 3 summarizes the performance of different machine learning models in correctly inferring the activities mentioned in Table 1, performed by the user during the interaction. It can also be seen from the results and Fig. **??** that our proposed framework can correctly identify multiple overlapping features with a good F1 score of nearly 80%, along with a high recall value. The results of the best performing model, i.e., CNN, is expanded further, *firstly,* in Table 4, which exhibits class-wise F1 score of all the classes mentioned in Table 1; *secondly,* in Table 5, which exhibits class-wise F1 score of all the activities resembling disengagement. Based on the predictions made by the machine learning models, the behavior categorization module of the pipeline (Fig. 4) can correlate these predictions with the state of the listener.

4.3.2 Benchmarking of enVolve. We benchmark the memory and bandwidth consumption by *enVolve* to test its usability over handheld devices. This benchmarking has been done for over 5 subjects in a live classroom setup. The online interaction was completely audio mode from the listeners' side; however, the speaker shared a presentation while speaking. The online classes lasted for 50 minutes, during which the bandwidth and memory consumption changed due to the speaker's presentation sharing, which added extra load on the browser to use more memory. The framework's use hasn't greatly increased resource consumption, according to this evaluation. After a few minutes of a minor increase from 8% to 27% RAM utilization, it achieved saturation. Overall Memory usage increased by 10% for 50 minutes. Bandwidth use is similar, with an

initial increment of 26%, saturation was achieved after a few minutes, and an overall 13% climb for 50 minutes. This benchmarking provides enough evidence that the system in its current state is useful and provides good foundational support for scalability.

5 RELATED WORK

In recent times, inertial sensing through earables has gained prominence in determining human activities and behavior [6, 10, 19]. The IMU data generated from earables such as eSense devices have exhibited promising results in characterizing gestures like head nodding, head shaking, chewing, speaking, etc. The user study performed by Choi *et al.* [4], conducted with the help of earable devices and wristbands, revealed head nodding and hand gestures can be utilized to determine a meeting's success. According to [4], a meeting has three stages: initiation, group discussion, and turning points. Hand and head gesture cues were calculated for these phases with 79% accuracy. However, unlike *enVolve*, their study was not concerned about providing feedback to the speaker. Also, the study participants were involved in marking the ground truth, which can break the flow of the audience to perceive the lecture as they are focusing on filling the survey during the lecture; hence, an automated method is required. Another similar work [11] has attempted to use earables to detect the participants' head movements and body posture in an online meeting. They, however, have not quantified the involvement level of the participants. Moreover, in this work, the participants were also involved in marking the ground truth; thus, the model may introduce subjectiveness in the evaluation.

Gashi *et al.* [5] have used earables to learn the movements of facial muscles to classify gestures such as smile, yawn, talk, confusion, etc., with an F1 score of 84%. However, independent classification of head gesticulations, body posture, and facial expressions will not address the problem of determining the level of involvement of passive listeners in an online meeting. Although some of the past efforts [11, 20] have attempted using inertial sensing to gauge the involvement of the meeting attendees, they lack in providing continuous feedback about the listeners to the speaker, which questions the effectiveness of the solutions in a real-time scenario.

Other approaches such as [2] and [7] have tried to provide real-time feedback and assistance by analyzing user behavior. However, their methodology uses audio and video data, which requires additional computation and resources. Our implementation works as a low-cost yet effective tool to analyze user behavior without significantly adding or increasing the load on the system. To the best of our knowledge, none of the existing works till now have tried to correlate the speaker-listener using earables during an online meeting in the absence of any *acoustic-visual cues*.

6 DISCUSSION AND CONCLUSION

enVolve demonstrates promising results in passively estimating the involvement level of a silent listener in an online meeting using the IMU data collected from eSense earable. Based on the positive results from this study, several other objectives have risen, such as an end-to-end system based on earables that can provide individual and global learning, and the attention level of a specific user, while improving the current visualizer to add more features. Another potential objective is to detect or derive different characteristics of individual users, such as their emotional well-being, using the same data. We plan to improve the current accuracy of the models and implement a novel technique based on the learning objectives of the models used in this study.

This investigation yielded relevant and fascinating findings. Cognitive expressions vary by person. Such signals can differ even for a single human. In such instances, it's difficult for machines to learn user context and personalize services. We need a gesture-invariant technique to model the system to eliminate subjective bias from individuals. Continuous and pervasive monitoring of behavioral gestures and psychophysiological markers can give individualized intelligent services on demand.

REFERENCES

[1] Ashwin Ahuja, Andrea Ferlini, and Cecilia Mascolo. 2021. PilotEar: Enabling In-ear Inertial Navigation. In *ACM EarComp*. 139–145.

[2] Bon Adriel Aseniero, Marios Constantinides, and Daniele Quercia. 2020. MeetCues: Supporting online meetings experience. In *2020 IEEE VIS*. IEEE, 236–240.

[3] Wilver Auccahuasi, Christian Ovalle, and Gloria Rojas. 2020. Analysis of the level of attention and meditation in children, in the development of online classes caused by the covid-19, through a brain-computer interface. (2020).

[4] Jun-Ho Choi, Marios Constantinides, and Daniele Quercia. 2021. KAIROS: Talking heads and moving bodies for successful meetings. In *HotMobile '21*. 30–36.

[5] Shkurta Gashi, Aaqib Saeed, and Silvia Santini. 2021. Hierarchical Classification and Transfer Learning to Recognize Head Gestures and Facial Expressions Using Earbuds. In *ICMI*. 168–176.

[6] Tahera Hossain, Md Shafiqul Islam, Md Atiqur Rahman Ahad, and Sozo Inoue. 2019. Human activity recognition using earable device. In *Adjunct ACM UbiComp and ACM ISWC*. 81–84.

[7] Ryo Iijima, Akihisa Shitara, Sayan Sarcar, and Yoichi Ochiai. 2021. Word Cloud for Meeting: A Visualization System for DHH People in Online Meetings. In *The 23rd ACM SIGACCESS*. 1–4.

[8] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Alessandro Montanari. 2018. Earables for personal-scale behavior analytics. *IEEE PerCom* 17, 3 (2018), 83–89.

[9] Fahim Kawsar, Chulhong Min, Akhil Mathur, Alessandro Montanari, Utku Günay Acer, and Marc Van den Broeck. 2018. eSense: Open Earable Platform for Human Sensing. In *16th ACM ENSS*. 371–372.

[10] Fahim Kawsar, Chulhong Min, Akhil Mathur, Alessandro Montanari, Utku Günay Acer, and Marc Van den Broeck. 2018. eSense: Open earable platform for human sensing. In *16th ACM SenSys*. 371–372.

[11] Dongwoo Kim, Chulhong Min, and Seungwoo Kang. 2021. Towards Automatic Recognition of Perceived Level of Understanding on Online Lectures using Earables. In *Adjunct ACM UbiComp and ACM ISWC*. 158–164.

[12] Matias Laporte, Preety Baglat, Shkurta Gashi, Martin Gjoreski, Silvia Santini, and Marc Langheinrich. 2021. Detecting Verbal and Non-Verbal Gestures Using Earables. In *Adjunct ACM PerCom and UbiComp*. ACM, Virtual USA, 165–170.

[13] Boon-Leng Lee, Boon-Giin Lee, and Wan-Young Chung. 2016. Standalone wearable driver drowsiness detection system in a smartwatch. *IEEE Sensors journal* 16, 13 (2016), 5444–5451.

[14] Mary L McHugh. 2012. Interrater Reliability: The kappa statistic. shorturl.at/aIKQR

[15] Carrie Anne Platt, NW Amber, and Nan Yu. 2014. Virtually the same?: Student perceptions of the equivalence of online classes to face-to-face classes. *Journal of Online Learning and Teaching* 10, 3 (2014), 489.

[16] David Pollreisz and Nima TaheriNejad. 2017. A simple algorithm for emotion recognition, using physiological signals of a smart watch. In *39th IEEE EMBC*. IEEE, 2353–2356.

[17] Juan Carlos Quiroz, Elena Geangu, and Min Hooi Yong. 2018. Emotion recognition using smart watch sensor data: Mixed-design study. *JMIR mental health* 5, 3 (2018).

[18] Dirk Robinson and Peyman Milanfar. 2003. Fast local and global projection-based methods for affine motion estimation. *JMIV* 18, 1 (2003), 35–54.

[19] Yushi Takayama, Shun Ishii, Anna Yokokubo, and Guillaume Lopez. 2021. Detecting forward leaning posture using eSense and developing a posture improvement promoting system. In *ACM EarComp*. 178–179.

[20] Matin Yarmand, Jaemarie Solyst, Scott Klemmer, and Nadir Weibel. 2021. It Feels Like I am Talking into a Void: Understanding Interaction Gaps in Synchronous Online Classrooms. In *ACM CHI*. 1–9.

Excerpt of ToothSonic: Earable Authentication via Acoustic Toothprint

Zi Wang
Florida State University
Tallahassee, FL, USA
ziwang@cs.fsu.edu

Yili Ren
Florida State University
Tallahassee, FL, USA
ren@cs.fsu.edu

Yingying Chen
Rutgers University
Piscataway, NJ, USA
yingche@scarletmail.rutgers.edu

Jie Yang
Florida State University
Tallahassee, FL, USA
jie.yang@cs.fsu.edu

ABSTRACT

Earables (ear wearable) are rapidly emerging as a new platform to enable a variety of personal applications. The traditional authentication methods thus become less applicable and inconvenient for earables due to their limited input interface. Earables, however, often feature rich around the head sensing capability that can be leveraged to capture new types of biometrics. In this work, we propose ToothSonic that leverages the toothprint-induced sonic effect produced by a user performing teeth gestures for user authentication. In particular, we design several representative teeth gestures that can produce effective sonic waves carrying the information of the toothprint. To reliably capture the acoustic toothprint, it leverages the occlusion effect of the ear canal and the inward-facing microphone of the earables. It then extracts multi-level acoustic features to represent the intrinsic acoustic toothprint for authentication. The key advantages of ToothSonic are that it is suitable for earables and is resistant to various spoofing attacks as the acoustic toothprint is captured via the private teeth-ear channel of the user that is unknown to others. Our preliminary studies with 20 participants show that ToothSonic achieves 97% accuracy with only three teeth gestures.

CCS CONCEPTS

• **Security and privacy → Biometrics**.

KEYWORDS

Biometrics, Toothprint, User Authentication, Earable

ACM Reference Format:
Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. Excerpt of ToothSonic: Earable Authentication via Acoustic Toothprint. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3544793.3563420

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09…$15.00
https://doi.org/10.1145/3544793.3563420

1 INTRODUCTION

Earables are rapidly emerging as a new platform to enable a variety of personal applications due to their rich around the head sensing capability. There also have been increasing research efforts to leverage earables to achieve tasks such as understanding our fitness and sleeps, accessing information, identifying contextual information, monitoring or tracking activities [2].

While earables show considerable promise, they also raise new questions in terms of security. This is because much of the value of the services offered by earables depends on the confidential and personal information they capture, process and transmit. However, adapting traditional authentication from other wearables or mobiles can be challenging. Quite simply, earables lack a suitable input interface to support rapid and reliable entry of passwords or most of the traditional biometrics. Voice-based authentication is convenient but has been proven vulnerable to voice spoofing attacks [6, 7]. Despite the issue, earables provide novel opportunities to improve or redesign approaches to authentication due to their rich around the head sensing capability. For example, recent work utilizes earable to sense ear canal and its deformation [5] for authentication. However, emitting acoustic sound to probe the ear canal could be intrusive for those who are sensitive to high-frequency sound.

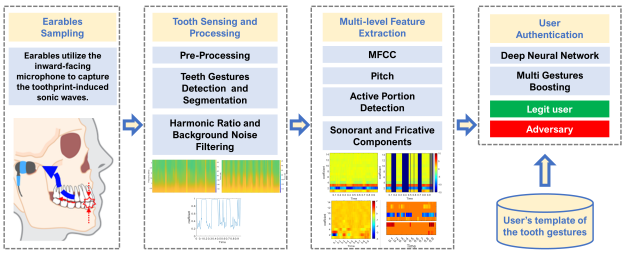


Figure 1: System flow of ToothSonic.

In this work, we propose ToothSonic, a secure earable authentication system that leverages the toothprint-induced sonic effect produced by a user performing teeth gestures for user authentication. In particular, when teeth slide or strike against each other, part of their mechanical energy is released in the form of sonic waves. The harmonics of the friction- and collision- excited sonic wave are

Biometric Factors / Tooth Gestures	Organ-level				Macro-level							Micro-level			
	Dental Mobility F/B	Dental Mobility U/D	Dental Mobility L/R	Propagation Channel	Dental arch shape	Depth of spee	Occlusion classes	Dental spacing	Incisor shape and size	Canine shape and size	Molar shape and size	Cusp	Enamel thickness	Enamel rod patterns	Tooth root
Occlusion Sliding			√	√	√	√	√	√	√	√	√	√	√	√	√
Molar Sliding	√			√		√		√			√	√	√	√	√
Canine Sliding	√			√				√	√	√			√	√	√
Incisor Sliding F/B	√			√			√	√	√	√			√	√	√
Incisor Sliding U/D		√		√			√	√	√	√			√	√	√
Incisor Sliding L/R			√	√			√	√	√	√			√	√	√
Occlusion Tapping		√		√	√	√	√		√	√	√	√	√		√
Molar Tapping		√		√		√					√	√	√		√
Canine Tapping		√		√		√			√	√	√		√		√
Incisor Tapping		√		√			√		√	√			√		√

Figure 2: Tooth gesture and related biometrics.

dependent on the teeth composition, the dental geometry, and the surface characteristics of each tooth [1]. The key insight is that the sonic waves produce from a teeth gesture carry the information of the toothprint. As every individual has a unique toothprint just like our fingerprint, two users perform the same teeth gesture will result in distinct toothprint-induced sonic waves, which could be sensed by the earables for user authentication. Compared with traditional biometrics, it has several advantages.

Anti-Spoofing. The friction- and collision- excited sonic waves are dependent on the toothprint, which is hidden in the mouth and skull, the head and skull. It is thus more resilient to spoofing attacks compared with traditional biometrics (e.g., fingerprint, face, and voice) that could be exposed to others. In addition, the sonic waves travel through the head tissues and skull channel, which hold the individual uniqueness acting as a hidden and encrypted channel that modulates the sonic waves. ToothSonic is thus resistant to sophisticated adversaries who can acquire the victim's toothprint, for example, via the dentist.

Wide acceptability. ToothSonic provides eye-free and hands-free authentication when hands and eyes are occupied (e.g., carrying objects or driving). It is also more socially acceptable than voice-based authentication in public places (e.g., offices and libraries) as the sonic waves of teeth gestures are much less perceptible and unobtrusive to others.

Implicit authentication. ToothSonic can also be exploited as an implicit authentication method when teeth gestures are used as a hands-free computer interface, for example potentially in "Switch Access" services, and for people with motor impairments.

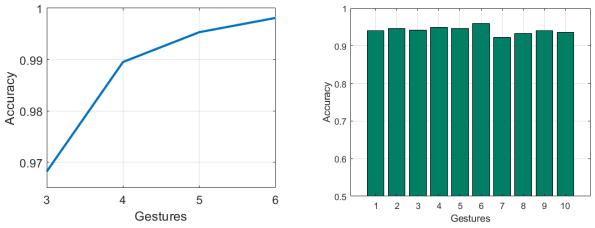
2 SYSTEM DESIGN

Our system comprises four major components, as shown in Fig. 1. The system uses energy-based event detection to locate the gestures. Next, our system segments the recorded signals into a sequence of gestures by utilizing the Munich Automatic Segmentation system [3]. To enhance the SNR, we apply the harmonic ratio to filter our background noise when no gestures are performing.

As shown in Fig. 2, we design 10 teeth gestures including 6 sliding gestures and 4 tapping gestures. The sliding gestures contain occlusion sliding, molar sliding, canine sliding, incisor sliding front/back, incisor sliding up/down, and incisor sliding left/right. And the tapping gestures are occlusion tapping, molar tapping, canine tapping, and incisor tapping. These gestures cover the major factors that affect the sonic waves of the toothprint when performing gestures.

3 PERFORMANCE EVALUATION

Fig. 3 (a) shows the accuracy that leverages multiple gestures. In sum, we could see that ToothSonic achieves high accuracy over 99% by combining a few gestures. In particular, our system could



(a) Multiple gestures **(b) Different gestures**

Figure 3: Authentication accuracy.

achieve authentication accuracy of 99.81%, 99.53%, 98.95%, 96.82% by with 6, 5, 4, 3 gestures, respectively.

Fig. 3 (b) shows the accuracy across 10 different gestures when using only one gesture for authentication. No.1 to No.6 stand for the six different sliding gestures and the left 4 gestures are tapping gestures. We observed that the performance of the sliding gesture is better than the tapping gestures. This is because sliding gestures have a longer duration and contain more tooth participants with different dimensions of information. Therefore, sliding gestures contain more features than tapping gestures, and thus could provide more accurate authentication. We could also find that the accuracy of canine gestures is the lowest. This is due to canine gestures only involve one side canine with less information, and such gestures are harder for users to perform in our experiments.

4 CONCLUSIONS

This paper proposed the excerpt of ToothSonic [4], a secure earable authentication system that leverages the toothprint-induced sonic effect produced by teeth gestures for user authentication. ToothSonic has several advantages over traditional biometric authentication including anti-spoofing, wide acceptability, and conditionally implicit authentication. We investigate representative teeth gestures that produce effective sonic waves carrying the information of the toothprint. Multi-level acoustic features are also extracted to represent intrinsic toothprint information. Our preliminary results demonstrate the effectiveness of ToothSonic in authenticating earable users.

REFERENCES

[1] Jean-François Augoyard. 2006. *Sonic experience: a guide to everyday sounds.* McGill-Queen's Press-MQUP.

[2] Romit Roy Choudhury. 2021. Earable computing: A new area to think about. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications.* 147–153.

[3] Florian Schiel, Christoph Draxler, and Jonathan Harrington. 2011. Phonemic segmentation and labelling using the MAUS technique. (2011).

[4] Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. ToothSonic: Earable Authentication via Acoustic Toothprint. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–24.

[5] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27.

[6] Linghan Zhang, Sheng Tan, Zi Wang, Yili Ren, Zhi Wang, and Jie Yang. 2020. VibLive: A Continuous Liveness Detection for Secure Voice User Interface in IoT Environment. In *Annual Computer Security Applications Conference.* 884–896.

[7] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on CCS.* 1080–1091.

Multimodal Attention Networks for Human Activity Recognition From Earable Devices

Jake Stuchbury-Wass
js2372@cam.ac.uk
University of Cambridge
Cambridge, UK

Andrea Ferlini
andrea.ferlini@nokia-bell-labs.com
Nokia Bell Labs
Cambridge, UK

Cecilia Mascolo
cm542@cl.cam.ac.uk
University of Cambridge
Cambridge, UK

ABSTRACT

Earables (a.k.a ear-worn wearable devices) are gaining traction in the wearables ecosystem for monitoring user health. Human activity recognition (HAR) is a promising use case of earables due to their placement on the head and the combination of sensors. In this paper, we explore using multimodal attention-based neural networks for HAR from the ear. Attention networks have had a large impact on other disciplines' machine learning tasks and we believe they present opportunities in HAR from earable data. Different methods of utilising attention mechanisms in the literature are discussed as well as the benefits and challenges of using such networks in the context of HAR on real systems.

1 INTRODUCTION

Earables are emerging as wearable user monitoring device [4]. These can contain multiple sensors which may be harnessed to perform, among other things, human activity recognition (HAR) [7]. Being mounted on the head, earables can extract information about movement of the head as well as the net movement of the body. Like with other wearable data, deep learning shows state-of-the-art results on earable data. This paper explores using an attention mechanism as part of a network for HAR on earable data for their improved performance, robustness and interpretability over other deep learning methods.

Human activity recognition is one of the most fundamental tasks in mobile sensing. It has applications in monitoring wellbeing and healthcare. HAR from wearables can be regarded as a time-series classification task, a long-standing problem with applications to healthcare and finance. In the past, this has been performed with traditional machine learning classifiers [12]. Normally, Inertial Measurement Unit (IMU) data is the main source of data for classification, but other data streams such as heart rate and body temperature can be used for HAR. Recently, deep learning techniques have been used to make these classifications and extract features in the same pipeline. This can incorporate multiple data streams more easily, and thus increase the robustness of the model [11]. As a lot of the deep learning models used for HAR previously have been inspired by advances in vision like the Convolutional Neural Network (CNN) or sequence modelling such as Recurrent Neural Networks (RNN), we argue that the latest and state-of-the-art technique, attention mechanisms, should be employed for HAR too and that the benefits of doing so outweigh the challenges.

Attention mechanisms [1] are designed to mimic human cognitive attention in a computer by learning what parts of the data are most relevant to each other and, therefore, finding the important features. Neural networks based on attention mechanisms have improved performance on many tasks in deep learning [10]. The transformer model architecture was proposed in 2017 [15] and led to the widespread use of attention-based models. The transformer is a sequence-to-sequence model, of which the main benefit is that it acts over the whole input sequence with an attention mechanism rather than having a limited window such as previous state-of-the-art such as RNNs. The transformer has led to successful large language models such as the Bidirectional Encoder Representations from Transformer model (BERT) [2] and its many derivatives. Besides its use in natural language processing, the transformer and attention mechanisms have led to the development of the vision transformer [3], which achieves state-of-the-art performance on computer vision tasks.

2 ATTENTION NETWORKS: A PRIMER

An attention network contains at least one layer that has learned parameters for the relevance of different inputs relative to the other inputs at each position. This relevance is calculated with a series of matrix multiplications for a query given certain keys and values. For sequence modelling, for example in translation, the input sequence is turned into a series of tokens with a positional embedding as the attention layer inputs. Transformers are an example of a successful purely attention-based network.

3 OPPORTUNITIES OF ATTENTION ON HAR

Opportunity 1: **Performance, robustness and versatile designs of attention layers.** Using attention networks for HAR from earable data gives improved performance as well as the robustness of using multimodal data such as IMU, audio, PPG and other sensors. An example is the combination of attention and convolution exploited for multimodal timeseries by Tripath et al. [14], which outperforms other deep learning methods on HAR datasets. Additionally, a promising method that resolves where to focus and what to focus on was proposed by Gao et al. [5]. They use both temporal and cross-channel attention layers in a Dual Attention network. This network was specifically developed for HAR tasks and shows an increased performance over other deep learning methods on HAR datasets. This combination of different attention mechanisms allows the network to outperform other attention-based networks while remaining end-to-end trainable, unlike some other networks.

Opportunity 2: **Multitask generalisation.** Another way to exploit the attention mechanism is with its generalisability. These

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09.
https://doi.org/10.1145/3544793.3563422

networks can be used to perform multiple tasks on the same data streams. This is particularly relevant to HAR where we may desire information such as gait and posture. Liu et al. developed a multi-task attention-based model that can run on-device [9] showing the possibility for this model to generalise and still be lightweight.

Opportunity 3: **Visualising the attention.** The interpretability of the model is also a benefit of attention networks, since it involves visualising the attention weights [13]. This has particular importance when applied to healthcare where HAR can be used. Gao et al. also discuss visualisation of channel importance by looking at attention weights [5]. This has applications in reducing the model complexity by not including data streams or features deemed to be less relevant, this could be important to running a model on-device in real-time.

4 CHALLENGES

In the previous section, we discussed the opportunities presented by applying attention networks to earable-based HAR. However, to achieve those, there are a number of challenges researchers have to face to implement these networks onto real systems. This section will reflect on these.

Challenge 1: **Keeping the model lightweight.** While developing the model, the end use case of running on a mobile device must be kept in mind. This means keeping the model lightweight enough for the target device. Attention models can have higher parameter counts than other deep learning models [16] so an increase in performance must be weighed against a loss in on-device inference time. Liu et al. [9] use a model with temporal shift attention and convolution layers that runs on-device using more efficient convolutions. Established techniques for reducing model size such as quantisation and pruning [8] are important for keeping the model lightweight. Additionally, hardware designed for edge on-device inference is being developed, bringing increased performance in mobile deep learning [17].

Challenge 2: **"Data Hungry" models.** Attention-based models can be described as "data hungry" and require larger datasets to train compared to other deep learning models [6]. This could reduce the performance of HAR from earables compared to models that are conventionally employed if sufficiently large datasets are not available. One solution to this problem could be to use larger datasets from other wearable devices and implement transfer learning, where a model trained on a larger dataset can learn the characteristics of the sensor data and then be retrained on a smaller, earable and task-specific dataset.

Challenge 3: **Adapting techniques from other ML domains and lack of established frameworks.** Implementing transformers and attention layers for domains such as natural language and vision is aided by well-developed frameworks. But for timeseries applications, these frameworks are in their infancy, which can make developing and experimenting with models time consuming. However, there are emerging frameworks, since this is an active area of research and timeseries tasks are well supported for other networks such as CNNs and RNNs. These can be adapted for attention layers.

5 OUTLOOK

Attention networks and transformers have revolutionised natural language and vision fields giving state-of-the-art performance. We have argued the opportunities presented by attention networks for HAR outweigh the challenges of implementing the models on real systems. Researchers will need to focus on keeping the models lightweight enough for mobile devices, the datasets used, as well as schemes to adapt techniques from other ML domains.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. https://doi.org/10.48550/arXiv.1409.0473 arXiv:1409.0473 [cs, stat]

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929 arXiv:2010.11929 [cs]

[4] Andrea Ferlini, Dong Ma, Lorena Qendro, and Cecilia Mascolo. 2022. Mobile Health With Head-Worn Devices: Challenges and Opportunities. *IEEE Pervasive Computing* 01 (2022), 1–9.

[5] Wenbin Gao, Lei Zhang, Qi Teng, Jun He, and Hao Wu. 2021. DanHAR: Dual Attention Network for Multimodal Human Activity Recognition Using Wearable Sensors. *Applied Soft Computing* 111 (Nov. 2021), 107728. https://doi.org/10.1016/j.asoc.2021.107728

[6] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. 2021. Escaping the Big Data Paradigm with Compact Transformers. *CoRR* abs/2104.05704 (2021). arXiv:2104.05704 https://arxiv.org/abs/2104.05704

[7] Tahera Hossain, Md Shafiqul Islam, Md Atiqur Rahman Ahad, and Sozo Inoue. 2019. Human Activity Recognition Using Earable Device. In *Adjunct 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 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 81–84. https://doi.org/10.1145/3341162.3343822

[8] Tailin Liang, John Glossner, Lei Wang, Shaobo Shi, and Xiaotong Zhang. 2021. Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing* 461 (2021), 370–403.

[9] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. 2020. Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 19400–19411.

[10] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A Review on the Attention Mechanism of Deep Learning. *Neurocomputing* 452 (Sept. 2021), 48–62. https://doi.org/10.1016/j.neucom.2021.03.091

[11] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D. Lane, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. 2018. Multimodal Deep Learning for Activity and Context Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (Jan. 2018), 157:1–157:27. https://doi.org/10.1145/3161174

[12] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. 2005. Activity Recognition from Accelerometer Data. In *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence - Volume 3* (Pittsburgh, Pennsylvania) *(IAAI'05)*. AAAI Press, 1541–1546.

[13] Blaž Škrlj, Sašo Džeroski, Nada Lavrač, and Matej Petkovič. 2020. Feature Importance Estimation with Self-Attention Networks. https://doi.org/10.3233/FAIA200256 arXiv:2002.04464 [cs, stat]

[14] Achyut Mani Tripathi and Rashmi Dutta Baruah. 2020. Multivariate Time Series Classification With An Attention-Based Multivariate Convolutional Neural Network. In *2020 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN48605.2020.9206725

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]* (Dec. 2017). arXiv:1706.03762 [cs]

[16] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2022. Transformers in Time Series: A Survey. https://doi.org/10.48550/arXiv.2202.07125 arXiv:2202.07125 [cs, eess, stat]

[17] Amir Yazdanbakhsh, Kiran Seshadri, Berkin Akin, James Laudon, and Ravi Narayanaswami. 2021. An evaluation of edge tpu accelerators for convolutional neural networks. *arXiv preprint arXiv:2102.10423* (2021).

Excerpt of Auritus: An Open-Source Optimization Toolkit for Training and Development of Human Movement Models and Filters Using Earables

Swapnil Sayan Saha, Sandeep Singh Sandha, Siyou Pei, Vivek Jain, Ziqi Wang, Yuchen Li, Ankur Sarker, and Mani Srivastava
swapnilsayan@g.ucla.edu
University of California, Los Angeles
Los Angeles, CA, USA

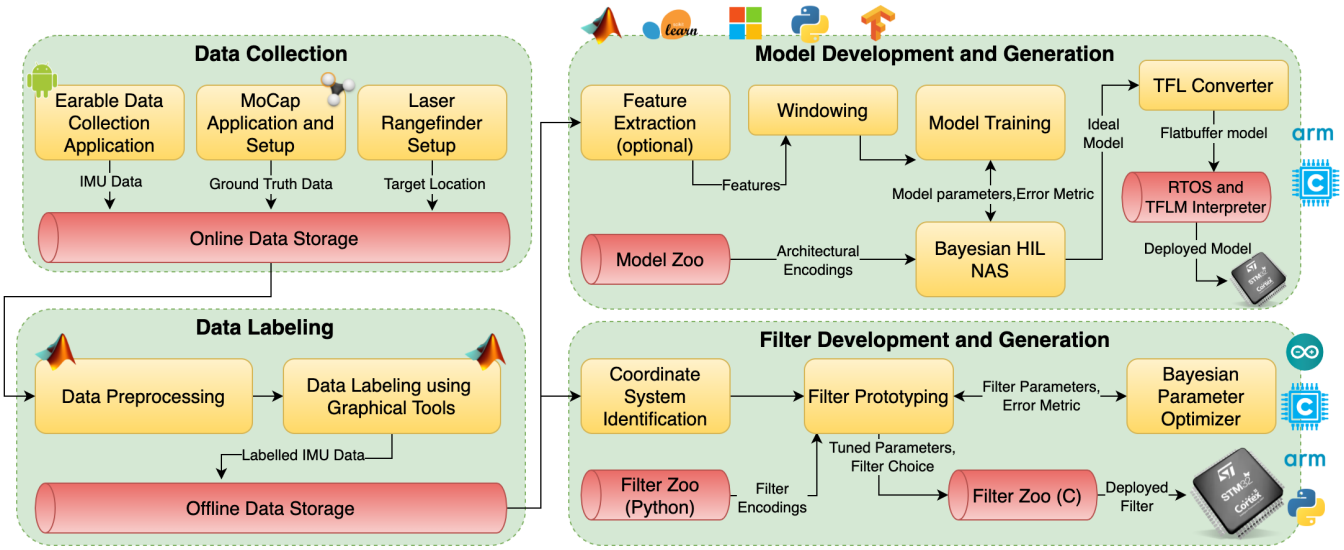


Figure 1: Architecture of Auritus [1]. The first two modules take care of collecting and labeling high resolution earable data interactively. The development and generation modules allow model and filter optimization through automated hardware-in-the-loop Bayesian neural architecture search and optimization, respectively.

ABSTRACT

Auritus is an extendable and open-source optimization toolkit designed to enhance and replicate earable applications. Auritus serves two primary functions. *Firstly*, Auritus handles data collection, pre-processing, and labeling tasks for creating customized earable datasets using graphical tools. The system includes an open-source dataset with 2.43 million inertial samples related to head and full-body movements, consisting of 34 head poses and 9 activities from 45 volunteers. *Secondly*, Auritus provides a tightly-integrated hardware-in-the-loop (HIL) optimizer and TinyML interface to develop lightweight and real-time machine-learning (ML) models for activity detection and filters for head-pose tracking. Auritus recognizes activities with 91% leave 1-out test accuracy (98% test accuracy) using real-time models as small as 6-13 kB. Our models are 98-740× smaller and 3-6% more accurate over the state-of-the-art. We also estimate head pose with absolute errors as low as 5 degrees using 20kB filters, achieving up to 1.6× precision improvement over existing techniques. Auritus is available at https://github.com/nesl/auritus.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Computing methodologies** → *Machine learning*; • **Computer systems organization** → Embedded systems.

KEYWORDS

earable, network architecture search, neural networks, machine learning, datasets, filters, human activity, head-pose, TinyML, optimization, hardware-in-the-loop

ACM Reference Format:
Swapnil Sayan Saha, Sandeep Singh Sandha, Siyou Pei, Vivek Jain, Ziqi Wang, Yuchen Li, Ankur Sarker, and Mani Srivastava. 2022. Excerpt of

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09.
https://doi.org/10.1145/3544793.3563423

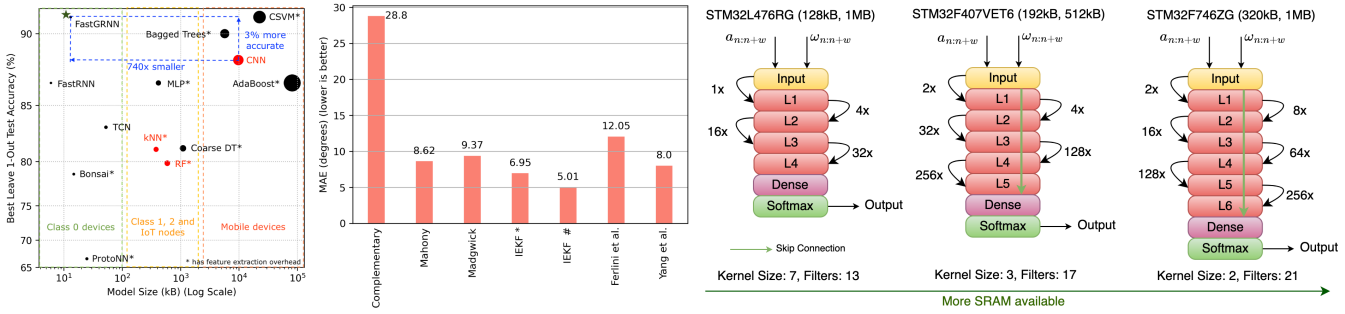


Figure 2: (Left) Accuracy vs. model size of Auritus models (black) vs. state-of-the-art models (red). (Middle) Error of Auritus filters vs. proposed earable head-pose filters. (Right) Our NAS adapts the same model to exploit full device capabilities.

Auritus: An Open-Source Optimization Toolkit for Training and Development of Human Movement Models and Filters Using Earables. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3544793.3563423

1 INTRODUCTION

The bulk of emerging innovations in earables builds upon ML advances for wearable activity detection. However, the data-hungry nature of ML training demands access to large-scale earable datasets, which are hard to come by. Moreover, given the tight memory, power, and compute constraints of earables (e.g., 56 kB SRAM, 16 MB flash), deploying on-device AI-driven activity detection and head-pose estimation is challenging. Auritus addresses the scarcity of earable datasets and software tools by providing:

- An open-source earable dataset from 45 volunteers containing 34 distinct head-poses and 9 classes of full-body activities with 2.43 million samples.
- Tools to enable data collection, processing and labeling.
- A zoo of 5 lightweight models, 5 conventional models, and 4 headpose filters that significantly outperform the state-of-the-art in terms of accuracy and resource usage.
- A HIL Bayesian neural architecture search (NAS) framework for training and deploying said models and filters on earables.

2 DATA COLLECTION AND LABELLING

For data logging, we used the eSense earable device from Nokia Bell Labs. The 6-channel inertial data was broadcasted at ~100 Hz to an in-house smartphone application we developed using eSense Android middleware backend. For sub-mm resolution ground truth collection, the participants wore a hat with OptiTrack Prime 17W MoCap infrared markers. The motion data of the head and the marker visual cues were tracked using Motive:Tracker and screen recorder applications, respectively. For the head-pose dataset, we collected 34 distinct head-poses from 27 targets per participant, totalling 45 participants. For the activity dataset, 9 classes of actions were recorded, namely walking, jogging, jumping, standing, turning left, turning right, sitting, lying, and falling. A total 6.75 hours of human movement data is available. To ease labeling the data collected in continuous chunks, we designed a graphical-user-interface to allow head-pose and activity data annotation via plots. The application developer selects points directly on a plot signifying the start and endpoints of regions of interests. After specifying all the endpoints and making any numerical adjustments to the data, the developer runs a script to perform automatic segmentation and labeling based on the endpoints.

3 MODEL AND FILTER DEVELOPMENT

To enable real-time activity classification on resource-constrained devices, we included several lightweight classifiers and filters suitable for onboard activity inference in the model and filter zoo. The models include temporal convolutional network, fast gated RNN, fast RNN, Bonsai and ProtoNN. The filters include complementary filter, Madgwick filter, Mahony filter and indirect extended Kalman filter. To find the ideal activity detection model candidate from the model search space for limited flash, RAM, and latency requirements, or to optimize filter parameters, we included a HIL Bayesian optimizer. The goal is to find a model/filter that maximizes the hardware SRAM and flash usage within the device capabilities while minimizing latency and error on the validation set. Our optimizer communicates with the target hardware during the optimization process to guarantee deployability and architectural adaptation based on resource constraints. We use Gaussian process as the surrogate model and gradient-free Monte Carlo sampling with Upper-Confidence Bounds as the acquisition function.

4 KEY RESULTS

Auritus generates models that are 98-740× smaller, yet 3-6% more accurate over the state-of-the-art. Our models are as small as 6-13kB with 91-98% accuracy, and our fall detection models are as small as 2kB with 98% accuracy. Our filters have ~5 degrees of error with 20 kB of code, providing 1.6× improvement over the state-of-the-art. Our NAS framework performs intelligent architectural adaptation and device capability exploitation based on resource availability.

REFERENCES

[1] Swapnil Sayan Saha, Sandeep Singh Sandha, Siyou Pei, Vivek Jain, Ziqi Wang, Yuchen Li, Ankur Sarker, and Mani Srivastava. 2022. Auritus: An Open-Source Optimization Toolkit for Training and Development of Human Movement Models and Filters Using Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–34.

Excerpt from "Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena"

Tobias Röddiger
Karlsruhe Institut of Technology
Karlsruhe, Germany
roeddiger@teco.edu

Christopher Clarke
University of Bath
Bath, United Kingdom
cjc234@bath.ac.uk

Paula Breitling
Karlsruhe Institut of Technology
Karlsruhe, Germany
breitling@teco.edu

Tim Schneegans
Karlsruhe Institut of Technology
Karlsruhe, Germany
schneegans@teco.edu

Haibin Zhao
Karlsruhe Institut of Technology
Karlsruhe, Germany
zhao@teco.edu

Hans Gellersen
University of Lancaster
Lancaster, United Kingdom
h.gellersen@lancaster.ac.uk

Michael Beigl
Karlsruhe Institut of Technology
Karlsruhe, Germany
beigl@teco.edu

ABSTRACT

By adding sensing capabilities to ear-worn devices, earables have emerged as a new platform. The ears are located closely to a number of important anatomical structures (e.g., brain, blood vessels). Also, the ear canal deforms upon facial movements and the ears can be comfortably touched by the hands. In a recent paper, we conducted a systematic literature review of 271 earable papers. We synthesized an open-ended taxonomy of 47 phenomena that can be sensed in, on, or around the ear. We identified 13 fundamental phenomena from which all other phenomena can be derived, and discuss sensing principles to detect them. The phenomena were reviewed in-depth in four main areas: (i) physiological monitoring and health, (ii) movement and activity, (iii) interaction, and (iv) authentication and identification. This breadth highlights the potential earables have to offer as a ubiquitous, general-purpose platform.

CCS CONCEPTS

• **General and reference** → **Surveys and overviews**.

KEYWORDS

earables, hearables; earphones; headphones; earbuds; ear wearable; earpiece; ear-worn; ear-mounted; ear-attached; ear-based

ACM Reference Format:
Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Excerpt from "Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena". In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3544793.3563413

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09…$15.00
https://doi.org/10.1145/3544793.3563413

1 INTRODUCTION

Earables are wearable that are worn in or near the ears. Hearing aids and personal speakers are examples of ear-worn electronics that have emerged for specific uses. As a new class of ubiquitous computing platform, we define "earables" as devices that integrate broader features. Due to their distinctive placement on the human body, earables (also known as "hearables") offer a unique platform for sensing a variety of properties, processes, and activities. At the heart of much of the research in this new field are questions of sensing - what can be detected and observed with earables, and what interactions and applications are enabled by sensing in or on the ear?

In a recent paper, we conducted a systematic and comprehensive literature review of 271 earable publications, representing the state of the art in earable sensing. This excerpt gives a brief introduction to article No. 135 published in Volume 6, Issue 3 in the Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies [1].

2 METHODOLOGY

Inspired by related reviews, we followed a structured process by collecting and filtering papers from the ACM and IEEE digital libraries using a number of defined selection criteria (e.g., sensing has to occur on or around the ears). This was supplemented with backward chaining using the same criteria applied to papers from other digital libraries (e.g., Springer). We used the following guiding definition of earables: "Earables are devices that attach in, on, or in the immediate vicinity of the ear to offer functionalities beyond basic audio in- and output." The described process resulted in 271 relevant articles which we analysed and clustered, and from which we derived a novel earable taxonomy consisting of different phenomena.

Figure 1: Overview map of earable sensing. The map is organised by phenomena that can be captured with earable sensors.

3 CONTRIBUTION

Each paper in the dataset was categorized with respect to the applied sensing principle, the corresponding phenomena which was sensed, and the application for which it was used. Through multiple iterations, we then synthesized a higher-level taxonomy of phenomena sensed with earables that is open-ended and extensible (new sensors might emerge, and further phenomena explored).

At the lowest level, we then identified and characterised phenomena that are directly sensed with sensors placed in or on the ear as *fundamental phenomena*, including for instance motion, body temperature and blood perfusion. Other phenomena are identified as indirectly observable and derived from fundamental phenomena, ranging from physiological parameters (e.g., heart rate) and lower-level cues (e.g., earable state; in or out of ear) to conditions (e.g., stress), actions (e.g. gestures), activities (e.g. daily tasks) and other context (e.g. user identity). In total, we identified and categorised close to 50 phenomena. We showed how higher-level phenomena build on 13 fundamental phenomena, and relate this to 21 different sensors and sensing principles that have been investigated for their observation.

Figure 1 provides a simplified map of these phenomena, clustered into related themes and grouped into four main areas of research: (1) physiological parameters and health; (2) movement and activity; (3) interaction; and (4) authentication and identification. For each of these research areas, we provide an extensive overview of the research conducted in the space.

Further sensing developments will likely expand the already rich set of phenomena sensed by earables, and new applications will emerge that leverage those currently available. However, to-date most earable sensing research has not been rigorously tested in-the-wild. Future work will have to demonstrate ecological validity and overcome robustness and engineering challenges to unleash the full potential that earables have to offer as a ubiquitous, general-purpose platform.

REFERENCES

[1] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (Sep 2022), 135.

Designing and evaluating a system for studying EarEEG signals

A Adarsh
TCS Research
Bangalore, India

Meghana S
Tata Consultancy Services
Bangalore, India

Kartik Muralidharan
TCS Research
Bangalore, India

Jayavardhana Gubbi
TCS Research
Bangalore, India

Ramesh Kumar Ramakrishnan
TCS Research
Bangalore, India

Arpan Pal
TCS Research
Kolkata, India

ABSTRACT

Electroencephalography (EEG) allows the study of the brain in humans with applications in areas of Psychology, Brain-Computer Interfaces, and Neuromarketing. The conventional systems limit the applications of EEG to lab environments, but recent developments towards portable EEG have allowed brain studies outside the laboratory in more realistic situations. One such direction is the Ear EEG, a wearable concept for recording the EEG using an ear worn device. This paper discusses a system design to study the Ear EEG compared to the scalp EEG using the OpenBCI. We evaluate the signal quality obtained with Ear EEG using a customized earpiece and demonstrate the feasibility of recording alpha attenuation using Ear EEG.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; *Mobile devices.*

KEYWORDS

EEG; EarEEG

ACM Reference Format:
A Adarsh, Meghana S, Kartik Muralidharan, Jayavardhana Gubbi, Ramesh Kumar Ramakrishnan, and Arpan Pal. 2022. Designing and evaluating a system for studying EarEEG signals. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3544793.3563417

1 INTRODUCTION

There exist several brain imaging modalities, such as Electroencephalography (EEG), Magnetoencephalography (MEG), Functional near-infrared spectroscopy (fNIRS), and functional Magnetic resonance imaging (fMRI), used to understand the anatomy and function of the brain for monitoring and diagnostic purposes. These modalities have their advantages and disadvantages in terms of spatial and temporal resolution, portability, and cost, and most are not well-suited for mobile and lifestyle integration. EEG is widely used to investigate the brain in applications such as neurophysiology, brain-computer interfaces, and neuromarketing.

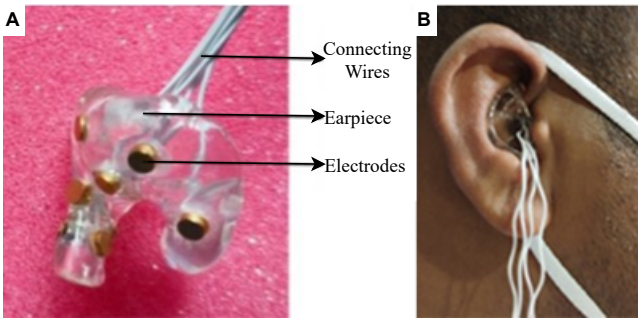


Figure 1: (A) Custom designed ear mould with electrodes placed and (B) A subject wearing the ear mould

EEG is a non-invasive recording of the brain's electrical activity, typically by placing electrodes on the scalp's surface. The EEG on the scalp estimates synaptic action at large scales closely related to behavior and cognition. Thus, EEG can provide a genuine window on the mind [10]. The clinical systems to record EEG require a conductive gel applied to the electrode sites. The use of conductive gels causes discomfort to the subjects and limits the recording time as the signal quality degrades with the drying of the gels. Although EEG analysis has advanced in academic and professional contexts, the complexity of clinical EEG systems and the requirement of specially trained personnel have made them unsuitable for mobile and lifestyle integration.

Researchers in recent years are working towards developing portable EEG technologies, offering a wireless, compact, and hassle-free EEG system to enable monitoring of the brain to understand the neural correlates of behavior, mental processes, and sleep. [12]. Wearable EEG is a technology that enables the translation of neuroscience from the laboratory to the real-life environment. The Ear EEG is a wearable concept of recording the EEG in the ear introduced by Looney et al. [8] demonstrating the feasibility of recording the ear EEG. The Ear EEG has shown to produce signal-to-noise ratios comparable to those from conventional EEG electrode placements and is robust to familiar sources of artifacts. The studies have been conducted on ear EEG to check its effectiveness in recording auditory and visual evoked potentials[7].

In this paper, we have presented the early results of the feasibility of Ear EEG and the methodology to build a rigorous and comprehensive system using OpenBCI to study a larger population of subjects and to span several protocols, and in this way, assess the use cases of Ear EEG compared to scalp EEG.

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09…$15.00
https://doi.org/10.1145/3544793.3563417

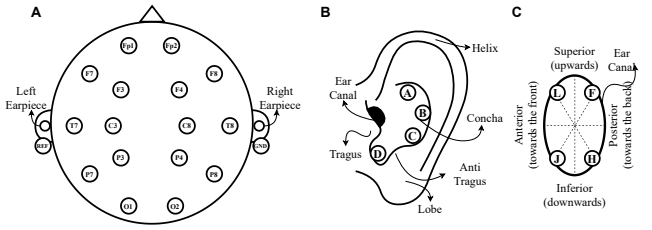


Figure 2: Electrode Configuration on Scalp and Ear. (A) Electrode placement on the scalp according to the 10-20 system. The reference electrode is placed on the left earlobe, and the ground electrode is placed on the right earlobe (B) Sketch of the exterior part of the left ear, showing the four regions corresponding to the electrode labels A through D. (C) Cross-sectional view of the left ear canal (sagittal plane) showing the electrode labels in the left ear canal. The electrodes are labeled based on the direction relative to the vertical axis and not based on the depth in the ear canal

2 SYSTEM DESIGN

2.1 Earpiece

This section discusses the design of customized earpieces. The earpieces were developed according to the anatomical shape of the subjects. An audiologist captured the impression of the ear by cleaning the ear and inserting a cotton ball with a thread inside the ear canal. Then, silicone was filled into the ear canal and the concha region. The impression was then carefully removed once the silicone had dried. The ear impression was then used to create a 3D ear mould. Then electrodes were placed according to the positions described in Figure 2. Various electrode materials were simulated [2], and gold-plated electrodes were selected. Gold-plated surface mount PCB contact was used as ear-EEG electrodes. The electrode had a diameter of 3mm and was 1.9mm thick. Figure 1 shows the earpiece design with electrodes mounted and a subject wearing the earpiece.

2.2 Electrode Configuration

Scalp. The ScalpEEG consisted of 16 electrodes according to the international 10-20 system and were FP1, FP2, F3, F4, F7, F8, C3, C4, T7, T8, P3, P4, P7, P8, O1, O2. The scalp electrode positions are shown in Figure 2A. Dry Comb electrodes were used for scalpEEG.

Ear. The EarEEG consisted of eight electrodes in each ear, with four electrodes placed in the ear canal and concha. The electrodes labeling convention follows [6]. The electrodes are labeled Exy, where x denotes the left(L) or right(R) ear, and y denotes the position in the ear. Positions A through D covers the concha region at four different locations. Positions F, H, J, and L, denote the ear canal electrodes. Figure 2B and 2C depicts the positions of the electrodes in the ear concha and canal.

2.3 Hardware

The OpenBCI Cyton-Daisy Biosensing Boards were used for recording EEG. The OpenBCI system is a low-cost open-source EEG amplifier with a proven record as an alternative to medical-grade EEG amplifiers [4] [11]. The OpenBCI Cyton is an eight-channel system with a sampling rate of 250Hz over Bluetooth, while 16 channels can be used with the add-on Daisy Board at a reduced sampling rate of 125Hz. The system uses the ADS1299, a 24-bit biopotential measurement integrated circuit developed by Texas Instruments. A PIC32-based microcontroller is implemented on the board, providing plenty of local memory and fast processing speeds. In addition, A 3-axis accelerometer (LIS3DH). A module for storing data in a micro SD card. The application to communicate with OpenBCI is an open-source application written with Processing language. The amplifier supports both wet and dry electrode configurations and has support for external sensors, and was chosen given the capability for customizing the recording configurations. There also exists a few issues with the OpenBCI, the main concerns being low sampling rate and high sensitivity of electrical noise to resolve which additional hardware is required.

3 METHODOLOGY

3.1 Experiment Design

Protocols. Five different protocols and the alpha attenuation task were chosen as stimuli for the subjects. The motivation for choosing these stimuli was to activate the five broad lobes and to study their influence on Ear EEG

Alpha Attenuation Task The subjects were instructed alternately to rest with open and closed eyes for 1 minute each. Auditory cues were presented at one-minute intervals, instructing the subjects to alternate between performing eyes-open and relaxing with eyes closed. This activity of opening and closing eyes is known to attenuate the alpha band (8-13Hz) in EEG.

Four Word Sentence Task The processing of words and sentences and memory tasks have been associated with the parietal lobe. The subject is presented with a set of four-letter words, following which the subject is to generate a meaningful sentence with words starting from each letter of the word [1].

Go-No-Go Task The role of the frontal cortex in response inhibition. The frontal cortex has been reported to be active during the infrequent stimulus. The Go/No-go task is an experimental protocol used to determine the response inhibition of the subject. The subject must respond by pressing a button when they see a "go" condition and not respond when they see the "no-go" condition.

Steady State Visual Evoked Potential (SSVEP) The human visual cortex can perceive light modulation to frequencies up to 75Hz, which is reported to be most sensitive around 10Hz. The SSVEP is typically induced by flashing light through a monitor or using a light source such as LED [7].

Motor Task Motor task can be defined as the subject's mental process of motor activity with actual motor movement. It is a well-reported fact that motor movement causes event-related synchronization (ERS) or event-related desynchronization (ERD) in the central *mu* (8-12Hz) and *beta* (13 - 35Hz) bands [9].

Auditory Steady State Response (ASSR) Auditory steady-state responses (ASSR) were reported as a method to assess the hearing threshold level objectively. Amplitude-modulated narrow or broadband sound signals can evoke the ASSRs. The neural encoding of the modulating frequency is visible in the frequency spectrum of the EEG [7].

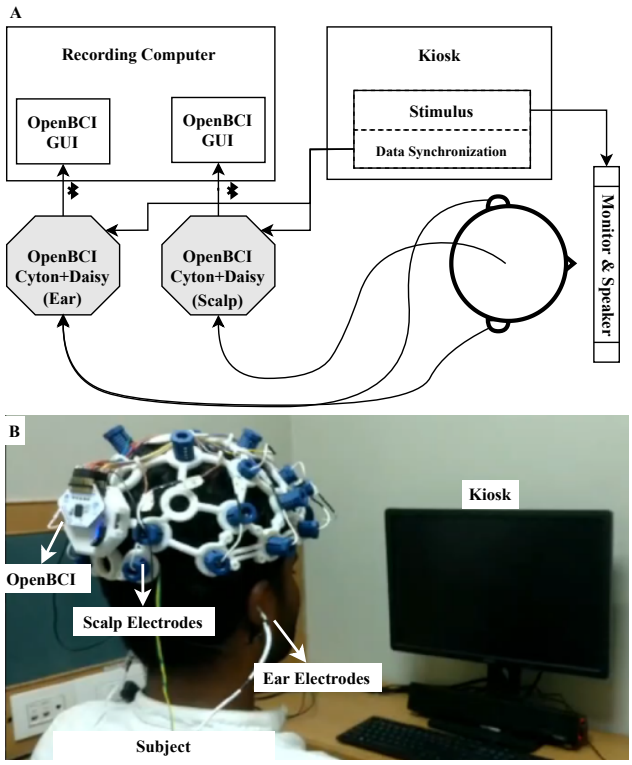


Figure 3: A) The block diagram of the recording setup, B) A subject in the recording setup wearing the scalp and ear electrodes

Kiosk. The protocols have been deployed on a kiosk developed in-house[5]. The kiosk has been developed predominantly in the Java platform and extensively tested on Microsoft Windows. Custom libraries were written to handle the stimulus drawing, hardware interfacing, data handling, etc. The kiosk is a subject-friendly tool and does not need the experimenter's intervention during the task to give the subject's privacy during the experiments.

3.2 Data Acquisition

The data was collected from 6 subjects after informed consent. For this study only the alpha attenuation paradigm is considered. The EEG recording was done using two OpenBCI Cyton-Daisy amplifier systems, each for the scalp and ear. The raw data was recorded using the OpenBCI GUI at a sampling rate of 125Hz. The electrode sites were cleaned using Isopropyl alcohol prior to electrode placement. The Ultracortex Mark iv headset was used to place the scalp electrodes and earpieces for the ear electrodes. The ground electrode of both the scalp and ear was placed on the right earlobe and the reference electrode on the left earlobe. Figure 3A shows the block diagram, and Figure 3B shows a subject in the recording setup.

Data Validation To validate the correctness of the EEG collected few checks were done visually on the OpenBCI GUI. Each channel data was visually inspected to check the signal by instructing the participants to blink their eyes and clench their jaws. Impedance was measured, and made sure low impedance was maintained for

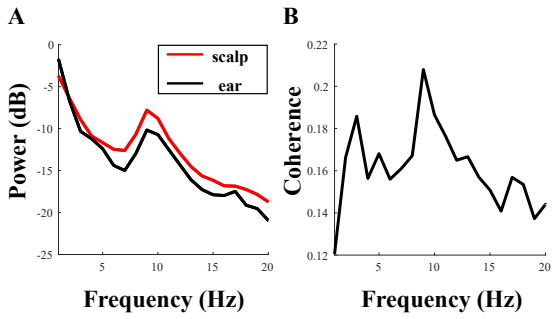


Figure 4: A) Average PSD results of the the scalp and ear electrodes for a Subject for the Baseline recordings, B) Average spectral coherence between the scalp and ear electrodes

good EEG signals. The railed information in the OpenBCI GUI was checked, and made sure that all channels were not railed.

Data Synchronization The data synchronization was at two levels i) to synchronize the scalp and ear EEG and ii) to synchronize the stimulus with EEG data. An Arduino-based system was developed to synchronize the OpenBCI systems with the stimulus. The kiosk communicated with the Arduino over serial communication, and the Arduino responded to these messages by toggling the digital output pin. The same digital pin output was fed to the OpenBCI systems through optical isolators, thus synchronizing the datastreams of the EEG with the stimulus. The event markers were recorded on the OpenBCI using the digital I/O pins in the AnalogRead mode.

3.3 Data Processing

The first step of the data processing was the synchronization of the scalp and ear datasets. The auxiliary analog channel recorded on the two OpenBCIs was used to calculate the event markers and these event markers was then used to align the data streams. Few datasets were excluded from analysis due to noisy recordings. The data preprocessing steps for scalp and ear EEG remained the same.

The data processing was done using EEGLAB, and the steps are as follows. The data was cropped to reject the first and last 10 seconds of data. The data was then highpass filtered to 1Hz. Additionally, the zapline toolkit[3] was used to remove the first and second harmonics of the line frequency (50Hz). Post filtering, manual inspection of the data was done to annotate bad data segments, and the bad segments were rejected. Bad channels were detected using the OpenBCI railed/not railed logic, and the channels were also rejected based on visual inspection.

4 EXPERIMENTAL RESULTS

Quality of Ear EEG. The quality of the Ear EEG was verified by comparing the power spectral density of the ear and scalp EEG. The power spectral density was computed with the multitaper method using a single taper. Figure 4A shows the average power spectral density of a subject's ear and scalp EEG. It was observed that the ear EEG showcased less power compared to the scalp. Spectral coherence was computed to determine the similarity between the scalp and ear EEG across frequencies. The spectral coherence ranges from 0-1, with values close to 1 depicting higher coherence at those frequencies. The spectral coherence between the scalp and ear EEG

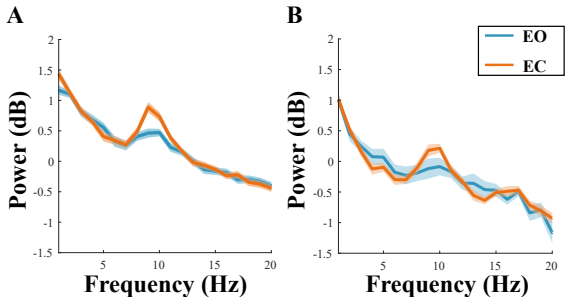


Figure 5: A comparision between eyes open (EO) and eyes close (EC) condition during baseline recordings in A) scalp electrode (O1) and B) ear electrode (ERL)

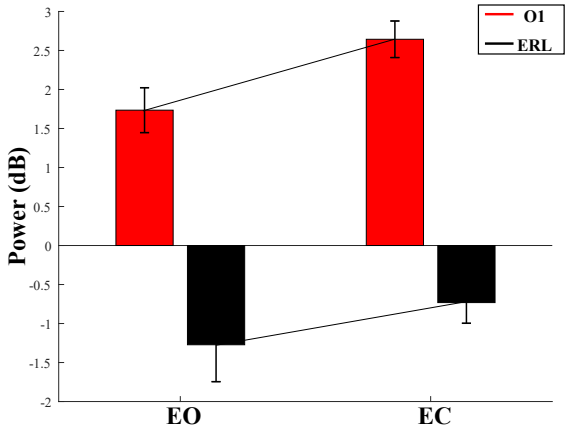


Figure 6: Bar plots depicting the change in total alpha power between eyes open (EO) and eyes close (EC) conditions in scalp electrode (O1) and ear electrode (ERL)

was the maximum around the alpha band, with averages around 0.2. Figure 4B shows the average spectral coherence between the scalp and ear electrodes.

Alpha Attenuation. The data was segmented based on the eyes open and eyes closed markers. The power spectral density was computed, and the absolute power in the alpha band was computed. The results of alpha attenuation for a typical subject (SUB003) are discussed here. Figure 5A for the scalp and Figure 5B for the ear shows the PSD plots for the respective EEG during the Eyes Open (EO) and the Eyes Close (EC) conditions. The mean alpha power of the Ear EEG was found to be lesser by half compared to scalp EEG. The change in power in alpha attenuation was computed as the ratio of absolute power in the alpha band during EC and EO. The Table1 tabulates the change in alpha power shown in Figure 6. The alpha power in the scalp EEG showed a change of 152%, whereas the Ear EEG showed a change of 57% between the EO and the EC conditions.

5 CONCLUSION AND DISCUSSIONS

In this paper, we discussed the system design to study Ear EEG. The methodology to design earpieces and interface them with OpenBCI was described. The scalp and ear EEG were recorded simultaneously

EEG Location	Scalp	Ear
Average Eyes Open alpha power (dB)	1.73	-1.27
Average Eyes Close alpha power (dB)	2.64	-0.73
Change in alpha power	1.52	0.57

Table 1: Alpha power in Scalp and Ear EEG for a subject

with common reference and ground. The subjects performed the alpha attenuation task of EO and EC.

It was observed that Ear EEG is effectively identical to similarly referenced conventional scalp EEG. The Ear EEG measurements had lower power in comparison with scalp EEG. The alpha activity during Eyes Open and Eyes Close in Ear EEG was found to have less discrimination but was completely acceptable. This demonstrates that the bio-potential acquired by the Ear EEG platform is similar to that of conventional scalp electrodes.

This motivates us to study the effect and correlation of the endogenous and exogenous cortical potentials generated in the different lobes due to various protocols, frequency bands, and frequency response characterization on the Ear EEG. Establishing reliable Ear EEG may lead the research toward sleep studies and brain-computer interfaces, which demands stable electrode performance, comfort, and scalable form factor. The aim is to develop an Ear EEG wearable with comfort levels comparable to the modern-day state-of-the-art wireless earbuds.

REFERENCES

[1] Mathias Benedek, Rainer J Schickel, Emanuel Jauk, Andreas Fink, and Aljoscha C Neubauer. 2014. Alpha power increases in right parietal cortex reflects focused internal attention. *Neuropsychologia* 56 (2014), 393–400.

[2] Abhranila Das, S Basu, Adarsh A, J Gubbi, Kartik Muralidharan, Meghana S, Mahendiran S, A Biradar, Ullas Pradhan, Tapas Chakravarty, Ramakrishnan Ramesh, and Arpan Pal. 2022. Surface Potential Simulation and Electrode Design for in-Ear EEG Measurement. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 937–940. https://doi.org/10.1109/EMBC48229.2022.9871926

[3] Alain de Cheveigné. 2020. ZapLine: A simple and effective method to remove power line artifacts. *NeuroImage* 207 (2020), 116356.

[4] Jérémy Frey. 2016. Comparison of an open-hardware electroencephalography amplifier with medical grade device in brain-computer interface applications. *arXiv preprint arXiv:1606.02438* (2016).

[5] Rahul D Gavas, Deepan Das, Tanuka Bhattacharjee, Mithun B Sheshachala, Lalit K Hissaria, Ramu R Vempada, Venkata S Viraraghavan, Anirban D Choudhury, Kartik Muralidharan, Ramesh K Ramakrishnan, et al. 2019. A sensor-enabled digital trier social stress test in an enterprise context. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1321–1325.

[6] P. Kidmose, D. Looney, and D. P. Mandic. 2012. Auditory evoked responses from Ear-EEG recordings. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 586–589. https://doi.org/10.1109/EMBC.2012.6345999

[7] Preben Kidmose, David Looney, Michael Ungstrup, Mike Lind Rank, and Danilo P Mandic. 2013. A study of evoked potentials from ear-EEG. *IEEE Transactions on Biomedical Engineering* 60, 10 (2013), 2824–2830.

[8] David Looney, Preben Kidmose, Cheolsoo Park, Michael Ungstrup, Mike Lind Rank, Karin Rosenkranz, and Danilo P Mandic. 2012. The in-the-ear recording concept: User-centered and wearable brain monitoring. *IEEE pulse* 3, 6 (2012), 32–42.

[9] Dennis J McFarland, Laurie A Miner, Theresa M Vaughan, and Jonathan R Wolpaw. 2000. Mu and beta rhythm topographies during motor imagery and actual movements. *Brain topography* 12, 3 (2000), 177–186.

[10] Paul L Nunez and Ramesh Srinivasan. 2006. A theoretical basis for standing and traveling brain waves measured with human EEG with implications for an integrated consciousness. *Clinical neurophysiology* 117, 11 (2006), 2424–2435.

[11] Srividya Pattisapu and Supratim Ray. 2021. Stimulus-induced narrow-band gamma oscillations in humans can be recorded using open-hardware low-cost EEG amplifier. *bioRxiv* (2021).

[12] Jiahui Xu and Baichang Zhong. 2018. Review on portable EEG technology in educational research. *Computers in Human Behavior* 81 (2018), 340–349.

Using Earables Platforms to Study Verbal Communication

Introducing earables to psycholinguistic research

Alejandro Pérez†
MRC Cognition and Brain Sciences Unit
University of Cambridge
Cambridge, UK
alejandro.perez@mrc-cbu.cam.ac.uk

Matthew H. Davis
MRC Cognition and Brain Sciences Unit
University of Cambridge
Cambridge, UK
matt.davis@mrc-cbu.cam.ac.uk

ABSTRACT

Earables provide a new opportunity to study conversation in the wild. They uniquely allow (*i*) accurate head motion tracking recorded synchronously with the speech signal and (*ii*) multiple people to simultaneously receive and stream conversational speech that is unconstrained by body movement. Here, our general aim is to introduce the use of earables for conducting psycholinguistic studies requiring audio and movement data jointly collected during verbal interaction in a natural setting. Specifically, we propose using earables platforms to address the relationship between head movement, speech and meaning transmission from single and multiple-person perspectives.

CCS CONCEPTS

• Applied computing~Law, social and behavioral sciences~Psychology • General and reference~Document types~General conference proceedings

KEYWORDS

Earables; head motion; conversation; linguistics; interpersonal; coordination; turn-taking; speech; verbal communication.

ACM Reference format:

Alejandro Pérez and Matthew H. Davis. 2022. Using earables platforms to study verbal communication: Introducing earables to psycholinguistic research. In *3rd International Workshop on Earable Computing (EarComp 2022). Cambridge, UK, 2 pages.*

https://doi.org/10.1145/3544793.3563414.

1 Procedure

We will collect accelerations, rotational velocities and audio signals with the eSense platform [1, 2] to measure head movements and turn-taking length during structured conversations. Data will be recorded while two participants play the board game "Taboo". This way, we can obtain concurrent linguistic measurements, along with a quantitative indicator of the success of an interaction, like the number of words correctly guessed, the number of errors and the number of conversational turns or the amount of time required to generate these guesses. Moreover, the task allows a similarly structured conversation across multiple participant dyads and the ludic aspect of the activity act as a prompt to facilitate the interaction. Potentially, we can include several groups, each of them using a different language. Lab Streaming Layer (LSL) software will synchronise the recordings across multiple participants. The relationship between signals will be measured (in bites) by computing the mutual information (MI) between time series (e.g., Gaussian Copula MI). Speech signals will be automatically transcribed to text (e.g., Google Cloud Speech-to-Text API), including speaker diarisation and words' time stamps to measure the turn-taking length. The entropy/predictability for each word and speaker intervention will be obtained by applying natural language processing (NLP) techniques (e.g., GPT-3 model).

1.1 Studying the relationship between head movements and speech

Head movements and verbal communication are interwoven; an inherent kinematic component is associated with speech production. Furthermore, by moving the head, we can transmit information (e.g., nods of agreement for approval). Tongue and mandibular movements provide visual cues that facilitate comprehension and interlocutor engagement [3]. These movements strongly track with the speech envelope. In turn, speech envelope information is essential for language comprehension and is closely tracked by listeners' brain responses to achieve speech intelligibility [4].

Earables offer considerable potential for quantitatively exploring head movement's communicative and linguistic function [5]. Here we will measure the mutual relationships between speech envelope and head motion. In recent work testing the motion tracking in the eSense platform, Ferlini et al. [6] describe that "speaking generates unwanted vibrations and micro-movements that are captured by the sensors". We hypothesise that these micro-movements captured by earables are associated with the speech envelope. Thus, speakers' (and listeners') head motion tracking of speech envelope will be measured. To test for a significance requirement of 95%, 1000 surrogates will be created for each participant. These surrogates will be obtained by perturbing the speech envelopes while preserving autocorrelation. The perturbed speech envelopes will be paired with intact head motion data, and the GCMI calculated. Then, we will compare if the group mean GCMI value is larger than the maximum null mean GCMI. In summary, we will identify whether the speech envelope provides information regarding head motion. If it is true that head micro-movements detected by the earables are correlated to the speech envelope, they could be removed from the motion signal and used as a backchannel to increase speech engagement in the listeners of the produced speech.

1.2 Studying interpersonal coordination during a conversation

Interpersonal non-verbal synchronization and smooth conversational turn-taking (i.e., with short gaps) have both been associated with successful face-to-face verbal interactions [7]. We propose to explore this association further using earables. Specifically, we will test whether head movement signals from two or more interlocutors can be used to measure the moment-by-moment degree of interpersonal synchronization. Moreover, we will test whether interpersonal synchronization correlates with the magnitude of the turn-taking lag. Finally, we will use these measurements to predict the "successfulness" of verbal exchange.

To test the validity of using earables motion signals to capture interpersonal synchronization, we will compare the two-person head motion mutual dependence against surrogate data created by re-pairing members of different, non-interacting dyads. Using the same measure, we will also evaluate the performance of the automatic classification of pairs of participants engaged in an interactive conversation as distinct from the randomly paired participants. Then, we will correlate these common head movement patterns with the gap length preceding the switch between listening and speaking. Finally, measures of the two heads' movement relationship and the length between conversational turns will be used to predict measures quantifying the efficacy of the verbal information exchange during the conversation. In summary, we will quantify the relationship between interpersonal coordination and the success of meaning transmission between conversational partners.

2 Applications and Significance

Conversational turn-taking is a critical type of coordination necessary for conversation. At any moment, there is a speaker and a listener, whose roles alternate to produce the two-way meaning transmission that characterizes a dialogue. We sense when to jump into the conversation, aided by grammatical cues, intonation and the 'backchannel' responses provided by head movements. For example, head movements can emphasize chunks of speech that are supposed to be uninterrupted [8]. However, there are situations where interlocutors have limited or no concurrent visual information. This can be the case with online interactions like avatar chats, in-game dialogues and telephone conversations. Through earables, we can synthesize spontaneous head movements to be delivered as a surrogate backchannel signal (e.g., vibrations) to help coordinate interlocutors' timing, supporting better conversational outcomes.

From head movement signals, we can determine which interpersonal interactions are most important for a given individual. This information can reveal which signals to prioritize in a multi-person situation to guide a directional hearing aid. Although there are other vision-based methods that are more reliable for head motion monitoring, earables are discrete and socially accepted, and allow real-life conversational recordings minimally supervised. We might also devise systems for providing feedback or training to individuals that struggle in everyday social situations, for example, by detecting poor timing in their conversational interventions or head movements that are at odds with prevailing cultural norms. Such systems might, in time, prove valuable for facilitating communication and enhancing speech comprehension.

We anticipate that measurements obtained from earables can reveal important cues to the status and quality of ongoing vocal communication. By combining interpersonal synchronization, turn-taking and speech predictability structure from NLP models, we will be able to tag and online monitor the nature of the interaction (e.g., socializing vs arguing) and the degree of bonding (empathy) between interlocutors.

In a nutshell, earables platforms enable the low-cost, unobtrusive monitoring of multiple people during prolonged naturalistic interaction. Altogether, allow researchers to conduct real-life experiments on interpersonal verbal communication. Here we emphasize that earables facilitate investigating how participants coordinate during verbal interaction.

ACKNOWLEDGMENTS

AP received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Sklodowska-Curie Individual Fellowship, grant agreement No 840885. MHD received intramural funding from the Medical Research Council (MC_UU_0005/5).

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom © 2022 Copyright is held by the owner/author(s). ACM ISBN 978-1-4503-9423-9/22/09. https://doi.org/10.1145/3544793.3563414

REFERENCES

[1] Kawsar F, Min C, Mathur A and Montanari A, 2018. Earables for Personal-scale Behaviour Analytics. *IEEE Pervasive Computing*, Volume: 17, Issue: 3.
[2] Min C, Mathur A, and Kawsar F, 2018. Exploring Audio and Kinetic Sensing on Earable Devices. In *WearSys 2018, The 16th ACM Conference on Mobile Systems, Applications, and Services (MobiSys 2018)*, Munich.x
[3] Medina S, Tome D, Stoll C, Tiede M, Munhall K, Hauptmann AG, Matthews I, 2022. Speech Driven Tongue Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20406-20416.
[4] Peelle JE, Gross J, Davis MH, 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb. Cortex,* 23(6), 1378-87.
[5] McClave EZ, 2000. Linguistic functions of head movements in the context of speech. *J. Pragmat.* 32, 855–878.
[6] Ferlini A, Montanari A, Mascolo C, Harle R, 2019. Head Motion Tracking Through in-Ear Wearables. In *Proceedings of the 1st International Workshop on Earable Computing (EarComp 2019),* pp. 8-13.
[7] Lahnakoski JM, Forbes PAG, McCall C, Schilbach L, 2020. Unobtrusive tracking of interpersonal orienting and distance predicts the subjective quality of social interactions. *R. Soc. Open Sci.* 7: 191815.
[8] Shockley K, Santana M-V, Fowler CA, 2003. Mutual interpersonal postural constraints are involved in cooperative conversation. *J. Exp. Psychol. Hum. Percept. Perform*. 29(2), 326–332.

OpenEarable: Open Hardware Earable Sensing Platform

Tobias Röddiger
Karlsruhe Institut of Technology
Karlsruhe, Germany
roeddiger@teco.edu

Tobias King
Karlsruhe Institut of Technology
Karlsruhe, Germany
king@teco.edu

Dylan Ray Roodt
Karlsruhe Institut of Technology
Karlsruhe, Germany
roodt@teco.edu

Christopher Clarke
University of Bath
Bath, United Kingdom
cjc234@bath.ac.uk

Michael Beigl
Karlsruhe Institut of Technology
Karlsruhe, Germany
beigl@teco.edu

ABSTRACT

Earables are ear-worn devices that offer functionalities beyond basic audio in- and output. In this paper we present the ongoing development of a new, open-source, Arduino-based earable platform called *OpenEarable*. It is based on standard components, is easy to manufacture and costs roughly $40 per device at batch size ten. We present the first version of the device which is equipped with a series of sensors and actuators: a 3-axis accelerometer and gyroscope, an ear canal pressure and temperature sensor, an inward facing ultrasonic microphone as well as a speaker, a push button, and a controllable LED. We demonstrate the versatility of the prototyping platform through three different example application scenarios. In sum, *OpenEarable* offers a general-purpose, open sensing platform for earable research and development.

CCS CONCEPTS

• **Hardware → Emerging technologies**; • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; **Ubiquitous and mobile devices**;

KEYWORDS

earables; hearables; open hardware

ACM Reference Format:
Tobias Röddiger, Tobias King, Dylan Ray Roodt, Christopher Clarke, and Michael Beigl. 2022. OpenEarable: Open Hardware Earable Sensing Platform. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3544793.3563415

1 INTRODUCTION

Earables are defined as "devices that attach in, on, or in the immediate vicinity of the ear to offer functionalities beyond basic audio in- and output." [6]. A broad spectrum of sensors have been used on the earable platform, ranging from motion, audio, and optical to biopotential, environmental, and electrical sensing principles. These have been used to detect a multitude of interesting phenomena that have been used in applications spanning several research areas including health monitoring, activity classification, interaction, and authentication and identification [6]. As a result, the earable platform has attracted attention from several, different research communities, and the number of research publications using the platform is increasing year-on-year.

A wide variety of hardware prototypes have been used in the earable research literature, ranging from commercial offerings such as Apple Air Pods[1] and cosinuss[2], prototype research platforms such as *eSense* developed by Nokia Bell Labs [5], and fully bespoke earable research devices (e.g., [3]). Of particular note is the eSense, which has accelerated the growth of earable research within the academic community. This in part was driven by the devices being freely distributed to academics across the world, providing a platform for which earable research can place and which others can openly contribute to. More recently, Chatterjee et al. [4] introduced *ClearBuds*, which has open hardware and comes equipped with dual microphones that can be used for speaker separation using beamforming. However, these earable platforms lack the extensibility that is required to take full advantage of the wide range of sensors that have been shown to be effective on the earable platform.

We introduce *OpenEarable*, the first fully open-source, Arduino-based earable research platform. *OpenEarable* aims to build upon the success of other earable prototyping platforms by providing a fully transparent and open hardware platform that enables researchers to push the boundaries of earable research. The main objective of *OpenEarable* is to provide an extensible platform that can be easily and cost-effectively manufactured for research and development purposes. In this paper, we present the first version of *OpenEarable* which features a 3-axis accelerometer and gyroscope, an ear canal pressure and temperature sensor, as well as an inward facing ultrasonic microphone and speaker. We provide an overview of the design process and an in-depth walkthrough of the hardware and software systems that make up the *OpenEarable* platform. Based on three exemplar applications from the research literature we highlight how the platform has to the potential to be used for motion-based activity tracking, detection of chewing events, and ear canal shape based authentication.

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09…$15.00
https://doi.org/10.1145/3544793.3563415

[1] AirPods Pro - https://www.apple.com/airpods-pro/
[2] cosinuss - https://www.cosinuss.com/en/technology/

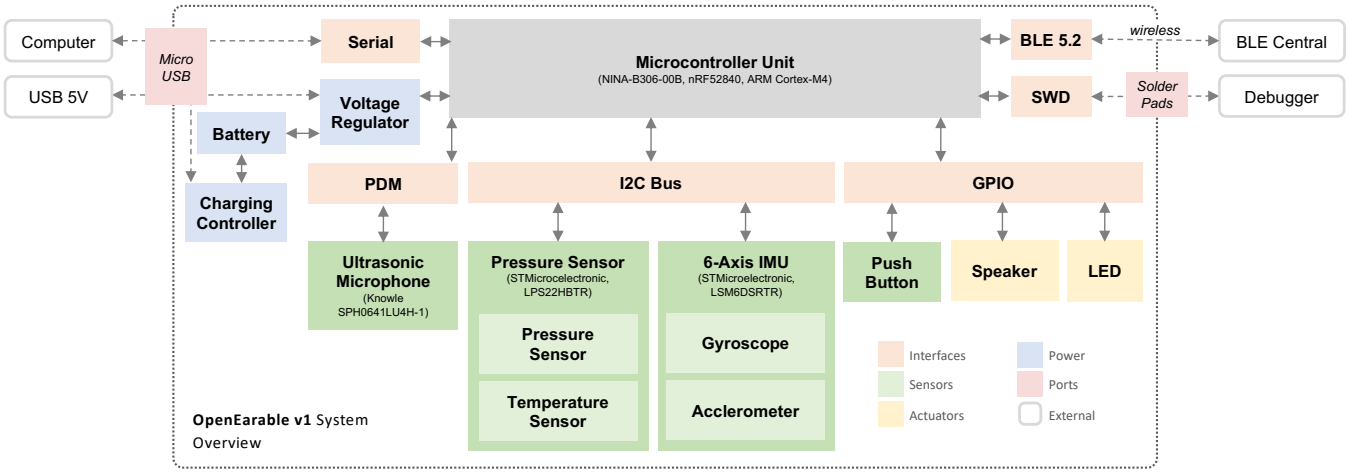


Figure 1: System overview of the *OpenEarable* system architecture. The microcontroller unit is the central hub which communicates with sensors, actuators, and external devices.

2 DESIGN PROCESS

In the following, we describe our guiding design principles and rationalise the sensor selection for the first version of *OpenEarable*.

2.1 Guiding Principles

Our main objective when developing the *OpenEarable* platform was to provide a general-purpose hardware sensing platform for the earable research community that allows for the exploration of state-of-the-art sensing capabilities on the ear. We were guided by the following principles throughout the design and development process:

2.1.1 Openness and Extensibility. The *OpenEarable* platform's hardware and software should be open to, and easily extensible by, others. The *OpenEarable* platform should provide the core infrastructure to enable the exploration of different sensing paradigms. As a result, all hardware design files, firmware, communication interfaces, and data recording tools should be made public and easily accessible so that others can modify and expand the platform in unique and novel ways. We also made a conscious effort to use development tools that are free-of-charge in the design of the hardware and software. We believe that as many people as possible should be provided with the opportunity to develop on, and for, the *OpenEarable* platform.

2.1.2 Manufacturability and Cost-Effectiveness. In order for people to leverage the *OpenEarable* platform for research, they must be able to easily manufacture the device at an affordable cost. To achieve this, we focused on commercial off-the-shelf components that require no specialized tools for manufacture. The PCB was specifically designed to be manufactured, and components assembled, by a self-service PCB assembly manufacturer. Additionally, all plastic parts are designed to be 3D-printable with a standard fused deposition modeling (FDM) printer, commonly available as consumer 3D printers or available to order online. The assembly of an *OpenEarable* should only require minimal equipment, with the first version only requiring a soldering iron, pliers and plastic-compatible power glue. All hardware components should be compatible with the open-source Arduino platform.

2.1.3 Attachment and Comfort. The *OpenEarable* platform, and any extensions, will need to be validated with users and therefore it should be easy to attach yet stable and robust against user movement. In addition, the earable should be comfortable to wear within the limitations of a general purpose prototyping device. For the first version of our design, we use an over-the-ear hook design that wraps around the auricle to encapsulate the electronics whilst providing mechanical stability. This provides an opportunity for sensors to be placed in, on, or around the ear.

2.2 Sensor Selection

For the first version of the *OpenEarable* platform we decided to incorporate both traditional and new sensing capabilities not currently available on other earable platforms. For basic input there is a push button, and a six-axis inertial measurement unit (IMU) for motion-based applications (e.g., gait analysis [2]) and to filter out motion artifacts.

For new sensing capabilities we chose to incorporate an ultrasound microphone and an in-ear pressure sensor combined with temperate sensor. Many earable platforms feature access to an external microphone for voice-based interaction, and most have a microphone inside the earbud for noise-cancellation. However, we could not identify an available platform that provides access to a microphone placed inside the ear canal. Therefore, *OpenEarable* features an inward facing ultrasonic microphone which can be used to detect ear canal shape and deformation based on measured sound reflection. We chose an ultrasound microphone to be able to detect both audible and inaudible sounds which do not disturb the wearer. We also incorporated a pressure sensor for in-ear barometry which provides information about the ear canal shape and deformation. In-ear barometry has gained traction in recent years across a range of applications and have been used to detect jaw and facial movements [1], blood pressure [10], and contraction of the

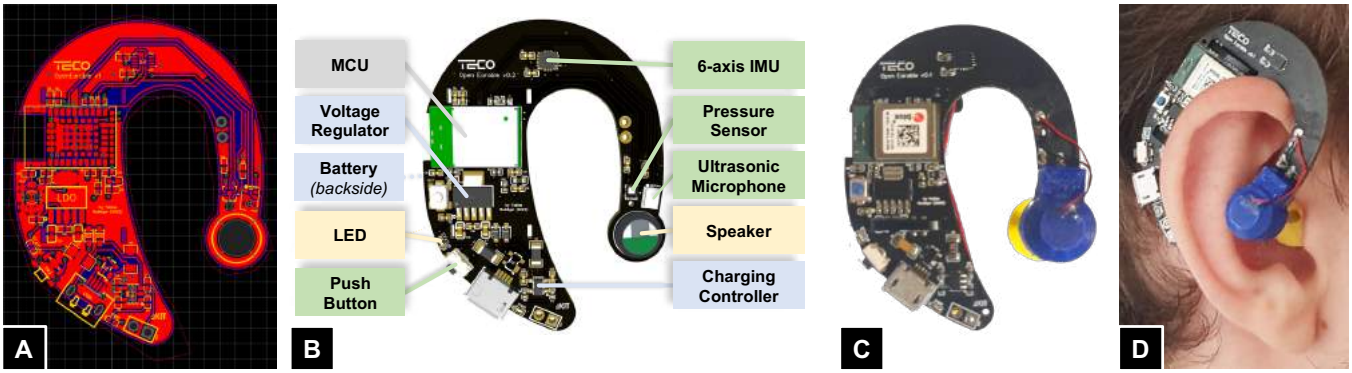


Figure 2: (A) PCB layout of *OpenEarable*; (B) 3D-rendering of the PCB and components; (C) assembled device with 3D-printed parts and battery; (D) a person wearing the device.

tensor tympani muscle (small muscle that actuates the eardrum) which can be used for interaction [7].

3 HARDWARE

The hardware of *OpenEarable* is inspired by existing works in the earable domain. We present the electronics, mechanical design, and production process. Step-by-step manufacturing instructions are available on the *OpenEarable* project's website [3].

3.1 Electronics

The following section describes the circuit layout, microcontroller unit, power architecture and sensors of *OpenEarable*. A schematic system architecture overview is shown in Figure 1.

3.1.1 Printed Circuit Board. We designed the *OpenEarable* PCB in an ear-hook form factor which makes it easy to attach the device to the ear. In addition, the shape of the PCB creates sufficient space to place all components behind the ear comfortably as this location was found to be most acceptable to place rigid components [8]. The PCB is 1.6 mm thick and is designed so that all surface mount device (SMD) components are on the top side only which simplifies assembly and makes it possible to have the components placed and soldered by a self-service PCB assembly manufacturer. Two holes in the PCB are designed specifically to let air and sound pass to the pressure sensor and ultrasonic microphone. In addition, the PCB has four holes for zip ties that hold the battery in place, and one large hole for the speaker. The design files of the PCB are open-source and released under a CC-BY license.

3.1.2 Microcontroller Unit. The microcontroller unit (MCU) of *OpenEarable* is a u-blox NINA-B306-00B module (NINA-B306-01B also compatible) which is based on the nRF52840 Bluetooth Low Energy (BLE) 5.2 system on chip (SoC). *OpenEarable* makes use of the Inter-Integrated Circuit (I2C) interface to communicate with the pressure and temperature sensor as well as accelerometer and gyroscope. The digital pulse density modulation (PDM) interface is used to read the microphone. Programming the MCU is possible using USB Serial or via Serial Wire Debug (SWD) on the backside of the PCB (e.g., to initially flash the USB Device Firmware Upgrade bootloader).

[3] *OpenEarable* project website - https://open-earable.teco.edu/

3.1.3 Power Architecture. In general, *OpenEarable* is intended to run from a single LiPo battery cell (*Renata ICP501230PS-03*, 135 mAh nominal capacitiy, 3.7 V nominal voltage). Charging is possible via a micro USB port with electrostatic discharge protection. For battery charging the board uses the *Microchip Technology MCP73831T* charging controller. As the MCU operates at 3.3V, *OpenEarable* also comes with a low dropout voltage regulator (*Texas Instruments TPS73733*). It is possible to also use the device while charging. When sampling all sensors and sending out the data via Bluetooth Low Energy a fully charged *OpenEarable* lasts roughly 10 hours which is well above the threshold for most research .

3.1.4 Ultrasonic Microphone and Speaker. An ultrasonic microphone (*Knowles SPH0641LU4H-1*) with bottom port is placed in close proximity above the speaker. By default, *OpenEarable* samples the microphone at approximately 44 kHz. The speaker inside *OpenEarable* is a standard true wireless stereo (TWS) 8 mm, 16 Ohm resistance earbud component that is available from many consumer electronics stores.

3.1.5 Pressure and Temperature Sensor. Pressure and temperature are measured in close proximity to the speaker and ultrasonic microphone. A hole in the PCB next to the pressure sensors redirects airflow from inside the ear canal. The pressure and temperature information are available from a single package inside the *STMicroelectronic LPS22HBTR* pressure sensor. The sensor can be configured to sample from 1 up to 75 Hz in an absolute pressure range of 260 to 1260 hPa. The temperature sensor supports a similar sampling rate range and has an absolute accuracy of ±1.5 °C.

3.1.6 Accelerometer and Gyroscope. *OpenEarable* has a 6-axis IMU (*STMicroelectronics LSM6DSRTR*) compromising of a 3-axis digital accelerometer and 3-axis digital gyroscope. Linear acceleration measurement range and angular measurement range can both be configured. By default, *OpenEarable* uses ±4 g linear acceleration range and ±500 dps angular rotation range. In theory, *OpenEarable* supports 1 Hz up to 6667 Hz accelerometer and gyroscope data. Limited by BLE bandwidth, *OpenEarable* currently reliably supports up to 104 Hz.

3.1.7 Light Emitting Diodes. *OpenEarable* features two LEDs for basic output. One static LED indicates the charging status when the

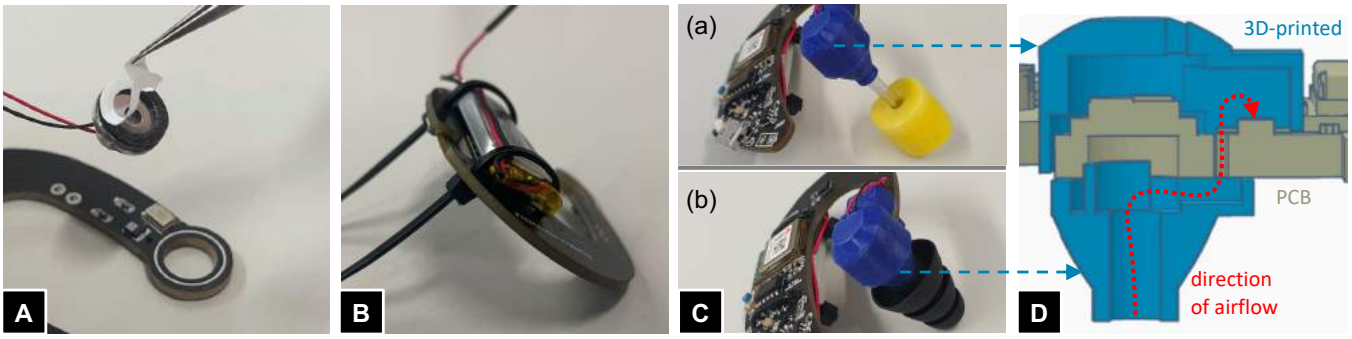


Figure 3: (A) Removing protective layer from adhesive foam of the speaker to glue it on the PCB; (B) Attaching the battery with zip ties through holes in the PCB; (C) 3D-printed shell of the earplug in two different configurations; (D) cross-section view of the earpiece illustrating the redirection of airflow and sound.

micro USB cable is plugged in (on: charging, off: fully charged or not charging). The second LED can be turned on and off or controlled in brightness using pulse-width modulation (PWM).

3.1.8 Push Button. A push button on the lower backside of *OpenEarable* can be used for simple, binary input. Another push button next to the MCU serves as reset button and can be used to enter the device firmware updates mode of the microcontroller by double pressing the button.

3.2 Mechanical Design and Assembly

The *OpenEarable* PCB is an integral part of the design as it functions as earhook. The assembly of the PCB was done by a contract manufacturer, see subsection 3.3. The parts that have to be self-assembled are described below.

3.2.1 Speaker and Battery. The speaker is an adhesive foam ring pre-installed so it can be glued onto the PCB while also sealing off the speaker (see Figure 3 A). The battery is attached onto open earable using 2 mm wide zip ties (see Figure 3 B).

3.2.2 Earplug. The *OpenEarable* earplug consists of two 3D-printed parts which are glued together and sealed off with PLA-friendly glue (Pattex instant glue). The front part sits above the speaker and the PCB through-hole and seals it of. Either, a foam type sealing earplug with plastic tube (*Etymotic Research* disposable eartip ER1-14A, 13mm diameter, see Figure 3 C (a)) to maximise ear canal sealing, or a triple flange conical silicone standard eartip can be put on the earplug (see Figure 3 C (b)). The backside separates the speaker cables and pressure sensor as well as microphone. Together, the 3D printed parts ensure that the ear canal is sealed for pressure sensing.

3.3 Production and Costs

OpenEarable was designed with the JLCPCB [4] parts library in mind. Therefore, almost all components are available as standard self-service SMD parts assembly order. The MCU and microphone have to be ordered specifically for assembly which, from our experience with JLCPCB, can two weeks lead time (depending on supplier availability) following which PCB manufacturing and assembly require an additional working week. For the 3D-printed parts we used an *Ultimaker 3*, however, there are also many inexpensive online 3D-printing services available that could be used to manufacture the earpiece plastic parts made-to-order.

The total costs excluding shipping for ten *OpenEarable* is roughly $400 (≈ $40 per device). The costs per device are split as follows: $0.50 PCB, $36.50 electric components, $0.10 zip ties, $4.10 foam earpieces (incl. 4 replacements), and $0.10 3D-printed parts. One-sided PCB assembly is free of charge.

4 SOFTWARE

All *OpenEarable* software is open-source and available on the project website under the MIT license.

4.1 Firmware

The *OpenEarable* firmware is implemented in C++ using the Arduino framework based on the implementation of the *Arduino Nano 33 Bluetooth Low Energy (BLE) Sense*. This makes it easily possible for others to change the firmware running on the device. The firmware reads out all sensors and makes them available via BLE. Due to bandwidth limitations, at least BLE 4.2 has to be supported by the device connecting to *OpenEarable*.

4.1.1 Generic Attribute Profile Specification. *OpenEarable*'s main interface for data transfer is a custom-defined Generic ATTribute Profile (GATT). Based on the profile, various functionalities of the earable can be controlled as well as sensors can be configured and read out. Table 1 gives an overview of the GATT specification for *OpenEarable* for regular data recording, as well as for recording and sending audio data. The sensor service is responsible for enabling sensors, configuring sampling rates and sending out sensor data. Using the device info service, a unique name and the device generation can be read out. The dedicated audio service sends out bursts of audio samples of roughly 1 second duration sampled at 62.5 kHz. At the moment, continuous audio streaming is not supported, however this is a software limitation that will be fixed in a future iteration (see subsection 5.4).

[4] JLCPCB - https://jlcpcb.com/

Table 1: BLE GATT profile services and characteristics overview of *OpenEarable*. A detailed documentation including UUIDs of the BLE API can be found on the project's website. The specification follows the schema for usage with edge-ml.org.

Service	Characteristic	Read / Write / Notify	Description
sensorService	sensorDataCharacteristic	Read / Notify	timestamped data of the different sensors
	sensorConfigCharacteristic	Write	enable sensors and configure sampling rate
deviceInfoService	deviceIdentifierCharacteristic	Read	unique identifier name of the device
	deviceGenerationCharacteristic	Read	generation of the device
audioService	audioCharacteristic	Read / Notify	burst chunks of ultrasonic audio data
	packageInfoCharacteristic	Read / Notify	info about total package amount and sending state

4.2 Recording Tool

Two options are available to record data with *OpenEarable*, a custom-built dashboard and an open-source and browser-based toolchain for machine learning on microcontrollers.

4.2.1 OpenEarable *Dashboard.* To make it easy to get started with recording sensor data, we have developed a dedicated dashboard for *OpenEarable* (see Figure 4). Users can connect to the device via their browser, configure sampling rates, enable sensor streams and record as well as export sensor data as CSV files.

Figure 4: *OpenEarable* dashboard that lets users configure sampling rates, enable sensors, and record data via WebBLE.

4.2.2 edge-ml. Out of the box, the *OpenEarable* firmware supports edge-ml[5], which is an open-source and browser-based toolchain for machine learning on microcontrollers. It offers recording, dataset management and labeling features. Using the default firmware installed on *OpenEarable*, users can simply connect to the device via WebBLE in their browser via edge-ml. In addition to data collection and labeling, it is also possible to train, validate, and export embedded machine learning models for *OpenEarable* using the edge-ml toolchain.

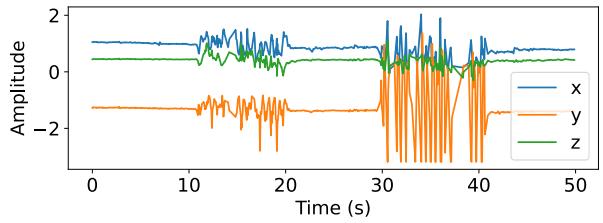
5 APPLICATION EXAMPLES

To gain an understanding that the *OpenEarable* platform is outputting valid data we used three example application scenarios from the earable literature.

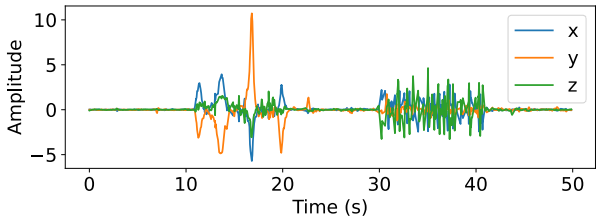
[5]edge-ml - https://edge-ml.org

5.1 Motion Tracking

Measuring motion on the ear is a common application in the earable space which can be used for a number of applications [6]. Figure 6 shows accelerometer and gyroscope readings for a test in which we recorded data from a single subject performing a sequence of three activities: (1) standing still, (2) walking, and (3) jumping jacks. We chose these activities to elicit distinct patterns, and the jumping jacks allow us to validate the mechanical stability of the *OpenEarable* during vigorous motion. The activities were performed for 10 seconds in the following sequential order: stand, walk, stand, jumping jacks, and stand.



(a) Data obtained with the integrated accelerometer.

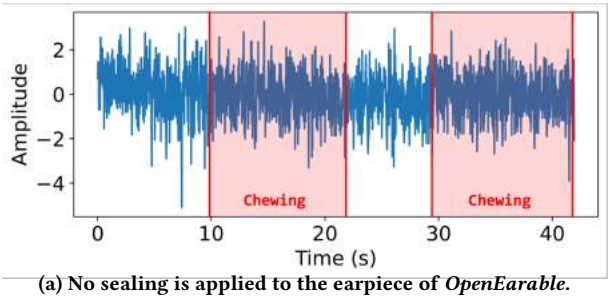


(b) Data obtained with the integrated gyroscope.

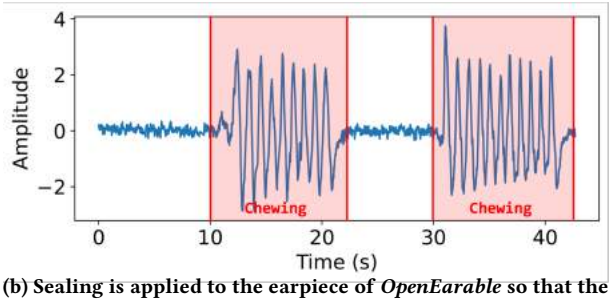
Figure 5: Motion activities recorded with *OpenEarable*. The following events can be seen in chronological order: Standing still, walking, standing still, jumping jacks and standing still. The data was z-normalized before plotting.

5.2 Ear Canal Pressure

A popular ear canal pressure application is the detection of jaw motions [1]. Figure 6a shows two sequences of chewing activities, with a break in-between. The importance of an air tight ear canal can be seen with the chewing events clearly visible when using the sealed ear buds. The distinct pressure signal demonstrates the feasibility of in-ear barometry using the *OpenEarable* platform.



(a) No sealing is applied to the earpiece of *OpenEarable*.



(b) Sealing is applied to the earpiece of *OpenEarable* so that the pressure sensor is air tight with the ear canal.

Figure 6: A sequence of ear canal pressure changes including chewing and not chewing with (a) a triple flange conical silicone standard eartip, and (b) *Etymotic Research* disposable eartip ER1-14A. The data was z-normalized before plotting.

5.3 Ear Canal Sound Reflections

It is possible for the ultrasonic microphone to pick up an inaudible signal from the speaker. This information can be used to understand the shape of the ear canal because the sound is reflected differently depending on the shape, a principle which can be used for authentication [9]. While more detailed evaluations are necessary to assess generalized authentication performance based on *OpenEarable*, Figure 7 shows the spectogram of a 18kHz tone played in the ear canal for 1 second.

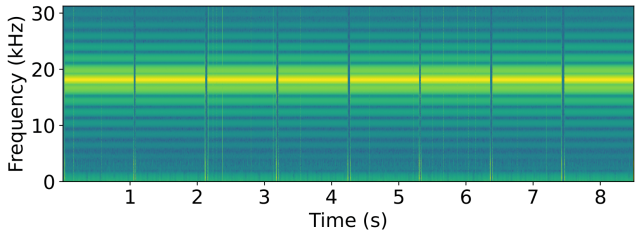


Figure 7: Spectrogram of a reflected ultrasonic signal which was emitted into the ear canal with multiple 1s long probings.

5.4 Future Work

As this is the first version of *OpenEarable*, there are a number of limitations with the prototype and improvements to be made. The current 6-axis IMU was selected due to stock limitations, and future versions will feature a 9-axis IMU with magnetometer. The earable is currently designed to be worn on one the right ear only and can not be paired with a second device. Currently, there are no libraries currently available for recording data from the *OpenEarable* platform on either Android or iOS devices which is a high priority item considering popular use cases of earable devices. While transferring a continuous audio signal over BLE is technically feasible, it is not yet implemented in the current *OpenEarable* firmware and the speaker only supports playback of a constant frequency. The Bluetooth classic advanced audio distribution profile (A2DP) is not yet supported. The standard ArduinoBLE library with current configuration achieved a transmission rates of 6.5 kB/s for the audio signal, we intend to use the NimBLE library, however, just recently support for the nRF52840 was added and compatibility with the bootloader of *OpenEarable* is pending but under active development. *OpenEarable* does not support reading out the battery level

6 CONCLUSION

OpenEarable is the first-if-its-kind open hardware initiative for earable research. In this paper we introduce a new device that features a series of sensors and actuators: a 3-axis accelerometer and gyroscope, an ear canal pressure and temperature sensor, an inward facing ultrasonic microphone as well as a speaker, a push button, and a controllable LED. We have shown the validity of the hardware based on three example application scenarios. Regarding the future development of *OpenEarable*, we are looking for feedback from the community and hope to bring parties together that are interested in pushing the platform further as a joint research effort. To stay up-to-date about the latest developments around *OpenEarable* we ask readers to refer to our project's website.

REFERENCES

[1] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. Canalsense: Face-related movement recognition system based on sensing air pressure in ear canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 679–689.

[2] L Atallah, A Wiik, B Lo, JP Cobb, AA Amis, and GZ Yang. 2014. Gait asymmetry detection in older adults using a light ear-worn sensor. *Physiological measurement* 35, 5, N29.

[3] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 2019. ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. In *The 25th annual international conference on mobile computing and networking*. 1–17.

[4] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 384–396.

[5] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Allesandro Montanari. 2018. Earables for personal-scale behavior analytics. *IEEE Pervasive Computing* 17, 3 (2018), 83–89.

[6] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 135.

[7] Tobias Röddiger, Christopher Clarke, Daniel Wolffram, Matthias Budde, and Michael Beigl. 2021. EarRumble: Discreet Hands-and Eyes-Free Input by Voluntary Tensor Tympani Muscle Contraction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[8] Tobias Röddiger, Christian Dinse, and Michael Beigl. 2021. Wearability and Comfort of Earables During Sleep. In *2021 International Symposium on Wearable Computers*. 150–152.

[9] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27.

[10] Jennifer Zeilfelder, Matthias Diehl, Christian Pylatiuk, and Wilhelm Stork. 2019. Concept for a Permanent, Non-Invasive Blood Pressure Measurement in the Ear. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 1–4.

A Preliminary Study for Detecting Visual Search Behaviors During Street Walking Using Earable Device

Kazuki Shimojo
The University of Tokyo
Tokyo, Japan
shimojo@mcl.iis.u-tokyo.ac.jp

Zengyi Han
The University of Tokyo
Tokyo, Japan
hzy@mcl.iis.u-tokyo.ac.jp

Yuuki Nishiyama
The University of Tokyo
Tokyo, Japan
yuukin@iis.u-tokyo.ac.jp

Kaoru Sezaki
The University of Tokyo
Tokyo, Japan
sezaki@iis.u-tokyo.ac.jp

ABSTRACT

Map applications on smartphones are powerful navigation tools for walking among places to visit for the first time and are widely used. On the other hand, checking the map applications tend to cause accidents on the road such as collisions with people, cars, and objects. To prevent this, we need to detect a walker's context regarding visual search behaviors and provide appropriate navigational information to the walker. In this paper, we propose a method to detect a walker's context regarding visual search behaviors by using motion sensors on an earable device. We collected and investigated motion and gaze data from an earable device and a gaze tracker respectively during street walking from five participants. Based on the investigation, we created a machine learning model for detecting looking around, smartphone, or normal during walking and stopping conditions. Our evaluations showed that our models can detect more than 95% walking and stopping conditions, and 71% of three detail conditions during walking, respectively.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile devices**.

KEYWORDS

Earable device, mobile sensor, behavior recognition, navigation, machine-learning

ACM Reference Format:
Kazuki Shimojo, Zengyi Han, Yuuki Nishiyama, and Kaoru Sezaki. 2022. A Preliminary Study for Detecting Visual Search Behaviors During Street Walking Using Earable Device. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3544793.3563416

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09…$15.00
https://doi.org/10.1145/3544793.3563416

1 INTRODUCTION

In recent years, wearable devices are being widely used on the arm, head, or abdomen [4, 9, 11]. Using the different sensors mounted on such devices, it has become possible to easily, and in detail, detect and collect information from daily activities, such as heart rate, number of steps, and calorie expenditure. In addition, wearable devices have become smaller, lighter, and less expensive. It is anticipated that we will use wearable devices in daily life more than ever before. Among wearable devices, those that are worn on the ear are called "Earable Devices" [4]. When earable devices are equipped with a motion sensor and a microphone and when these capabilities are combined with voice recognition functions on a smartphone, a new hands-free user assistance service can be realized [8]. For example, if earable devices can detect a walker lost in a city, it can be possible to realize efficient and safe route guidance by combining it with voice guidance. On the other hand, to realize this service, a high degree of accuracy in behavior recognition is important.
The purpose of this study is to verify whether data collected from motion sensors mounted on earable devices can be used to correctly detect a walker's state while street walking. If the walker's walking condition can be correctly detected using earable devices, it will lead to the realization of new services such as voice navigation at the appropriate time. The contributions of this paper are as follows:

- The feasibility of using earable devices in detecting a walkers' visual search behavior during street walking is explored.
- Motion-sensor data are labeled using Tobii's eye gaze recordings.
- Machine learning models are used to classify whether a walker is walking or stopping from motion-sensor data collected by earable devices.

2 RELATED WORKS

Research has been conducted to estimate user behavior by using sensors in wearable devices, in addition to smartphones. Earable devices, which are devices worn on ears like AirPods, are expected to become important in the development of wearable devices in the future. Earable devices will provide various services such as continuous sensing of human behavior, the realization of AI (augmented reality) through sound, information transmission by AI voice assistants such as Alexa and Siri, and tracking of health status. To recognize behavioral and physical activities through earable devices,

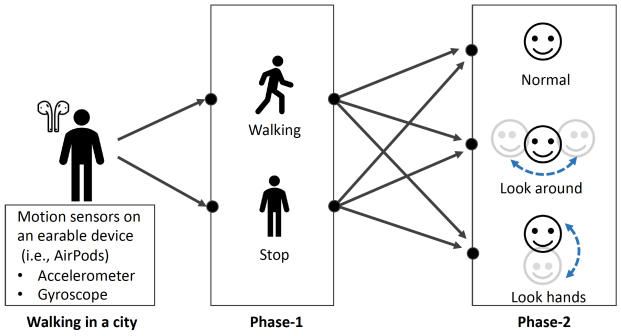


Figure 1: Overview of a method for detecting visual search behaviors

the aggregation of collected open data is essential. In 2018, Nokia Bell Labs at the University of Cambridge developed eSense [4], a research and development platform, to facilitate the activation of research on earable devices.

Several studies have focused on behavior recognition through earable devices [1–3, 5]. Compared with a smartphone, which is shaken in the pocket, and a smartwatch, which is affected by arm swing, an earable device is more stable due to being on the ear. The accuracy of detecting head movement with earable devices using just an accelerometer has shown to be higher than using both an accelerometer and a gyroscope together [6]. These results indicate the usefulness of sensing design in earable devices. In addition, by leveraging the superiority of earable devices in accurately tracking the state of a head, some studies have shown that an inertial navigation system using an inertial measurement unit in an earable device could enable voice navigation on behalf of a visually impaired person [1].

3 VISUAL SEARCH DETECTION

In this section, we present a detailed framework and classification method.

3.1 Overview

Figure 1 illustrates the overview of our approach for detecting visual search behaviors while walking in a city. As illustrated, we designed a two-phase approach for detecting visual search behaviors. First, our method detects whether a walker is walking or has stopped. After that, our method classifies three types of visual search behaviors: look around, look hands (i.e., smartphone), and normal.

We categorized the walker's visual search behaviors as follows: First, we classified whether participants were walking or stopping, and each of the two states were further subdivided into three states. As a result, six types of states were defined.

- *walking* (normal walking.)
- *look_sp_while_walking* (participant walking while using their smartphone.)
- *look_around_while_walking* (participant looking around while walking.)
- *stop* (participant standing, such as waiting for a traffic light.)

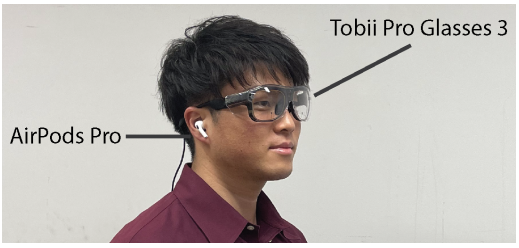


Figure 2: Wearing AirPods Pro and Tobii Pro Glasses 3

- *look_sp* (participant standing still while using their smartphone.)
- *look_around* (participant standing still while looking around.)

3.2 Devices

Participants wore AirPods Pro[1], as it is an earable device with a motion sensor to detect a walker's state. AirPods Pro is equipped with an accelerometer and a gyro sensor that can collect motion data in real-time via the iOS CoreMotion API. In this experiment, motion data from AirPods Pro was continuously collected and stored by using the AWARE Framework for iOS [2] [7]. We collected motion sensor data at 30 Hz and saved it in CSV format. In addition to AirPods Pro, subjects wore Tobii Pro Glasses 3 [10] to track the walker's gaze in real time because this experiment required classification and labeling of the walker's gaze during street walking. Tobii Pro Glasses 3 is an eye tracking device that is widely used in visual-related scientific research. Figure 2 shows an example of a participant wearing AirPods Pro and Tobii Pro Glasses 3.

3.3 Classification Method

For the classification of two states, we created multiple machine learning methods. Specifically, we usd five types of machine learning models: K-nearest neighbor method, logistic regression, gradient boosting, random forest, and multi-class support vector machine (SVM). These machine learning models were implemented using Scikit-learn[3]. In addition, we performed cross-validation on all the machine learning models, thereby creating five patterns with one of the five subjects as the test data and the other four as the training data. We then evaluated each evaluation index using the average value.

4 EVALUATION

In this section, we describe the data collection process, and then we evaluate the feasibility of detecting the user's state using motion-sensor data obtained from an earable device.

4.1 Data Collection

Five male college students participated in the data collection process (IRB-approved). Each participant wore the devices, as shown in Figure 2, and walked along the designated route. Figure 3 shows the route of the data collection, which is located in an intricately

[1]https://www.apple.com/jp/airpods-pro/
[2]https://github.com/tetujin/AWAREFramework-iOS
[3]https://scikit-learn.org/

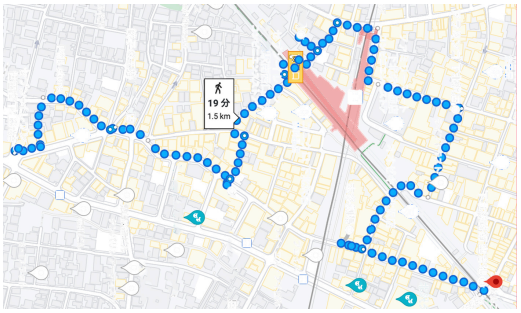


Figure 3: Route map of the street walking experiment.

Figure 4: Walker's gaze recorded by Tobii Pro Glasses 3.

laid out residential neighborhood with streets that are hard to learn to navigate correctly. Six destinations were set between the start and finish points and the walking time was designed to take approximately 20 minutes.

Participants were given a map of the route through six destinations and they were asked to walk the route accordingly. Over 20 minutes of walking, we collected the motion data and eye gaze recordings of participants. By having the participants walk along the route, we aimed to collect motion data as naturally as possible. Figure 4 shows an example image of video and gaze recordings using Tobii Pro Glasses 3.

4.2 Evaluation 1: Detecting Walking and Stopping Conditions

First, we counted the number of each of the two states per subject. The highest number was approximately 26,000 samples of motion-sensor data and the lowest was approximately 1,300. To align with the lowest number, 1,280 samples were randomly selected for each subject state. We divided the data into 64 samples and calculated the mean value, variance, maximum value, minimum value, and standard deviation of each sample.

Walking and stopping could be precisely classified by of five types of machine learning models. The f1-scores are shown in Table 1. It is apparent taht the random forest model is most suitable for use. The values of the random forest model are 0.95, 0.97, 0.95, and 0.95 in the order of accuracy, precision, recall, and f1-score.

Table 1: F1-scores of five types of machine learning

Machine Learning Model	F1-score
random forest	0.950
gradient boosting	0.950
logistic regression	0.843
multi-class SVM	0.515
K-nearest neighbor	0.495

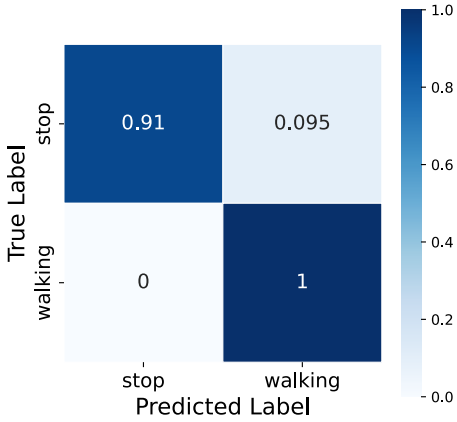


Figure 5: Estimation accuracy of random forest. The vertical axis represents the true label, and the horizontal axis is the predicted label.

4.3 Evaluation 2: Detecting Detailed Conditions During Walking and Stopping

After classifying the data as walking or stopping, the next step was to classify the motion-sensor data in detail. There are three detailed states to walking, as defined in the previous section: *walking*, *look_sp_while_walking*, and *look_around_while_walking*. We estimated the number of each of the three states per subject. The data of one of the subjects were much less than others; therefore, we used only four subjects' data. We used the five types of machine learning models. 512 samples were randomly selected, and we divided the data into 64 samples. Regarding stopping, classifying into these three states: *stop*, *look_sp*, *look_around*. We evaluated the data in the same way. 128 samples were randomly selected and divided into 16 samples.

The results of classifying each walking and stopping into three states were not sufficient for the effectiveness of the machine learning model. About the result of classifying walking into three detailed states, the random forest model exhibited the best performance: the f1-score was 0.66, and the recall was 0.71. The recall values are shown in Figure 6. All of the machine learning models performed poorly while classifying the stopping condition. The results of the random forest model, the best performer, are shown in Figure 7.

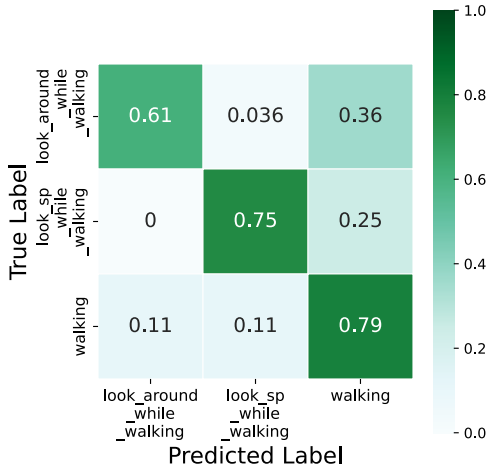


Figure 6: Estimation accuracy of three states of walking

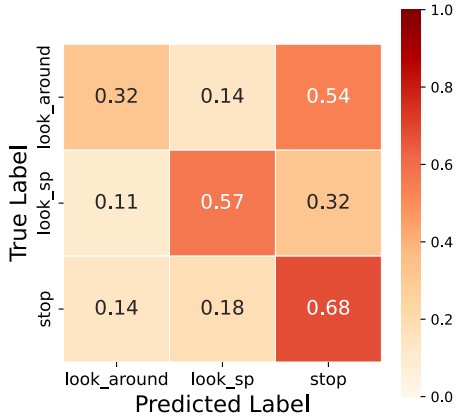


Figure 7: Estimation accuracy of three states of stopping

5 DISCUSSION

There are two reasons why all machine learning models performed poorly in classifying each walking and stopping into three detailed states. First, the amount of data is not adequate for the experiment. The number of subjects need to be increased to gather more motion data. In particular, the stopping time was much shorter than walking time. Therefore, it may be necessary to collect data not only in a natural state but also in a controlled state. Second, we should reconsider the execution of the machine learning models. We used data picking randomly. However, it is very important to deal with the waveform as a characteristic. For example, when subjects turn around, a unique waveform of the rotation sensor recordings appears. Therefore, the method of using a sliding window or bump detection may be effective for keeping the characteristics of the motion data.

6 CONCLUSION

Focusing on the realization of navigation services for street walking with earable devices, we investigated the possibility of classifying user visual search behavior. We evaluated the feasibility of classifying a walker's visual search behavior by conducting a preliminary study. The results showed that whether a subject is walking or stopping can be detected by machine learning models and that the random forest model especially had a high performance of 0.95 and approximately 1.0 in terms of f1-score and recall, respectively. However, the classification of walking and stopping remains a problem, and it is necessary to realize more detailed classification. We will collect a larger amount of data and find a more appropriate method for detection in the future. Also, we will provide appropriate navigation information based on a walker's context detection.

ACKNOWLEDGMENTS

This work was partly supported by NICT, Japan (01101).

REFERENCES

[1] Ashwin Ahuja, Andrea Ferlini, and Cecilia Mascolo. 2021. PilotEar: Enabling In-Ear Inertial Navigation. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (Virtual, USA) *(UbiComp '21)*. Association for Computing Machinery, New York, NY, USA, 139–145. https://doi.org/10.1145/3460418.3479326

[2] Md Islam, Tahera Hossain, Md. Atiqur Rahman Ahad, and Sozo Inoue. 2021. *Exploring Human Activities Using eSense Earable Device*. 169–185. https://doi.org/10.1007/978-981-15-8944-7_11

[3] F. Kawsar, C. Min, A. Mathur, and A. Montanari. 2018. Earables for Personal-Scale Behavior Analytics. *IEEE Pervasive Computing* 17, 03 (jul 2018), 83–89. https://doi.org/10.1109/MPRV.2018.03367740

[4] Fahim Kawsar, Chulhong Min, Akhil Mathur, Alessandro Montanari, Utku Günay Acer, and Marc Van den Broeck. 2018. ESense: Open Earable Platform for Human Sensing. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 381–383. https://doi.org/10.1145/3267305.3267640

[5] 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* (Reims, France) *(AH2019)*. Association for Computing Machinery, New York, NY, USA, Article 25, 4 pages. https://doi.org/10.1145/3311823.3311869

[6] Chulhong Min, Akhil Mathur, and Fahim Kawsar. 2018. Exploring Audio and Kinetic Sensing on Earable Devices. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications* (Munich, Germany) *(WearSys '18)*. Association for Computing Machinery, New York, NY, USA, 5–10. https://doi.org/10.1145/3211960.3211970

[7] Yuuki Nishiyama, Denzil Ferreira, Yusaku Eigen, Wataru Sasaki, Tadashi Okoshi, Jin Nakazawa, Anind K Dey, and Kaoru Sezaki. 2020. iOS Crowd-Sensing Won't Hurt a Bit!: AWARE Framework and Sustainable Study Guideline for iOS Platform. In *Distributed, Ambient and Pervasive Interactions*, Norbert Streitz and Shinichi Konomi (Eds.), Vol. 12203. Springer International Publishing, Cham, 223–243. https://doi.org/10.1007/978-3-030-50344-4_17

[8] Yuuki Nishiyama and Kaoru Sezaki. 2021. Experience Sampling Tool for Repetitive Skills Training in Sports Using Voice User Interface. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (2021-09-21) *(UbiComp '21)*. Association for Computing Machinery, Virtual, USA, 54–55. https://doi.org/10.1145/3460418.3479283

[9] Ivan Miguel Pires, Nuno M. Garcia, Nuno Pombo, Francisco Flórez-Revuelta, and Susanna Spinsante. 2018. Approach for the Development of a Framework for the Identification of Activities of Daily Living Using Sensors in Mobile Devices. *Sensors* 18, 2 (2018). https://doi.org/10.3390/s18020640

[10] Tobii Pro AB. 2014. Tobii Pro Lab. http://www.tobiipro.com/ Computer software.

[11] Kai Zhan, Steven Faux, and Fabio Ramos. 2015. Multi-scale Conditional Random Fields for first-person activity recognition on elders and disabled patients. *Pervasive and Mobile Computing* 16 (2015), 251–267. https://doi.org/10.1016/j.pmcj.2014.11.004 Selected Papers from the Twelfth Annual IEEE International Conference on Pervasive Computing and Communications (PerCom 2014).