Ph.D. Thesis

Andrea Ferlini

Exploring the Potential of Earables for Personal-Scale Sensing

Churchill College
University of Cambridge

Email: af679@cl.cam.ac.uk

September 15, 2022

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. It does not exceed the prescribed word limit of 60,000 words for the Computer Science Degree Committee, including appendices, footnotes, tables and equations.

Andrea Ferlini

September 15, 2022

Abstract

Earables (in-ear wearables) present a new frontier in wearables. Acting both as leisure devices, providing personal audio, and as computing and sensing platforms, earables can collect sensory data from the head. However, earables, more than the majority of other wearables (like smartwatches), have inherent size, shape, and usability constraints arising from their small form factor. These restrictions inevitably limit the number, type, and placement of the sensors that could be added to such platforms. Because of that, earables' full potential has yet to be truly unlocked. In this dissertation, we characterize and explore three different sensors that realistically will be integrated into future earables. With usability and feasibility constraints in mind, we investigate (i) in-ear, 9 degrees of freedom (DoF), inertial measurement units (IMU); (ii) microphones (facing in-the-ear-canal); (iii) photoplethysmography sensors (PPG). Data collected by earables augmented with these sensors in and around the ear may enable several personal-scale sensing applications such as augmented/virtual reality, improved navigation, medical rehabilitation, fitness tracking, identification, and health condition screening. This is facilitated by the human head being subject to fewer vibrations and random movement variations than the lower parts of the body, thanks to the inherent damping in the musculoskeletal system.

First, we study inertial sensing for head movements tracking. However, an absolute reference is key in order to re-calibrate IMUs and track absolute rotations, thus enabling more advanced applications. In mobile devices, this is done by coupling accelerometers and gyroscopes with a magnetometer. Though, as of today, no earables are equipped with a magnetometer. Hence, we investigate how to add one to an existing pair of earables (eSense). After characterizing the source of interference that would affect a magnetometer in an earable, and understanding how traditional calibrations fall short, we devise a user-transparent magnetometer calibration for earables. This sheds light on the potential of earables for both relative motion tracking as well for more advanced use cases such as

navigation.

The second part of this dissertation features another commodity sensor: the microphone. We focus on *in-ear facing microphones*, exploiting the unique positioning of earables. Unlike IMUs, not yet readily available in all consumer earables, in-ear facing microphones are already present in both high-end leisure earbuds (e.g., Apple AirPods Pro) and in hearing aids for noise cancellation purposes. Leaning on that, we research in-ear acoustic sensing for both motion sensing (step counting, hand-to-face gesture interactions, and human activity recognition) and user-identification (based on acoustic-gait tracking). This exploration paves the way to new interactions between users and earables, whilst increasing the security of sensory earables (through identification).

Finally, we investigate ear-worn photoplethysmography (PPG) sensing. PPGs, like IMUs and microphones, are easily integrated into an earable form factor. First, we identify the optimal sensor placement for ear-worn PPG sensing, looking both at a resting baseline as well as the impact of motion artifacts. Then, we focus on head movements and facial expressions that people perform naturally when wearing earables, causing skin and tissues displacements around the ear and inside the ear canal. Understanding such artifacts becomes key to the success of earables as the next wearable platform for accurate and reliable cardiovascular health monitoring.

The prototypes developed, the data collected, the analyses performed, and the insights drawn in this dissertation provide evidence of the potential of earables as a disruptive platform for mobile personal-scale sensing.

Acknowledgments

Pursuing a P.h.D. is surely not the easiest of endeavors, and writing this thesis has proven rather challenging in many aspects. Yet, here I am writing the only bit of this dissertation that people (other than my examiners and supervisor) will read. It is finally time to thank all the people that walked this journey alongside me and helped me go through the tough moments.

This thesis goes to my parents, Daniela and Marco; without them, I would not be here writing these very same words. Without your sacrifices and help, I would have never arrived where I am today. A special mention has to go to my grandparents, Dada (I wish you were still here witnessing this moment), Enrica, and Luigi, who always cheered for and believed in me. I love you.

First and foremost, I want to thank my supervisor, Cecilia Mascolo. Without your guidance, I would have never reached the academic maturity required to put together a P.h.D. dissertation. You taught me rigor and gave me a mental structure to follow. Thank you for your patience, encouragement, trust, and leadership. I also want to spend gratitude words for Robert Harle. When, in my first year, I had thought about quitting, your words of wisdom convinced me to endure. To Fahim Kawsar, your advice and mentorship during my internship were pivotal to my career development. I am grateful to Nokia Bell Labs for supplying the funds to carry out this research, along with Churchill College and the Computer Laboratory, whose support also made this work possible.

To Giovanni Pau, if I decided to pursue a P.h.D. it is all your fault. You know more than I should write. To Davide Giachi, the only high school professor who believed in me. Without your endorsement, I would have probably been lost long ago.

To Alessandro. You have been, and still are, a great mentor, colleague, and, most importantly, a special friend. To Lorena. Thank you for being like a second mother and an older

sister, thanks for the endless chats and bearing with me, and for my Blue Mondays. To all my other collaborators. A sincere thank you for your guidance, help, and support. A special mention to Andreas and Dong for their hard work and friendship. Andreas, working late nights with you gave me the strength to finish what I started. To all my other lab mates, thank you for being part of this journey and contributing to my growth as a person and researcher. A particular mention has to go to Dimitris, Erika, Ignacio, Kayla (best intern ever), Ting, and Young.

To all the friends I have met in Cambridge, Albane, Alex S., and Shyam in particular. To all the friends from Churchill College MCR. I will always be grateful to have the pleasure of calling you, friends. To Davide, I could not have asked for a better housemate.

To all my CUCC mates I have been logging miles with during the past two years. In particular, to Fabio, Matt, and Harry. To Alex P., you are one of the kindest people I have ever met. Our chats got me through one of the hardest moments of my life.

To the members of the One Men's Tree, having a second family when far from home is something I will never take for granted. Your friendship got me through isolation and the daily struggles of the past three years. Even now that we are inevitably parting ways, I am sure we will always be there for each other. I will miss you.

To Giorgio. If I listed all the reasons I have to thank you for, I would have to write another full manuscript. Without you, I would not be the person I am today. To my friends back home and in Paris, particularly Angelo, Davide, Elia, Emilie, Francesca, Francesco O., Francesco G., Giacomo I., Joel, Luca B., Luca C., Luca D.M., Margareth, Mattia, Matteo, Miriam, Riccardo, and Wei. It is good knowing there is always someone I can count on. To Alex G., thank you for coaching me on and, more often than not, off the bike.

Finally, Roxana, the love of my life. Thank you for everything. Thank you for supporting me each and every day. Thank you for helping me become a more rational person and a better human being. Thank you for loving me. Thank you for bearing with me, my anxiety, stress, and sleepless nights. I owe you more than you possibly imagine. Ti amo.

Contents

1 Introduction **1**

1.1 Challenges in earable research for personal-scale sensing 2
1.2 Thesis and substantiation 5
1.3 Contribution and chapter outline 6
1.4 List of publications 10

2 Related Work **13**

2.1 Inertial sensing 13
2.1.1 Head motion tracking 14
2.1.2 Magnetometer calibration 14
2.2 Acoustic motion sensing 16
2.2.1 Facilitating in-ear acoustic sensing: the Occlusion Effect 17
2.2.2 Activity recognition 18
2.2.3 Hand-to-face interactions 19
2.2.4 Gait tracking 19
2.3 Photoplethysmography sensing 21
2.3.1 Ear anatomy 21
2.3.2 Photoplethysmography primer 22
2.3.3 Wearable applications of PPG sensors 23
2.3.4 Motion artifacts on in-ear PPG 26
2.4 Summary 28

3 In-ear Inertial Sensing **29**

3.1 Introduction 29
3.2 Platform overview 32
3.2.1 eSense Platform 32

3.2.2 Challenges of inertial tracking . . . 32
3.3 Head tracking methodology . . . 33
3.4 User study and head motion tracking performance . . . 36
3.4.1 Data collection methodology . . . 36
3.4.2 Baseline: silent subject . . . 37
3.4.3 Impact of chewing activity . . . 39
3.4.4 Impact of Speech . . . 39
3.5 Why is the Magnetometer Missing? . . . 40
3.6 Enabling in-ear magnetic sensing: motivation . . . 43
3.7 Magnetometer Calibration and Interference . . . 45
3.7.1 Magnetometer Calibration . . . 46
3.7.2 Interference Characterization . . . 46
3.8 A Novel Calibration Algorithm . . . 49
3.8.1 Overview . . . 49
3.8.2 Calibration Approximation . . . 50
3.8.3 Calibration Algorithm . . . 52
3.9 Calibration evaluation . . . 54
3.9.1 Micro benchmarks . . . 54
3.9.2 In-the-wild example case-study: navigation . . . 58
3.9.3 Computational and power consumption considerations . . . 60
3.10 Discussion and limitations . . . 61
3.11 Conclusion . . . 62

4 In-ear Microphone Sensing: OESense 64
4.1 Introduction . . . 64
4.2 Motivation . . . 65
4.2.1 Preliminary exploration: accelerometers . . . 66
4.2.2 Preliminary exploration: traditional external microphones . . . 67
4.3 OESense: system design . . . 68
4.3.1 Overview . . . 68
4.3.2 Impact of the occlusion effect . . . 68
4.3.3 Initial exploration . . . 70
4.3.4 Sensing Pipelines . . . 71
4.4 Implementation . . . 74
4.4.1 Hardware prototyping . . . 75

4.4.2 Earbud fit test . . . 75
4.5 Data collection . . . 76
4.5.1 Step counting . . . 76
4.5.2 Human activity recognition . . . 77
4.5.3 Hand-to-face gesture recognition . . . 77
4.6 Evaluation . . . 78
4.6.1 Baselines benchmarking . . . 78
4.6.2 Step counting . . . 79
4.6.3 Human activity recognition . . . 80
4.6.4 Hand-to-face gesture recognition . . . 83
4.6.5 Impact of music playback . . . 85
4.6.6 Power and latency measurement . . . 86
4.7 Discussion and limitations . . . 87
4.8 Conclusion . . . 88

5 In-ear Microphone Sensing: EarGate 89
5.1 Introduction . . . 89
5.2 Preliminaries . . . 92
5.2.1 Out-ear microphone vs. in-ear microphone . . . 92
5.2.2 Feasibility exploration . . . 94
5.3 System design . . . 94
5.3.1 EarGate: system at a glance . . . 94
5.3.2 Signal processing . . . 96
5.3.3 Feature extraction . . . 98
5.3.4 Identification methodology . . . 98
5.4 Implementation . . . 99
5.4.1 EarGate prototype . . . 100
5.4.2 Data collection . . . 100
5.5 Performance Evaluation . . . 100
5.5.1 Metrics . . . 102
5.5.2 Training-testing protocols . . . 102
5.5.3 Overall performance . . . 104
5.5.4 Impact of data imbalance . . . 104
5.5.5 Impact of walking conditions . . . 105
5.5.6 Impact of human speech . . . 106

5.5.7 Impact of music playback . . . 107
5.5.8 Sensor multiplexing . . . 108
5.5.9 Impact of training size . . . 109
5.5.10 Transfer learning . . . 109
5.5.11 Contribution of specific features . . . 110
5.6 System considerations . . . 111
5.7 Discussion . . . 114
5.8 Conclusion . . . 115

6 In-ear Photoplethysmography Sensing 116
6.1 Introduction . . . 116
6.2 Ear anatomy and sensor placement . . . 120
6.3 PPG and vital signs extraction . . . 121
6.3.1 Vital signs estimation from PPG . . . 121
6.4 Experimental setup . . . 124
6.4.1 Study population . . . 124
6.4.2 Devices . . . 125
6.4.3 Data collection protocol . . . 125
6.4.4 Data analysis . . . 126
6.5 Benchmarks . . . 126
6.5.1 PPG placement . . . 126
6.5.2 Impact of motion artifacts . . . 130
6.6 Outlook: in-the-canal PPG . . . 132
6.6.1 Form factor and ear-tip design . . . 133
6.6.2 Signal processing pipeline . . . 133
6.7 Existing PPG motion artifacts datasets . . . 134
6.8 EarSet: methodology . . . 136
6.8.1 Study population . . . 136
6.8.2 Data collection devices . . . 137
6.8.3 Data collection protocol . . . 140
6.9 Results . . . 142
6.9.1 Dataset outlook and template matching . . . 142
6.9.2 Handcrafted metrics extraction and statistical analysis . . . 143
6.9.3 DNN-based motion artifacts classification . . . 144
6.10 Conclusion . . . 147

7 Final Remarks **149**
7.1 Summary of contributions 149
7.1.1 Head motion tracking and addition of a calibrated magnetometer 150
7.1.2 In-ear microphone-based general motion sensing and user identification 150
7.1.3 In-ear photoplethysmography: best location for in-ear PPG sensing and robustness to motion artifacts 151
7.2 Limitations and future directions 152
7.2.1 Dataset size 152
7.2.2 Machine learning algorithms 153
7.2.3 New modalities and forms 154

'So it begins.'

King Théoden

Chapter 1

Introduction

Since the first smartphone hit the market, we have witnessed a rapid rise in mobile computing and mobile sensing [1]. The widespread diffusion of wearable technologies that followed has ultimately led to the development of what is known as personal-scale sensing. Personal-scale sensing, or the collection of fine-grained individual-scale data for sensing applications, has become a key enabler of a number of applications to augment human perception and enhance experience. With smartwatches and smart-rings, it is now possible to collect an unprecedented wealth of data, at an individual-scale, from different locations on the human body. Personal-scale sensing relies on modern mobile and wearable devices to monitor fitness, human activities, vital signs, provide context-aware information, and to augment user experience and awareness in general.

Continuously tracking users' behaviors enables a variety of services and applications. These could be related to fitness, health, and mental well-being. Additionally, monitoring users over time could also facilitate cognitive assistance (with context-aware recommendations, navigation, and guidance), behavioral monitoring, augmented reality, seamless identification, and life-logging [2]. Users leverage their mobile devices (e.g., smartphones, smartwatches, smart-rings, smart-earbuds, glasses, etc.) equipped with multiple sensors (e.g., inertial measurement units, microphones, photoplethysmography, etc.) to measure their vital signs, keep a record of their daily activities, and get insights on how to improve their lifestyle.

Recent years have shown a new trend in wearables, with the steady diffusion of wireless head-worn devices, such as helmets, glasses, brain computer interfaces (BCIs), and earbuds

—forecasted to account for the largest share of the wearable market[1]. *Head-worn* devices offer fascinating sensing opportunities as they are positioned at an extremely promising vantage point on the body: the user's head. Sensing around the human sensoria (i.e., brain, ears, eyes, nose, mouth) could open the door to tracking various vital functions which could not be monitored through more conventional wearables like smartwatches. For instance, head-worn wearables could be useful to sense vital functions such as neurological activity [3], cardiovascular health [4, 5], hearing impairments, but also eating habits and dietary signs [6, 7]. Inevitably, the sensing capabilities of head-worn wearables lead to the birth of a new core component in the wearable personal-scale sensing ecosystem, thus far mostly composed of smartphones and smartwatches (plus a few, less common, other additions). With data coming from a new less-explored vantage point (i.e., the head), the granularity of the information the user can benefit from dramatically increases.

This dissertation will focus on augmenting and pushing the envelope of what it is possible to sense by leveraging *earables* (also known as smart-earbuds). Compared to existing devices commonly adopted to monitor people's health, fitness and activities, earables possess a number of advantages:

i. unlike smartphones that do not have a location known a priori on the body (in the hand of the user, left in whichever pocket of their jeans, etc.), earables, when worn, have a stable position in the user's ears;

ii. unlike smartwatches that suffer from continuous (random) wrist movements, earables are less susceptible to motion artifacts as vibrations are naturally dampened by the musculoskeletal system;

iii. finally, unlike other wearables worn at the extremities of the body, like smart-rings and pulse oximeters, earables are not affected by phenomena like vasoconstriction, which is often suffered in the fingers in cold conditions.

1.1 Challenges in earable research for personal-scale sensing

Earables are certainly a new comer in the wearable ecosystem. Nevertheless, they have the capability and potential to become staple platforms for personal-scale monitoring.

[1] https://www.statista.com/statistics/385658/electronic-wearable-fitness-devices-worldwide-shipments/

To augment earables with new sensors and thus unlock their true potential for personal-scale sensing, researchers are called to face a number of limitations. These, are further complicated by data analysis issues, tight systems constraints, and usability requirements. This section first surveys the general sensing and systems/data processing limitations that arise in today's earable research before focusing on the challenges specifically related to a subset of sensors critical to enable personal-scale applications.

Sensing: Enabling personal-scale applications means being able to sense and aggregate a multitude of different bio-signals simultaneously. Doing so on earables is still an open research problem. It is still unclear which types of sensors can be integrated in earables without altering their main functioning, i.e., audio playback. In particular, it is hard to quantify to what extent these new modalities can provide data of high-enough-quality to facilitate personal-scale applications. This stems from two main factors: (1) earables are inherently small in form factor, making it hard for the manufacturers to accommodate a set of different sensors; (2) due to the complex nature of human physiology and human activities, bio-signals are inevitably noisy, with measurements often having poor signal-to-noise ratio (SNR), ultimately resulting in low-quality data.

Systems/Data processing: Augmenting earables with multiple sensors opens the door to a number of interesting scenarios, however, it also introduces the difficulty of processing the continuous stream of data generated, without draining battery life. It then becomes key to craft dedicated solutions that can guarantee not only low latency (often *quasi-real-time*) in processing the collected data but, more importantly, ensuring minimal energy waste. The need for low power consumption stems from the fact that augmenting user experience through personal-scale sensing requires continuous and long-term monitoring of various bio-markers and human activities. Doing so inevitably demands high power consumption, quickly draining the device battery. For instance, existing commercially available wireless earbuds can, on average, only operate continuously for 4 to 5 hours, with a full charge. With the inclusion of more sensors, the need for more energy inevitably increases, thus leading to even shorter operational time (i.e., duration of a battery cycle before you charge the earbuds). Therefore, developing algorithmic pipelines which do not dramatically reduce the battery-life of earables is yet another pressing challenge.

This dissertation deals with three specific sensing modalities, key enablers for personal-scale applications: IMUs, microphones and PPG. The remainder of this section will break

down the broad challenges facing earables discussed above, and relate them to the sensors of interest in this thesis.

Inertial Measurement Unit

Research efforts are ongoing to determine the extent to which kinetic earables [2, 8, 9], i.e., earables equipped with inertial measurement units (IMUs), can be used to track human activities. Early works have seen earables featured as a fitness tracker (e.g., step counting [10]), as well as to monitor dietary habits (e.g., food and beverage intake and classification [7]). Additionally, by looking at inertial data recorded from the ears, earables have shown potential to detect medical conditions like teeth grinding or jaw clenching (Bruxism) [11, 12]. Yet, prior to the writing of this dissertation, no one had explored the potential of earables in tracking head movements [13]. By only being equipped with accelerometers and gyroscopes, existing kinetic earables lack the presence of a magnetometer, key to ensure the accuracy of the IMUs [14]. Besides, without an absolute reference, like that provided by the Magnetic North, it is very hard to correctly initialize any inertial motion tracking system [14, 15]. The reason for absence of magnetometers in earables, and the consequent little investigation around head motion tracking, mainly lies in the inherit device form constraints of today's earable platforms, and the consequent difficulty of ensuring high-quality data. For instance, to embed a magnetometer in an earable, it is key to avoid magnetic interference. Normally, this could be achieved by placing the sensor as far as possible from, for example, sources of interference such as speakers and Bluetooth circuitry. However, as we have discussed, the small size of earables is such that this is not a viable option (Chapter 3).

In-ear microphone

Although limited, we have seen previous works where 6 degrees of freedom (DoF) IMUs have been used in earables to sense motion [2, 10]. Yet, interestingly, the literature lacks research efforts related to exploring similar sensing opportunities with alternative modalities that might be more commonly available in earables. For instance, earables are inherently equipped with microphones (both external facing as well as in-ear facing) both for placing calls and for adaptive noise canceling purposes. While embedding an in-ear microphone in an earable is not as challenging as augmenting an earable with a magnetometer, ensuring high-quality in-ear audio recordings presents a number of hurdles from a data process-

ing point of view. Sound attenuation and background noise inevitably result in very low signal-to-noise ratio (SNR), which ultimately makes sensing personal-scale motion with external facing microphones extremely challenging. Besides, it is often very hard to discern and distinguish different types of motion from noise (as noise artifacts may appear superimposed to the actual signal). In addition, while in-ear microphones could become useful for motion-sensing personal-scale applications (Chapter 4 and Chapter 5), today's earable platforms lack application programming interfaces (APIs) to access those data. Finally, from a systems perspective, it becomes key to process the audio recordings efficiently and often in a quasi-real-time fashion (e.g., to enable continuous step counting, activity recognition, etc.), while trying to minimize the power consumption and preserve all the information present in the data.

In-ear photoplethysmography:

If equipped with photoplethysmography sensors (PPG), earables can be used to continuously monitor vital signs like resting heart rate and respiratory rate, as well as other common fitness-related bio metrics (e.g., heart rate variability, blood oxygen saturation, energy expenditure) [16–18]. Yet, as we will discuss in Chapter 6, the literature fails to deliver systemic studies of motion artifacts experienced at different vantage points around the ear. In particular, to facilitate earable-based photoplethysmography, there are a number of sensing challenges to overcome. First and foremost, it is not clear where is the best location for sensing PPG in/around the ear. This has implications not only related to the quality of the collected PPG data itself, but also on the form of the device and the overall manufacturing of the earable. Additionally, the extent to which motion (both full body as well as more specifically related to the head/face) affects the quality of the earable-sensed PPG signal is still largely uncharted territory. The presence of motion induces artifacts heavily impacting the PPG trace by altering the morphology of the PPG pulse (Chapter 6), as well as jeopardizing the quality of the vital signs estimated from the PPG data.

1.2 Thesis and substantiation

We have seen how earables research, and specifically that for person-scale sensing, is still an immature field under development and, as such, it is characterized by a myriad of limitations and challenges yet to be tackled. To address the challenges and limitations in existing earables research, our thesis is as follows: *to enable earables for reliable personal-*

scale sensing applications, we need to investigate the real capabilities and limitations of a set of sensors that could be capable of accurately monitoring different dimensions of human activities, motion, and vital signs. To corroborate our thesis, we evaluate the performance of three sensors that could realistically be the pillars of the next generation of earables. In particular, we look at both their pitfalls as well as at the possible applications they would enable. Ultimately, this dissertation answers the following research questions:

- **Research Question 1:** How can we track head movements with earables and how can we improve the accuracy of the inertial tracking by adding a *calibrated* magnetometer?
- **Research Question 2:** How can we leverage in-ear microphones for general motion sensing and can we exploit acoustic gait to identify users?
- **Research Question 3:** How can we enable in-ear photoplethysmography: where is the best location around the ear to sense PPG and to what extent is in-ear PPG robust to motion artifacts?

To address these questions we first assess the potential of in-ear accelerometer and gyroscopes for tracking head movements. Based on our findings, we explore how we can embed a magnetometer in earables to improve the performances of continuous inertial motion tracking. We then investigate the capabilities of in-ear microphones for an efficient general motion sensing framework. On top of that, we devise an end-to-end system for user identification based on acoustic gait. Subsequently, we study the quality of PPG when collected in three different locations around the ear. Finally, after having identified the best placement for in-ear PPG, we research the impact of a set of motion artifacts that could particularly affect earables (i.e., head and facial movements).

1.3 Contribution and chapter outline

Acknowledging the limitations and challenges of earable sensing, in this dissertation, we characterize and explore the potential of a set of three different sensors when embedded in earables. We discuss the applications these new *in-ear* sensing modalities could enable. Ultimately, this Ph.D. thesis investigates in-ear, 9 degrees of freedom (DoF), inertial measurement units (IMU), microphones facing in-the-ear-canal, and in-ear photoplethysmography sensors (PPG). We summarize the state-of-the-art earable sensing in Chapter 2, while the rest of the dissertation is articulated in three main technical chapters addressing

the research questions posed in the previous section. The three major contributions of this work are outlined as follows:

Contribution 1: In-ear inertial sensing

In Chapter 3, we investigate earables for inertial sensing. Specifically, we look at head tracking. Head tracking is a fundamental component in visual attention detection, which, in turn, could be beneficial to improve the state of the art of hearing aid devices. Unfortunately, earables have inherent size, shape, and weight constraints limiting the type and position of the sensors on such platforms. For instance, lacking a magnetometer in all earables reference platforms, earables lack reference points. Thus, it becomes harder to work with absolute orientations — key to track head rotations. We evaluate the performance of eSense [2], a representative earable device, to track head rotations. By leveraging two different streams (one per earbud) of inertial data (from the accelerometer and the gyroscope), we achieve an accuracy up to a few degrees. We further investigate the interference generated by a magnetometer in an earable to understand the barriers to its use in these types of devices.

Embedding magnetometers in earables is challenging, as earables heavily rely on radio (mostly Bluetooth) communication (RF) and contain magnets for magnetic-driven speakers and docking. We explore the feasibility of adding a built-in magnetometer in an earbud, presenting the first comprehensive study of the magnetic interference impacting the magnetometer when placed in an earable: both that caused by the speaker and by RF (music streaming and voice calls) are considered. We find that appropriate calibration of the magnetometer removes the offsets induced by the magnets, the speaker, and the variable interference due to Bluetooth communications. Further, we present an automatic, user-transparent adaptive calibration that obviates the need for alternative, expensive, and error-prone manual or robotic calibration procedures. Our evaluation shows how our calibration approach performs under different conditions, achieving convincing results with errors below 3° for the majority of the experiments.

Contribution 2: In-ear microphone sensing

In Chapter 4 and Chapter 5, we explore the potential of in-ear facing microphones for personal-scale human motion sensing. Interestingly, due to the interference from head movements and/or background noise, commonly-used modalities (e.g., IMUs and tradi-

tional, outward-facing microphones) fail to reliably detect both intense and light motions. To obviate this, in Chapter 4 we propose OESense, an acoustic-based in-ear system for general human motion sensing. The core idea behind OESense is the joint use of the *occlusion effect* (i.e., the enhancement of low-frequency components of bone-conducted sounds in an occluded ear canal) and an inward-facing microphone, which naturally boosts the sensing signal and suppresses external interference. We prototype OESense as an earbud and evaluate its performance on three representative applications: step counting, activity recognition, and hand-to-face gesture interaction. With data collected from 31 subjects, we show that OESense achieves 99.3% step counting recall, 98.3% recognition recall for 5 activities, and 97.0% recall for five tapping gestures on human face, respectively. We also demonstrate that OESense is compatible with earbuds' fundamental functionalities (e.g. music playback and phone calls). In terms of energy, OESense consumes 746 mW during data recording and recognition and it has a response latency of 40.85 ms for gesture recognition. Our analysis indicates such overhead is acceptable and OESense is potential to be integrated into future earbuds.

Building on top of the remarkable performances showed by OESense in counting steps, in Chapter 5, we decided to investigate the potential of earables in tracking gait (i.e., the walking style of a person). Human gait is a widely used biometric trait for user authentication and identification. Given the wide-spreading, steady diffusion of earables as the new frontier of wearable devices, we investigate the feasibility of earable-based gait identification. Specifically, we look at gait-based identification from sounds as induced by walking and propagated through the musculoskeletal system in the body. Once again, our system, EarGate, leverages an in-ear facing microphone which exploits the earable's *occlusion effect* to reliably detect gait from inside the ear canal, without impairing the general usage of earphones. With data collected from 31 subjects, we show that EarGate achieves up to 97.26% Balanced Accuracy (BAC) with very low False Acceptance Rate (FAR) and False Rejection Rate (FRR) of 3.23% and 2.25%, respectively. Further, our measurement of power consumption and latency investigates how this gait identification model could live both as a stand-alone or cloud-coupled earable system.

Contribution 3: In-ear photoplethysmography sensing

In Chapter 6, we study in-ear photoplethysmography (PPG) for uncovering exciting opportunities for sensing and cardiac healthcare research with earables. Photoplethysmography

is a simple, yet very powerful, technique to study blood volume changes by looking at light intensity variations. PPG sensors are mechanically straightforward and have become an important enabler to detect various vital signs (including heart rate) across a range of wearables like smart-watches and, more recently, earables. However, it is critical to understand and characterize sensory measurements' accuracy in earables impacting healthcare decisions.

Therefore, we report a systematic characterization of in-ear PPG in measuring vital signs: heart rate (HR), heart rate variability (HRV), blood oxygen saturation (SpO_2), and respiration rate (RR). First, we explore in-ear PPG inaccuracies stemming from different sensor placements and motion-induced artifacts. We observe statistically significant differences across sensor placements and between artifact types, with in-the-canal placement showing the lowest inter-subject variability. However, our study shows the absolute error climbs up to 29.84%, 24.09%, 3.28%, and 30.80% respectively for HR, HRV, SpO_2, and RR, during motion activities (like walking and running). Our results suggest that in-the-canal in-ear PPG is reasonably accurate in detecting vital signs but demands careful mechanical design and signal processing treatment.

Besides common activities like walking and running, PPG signals are also severely affected by motion artifacts caused, for example, by skin and tissue deformations, blood flow dynamics or movement of the sensor over the skin. Generally, motion artifacts typically manifest as significant changes to the signal's morphology (e.g., missing heartbeats), hindering the trustworthiness of vital signs extraction. This problem is particularly important on earables. Despite being on the head, the most stable part of the human body, head movements and facial expressions that people perform naturally when wearing earables, cause skin and tissues displacements around the ear and inside the ear canal. Understanding such artifacts is fundamental to the success of earables as the next wearable platform for accurate and reliable cardiovascular health monitoring. However, the lack of commercially available open-source PPG-equipped earables, together with the lack of existing in-ear PPG datasets, are preventing the research community from tackling such pressing issue.

To this end, in this dissertation, we first report on the design of a novel ear tip which includes a 3-channel PPG coupled with a 6-axis motion sensor (IMU) co-located on the same tip. This enables the sensing of synchronized and spatially distant (i.e., one tip in the left and one in the right ear) PPG data at multiple wavelengths and the corresponding

motion signature, for a total of 18 data streams. Leveraging our device we collected data from 30 participants while performing 16 natural motions including both head/face motions (ranging from nods/shakes to eyes and mouth movements) and full body movements (i.e., walking, running). Preliminary analysis on the dataset show interesting differences across the various motion artifacts, hinting that is possible to investigate automatic recognition tasks. This unique dataset will greatly support research towards making in-ear vital signs sensing more accurate and robust, hence unlocking the full potential of the next-generation PPG-equipped earables. The research presented in Chapter 6 has been done in collaboration with researchers at Nokia Bell Labs, Cambridge (UK).

1.4 List of publications

Some of the work related to this dissertation has been published in peer-reviewed journals, conferences and workshops as listed below.

Works related to this dissertation

[13] **Head Motion Tracking Through in-Ear Wearables**
Andrea Ferlini, Alessandro Montanari, Cecilia Mascolo, and Robert Harle.
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 (EarComp 2019).

[5] **Enabling In-Ear Magnetic Sensing: Automatic and User Transparent Magnetometer Calibration**
Andrea Ferlini, Alessandro Montanari, Andreas Grammenos, Robert Harle, and Cecilia Mascolo.
Proceedings of the 19th International Conference on Pervasive Computing and Communications (PerCom 2021).

[19] **OESense: Employing Occlusion Effect for In-ear Human Motion Sensing**
Dong Ma, Andrea Ferlini, and Cecilia Mascolo.
Proceedings of the 19th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys 2021).

[20] **EarGate: gait-based user identification with in-ear microphones.**

Andrea Ferlini, Dong Ma[2], Robert Harle, and Cecilia Mascolo.
Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom 2021).

[4] **In-Ear PPG for Vital Signs**
Andrea Ferlini, Alessandro Montanari, Hongwei Li, Ugo Sassi, Chulhong Min, and Fahim Kawsar.
IEEE Pervasive Computing, Special Issue on Computational Materials 2021.

Papers submitted

Mobile Health with Head-Worn Devices: Challenges and Opportunities
Andrea Ferlini[3], Dong Ma[3], Lorena Qendro[3], and Cecilia Mascolo.
Invited Paper at IEEE Pervasive Computing.

EarSet: A Multi-Modal Dataset for Studying the Impact of Head and Facial Movements on In-Ear PPG Signals
Andrea Ferlini, Alessandro Montanari[4], Ananta Balaji, Cecilia Mascolo, and Fahim Kawsar.
Dataset paper at ACM IMWUT (Interactive, Mobile, Wearable and Ubiquitous Technologies).

Other works

[21] **Revisiting WiFi offloading in the wild for V2I applications.**
Furong Yang, Andrea Ferlini, Davide Aguiari, Davide Pesavento, Rita Tse, Suman Banerjee, Gaogang Xie, and Giovanni Pau.
Computer Networks 202 (2022): 108634.

[22] **PilotEar: Enabling In-ear Inertial Navigation.**
Ashwin Ahuja, Andrea Ferlini, and Cecilia Mascolo.
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 (EarComp 2021).

[2]Dong Ma, Postdoctoral student at the University of Cambridge, contributed equally to the research work reported in this paper.
[3]First Authors with equal contribution.
[4]Alessandro Montanari, Research Scientist at Nokia Bell Labs Cambridge, contributed equally to the research work reported in this paper.

[23] **Motion-resilient Heart Rate Monitoring with In-ear Microphones.**
Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo.
arXiv preprint arXiv:2108.09393 (2021).

[24] **Corner-3D: A RF simulator for UAV mobility in smart cities.**
Andrea Ferlini, Wei Wang, and Giovanni Pau.
The ACM SIGCOMM 2019 Workshop on Mobile AirGround Edge Computing, Systems, Networks, and Applications (MageSys 2019).

[25] **Leader-Follower Formations on Real Terrestrial Robots.**
Alexandru Solot, and Andrea Ferlini.
The ACM SIGCOMM 2019 Workshop on Mobile AirGround Edge Computing, Systems, Networks, and Applications (MageSys 2019).

[26] **C-Continuum: Edge-to-Cloud computing for distributed AI.**
Davide Aguiari, Andrea Ferlini, Jiannong Cao, Song Guo, and Giovanni Pau.
IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS).

'A full belly is little worth where the mind is starved.'

Mark Twain

Chapter 2

Related Work

In the previous chapter we discussed the potential of earables for future personal-scale sensing applications. After discussing the limitations and challenges of today's earable platforms, in this chapter we proceed by reviewing techniques and applications of earables for personal-scale sensing. Specifically, we discuss prior work on kinetic earables, i.e., earables augmented with IMU (Section 2.1) and sensing attempts leveraging earables' microphones (Section 2.2). We briefly discuss how inertial motion sensing differs from audio-based motion sensing. Finally, we conclude by describing the efforts that leverage in-ear/around-the-ear PPG for vital signs monitoring (Section 2.3).

2.1 Inertial sensing

Earables are a relatively new concept. One of the key research opportunities facilitated by earables, is the possibility of tracking various sources of motion from a new vantage point: the human head [27]. While off-the-self earables often do not feature inertial sensors, or at least do not expose application programming interface (API) and data, the earable research community has greatly benefited from the release of *eSense* [9]. Emerging from the efforts of Kawsar et al., eSense is a kinetic earable research platform which allows researchers to access accelerometer and gyroscope data [9]. The research community has leveraged earable-based inertial sensing for activity recognition [2], as well as step counting [10]. IMU-equipped earable prototypes have also proven effective for health oriented applications like dietary habits monitoring [6, 7, 28], jaw clenching [11] and teeth grinding (Bruxism) detection [12]. Among these, Bedri et al. [6] and Amft et al. [28] combine IMUs and

microphones to detect and classify eating activities. The former aims to detect in-the-wild chewing activities, whereas the latter focused on the classification of four different types of food through the analysis of eating activities.

2.1.1 Head motion tracking

Inertial motion tracking is a known challenge and a well explored area [14, 15, 29]. One of the most recent works in the field, and the state-of-the-art approach when it comes to IMU-based motion tracking, is the research presented by Shen et al. [14]. In their paper, the authors widely discuss the 3D orientation problem (i.e., correctly estimate the object's 3D orientation a global reference frame) and present MUSE, a magnetometer-centric sensor fusion algorithm for orientation tracking. According to their results, MUSE raises the bar, outperforming all the previous state-of-the-art orientation tracking approaches. Prior to their work, the other state-of-the-art techniques, like A^3 [15], heavily relied on the acceleration of gravity (g) to determine the object orientation in the space and leveraged magnetometer readings to re-calibrate the system. However, as reported by Shen et al. [14], those previous works rely on the following three assumptions:

1. slow linear motion, with accelerometer data that have g as average;
2. slow rotational motion, with Gaussian errors that preserve the linearity of the system;
3. motion with frequent, fairly long pauses, needed to reset the gravity estimation.

Yet, because of the unpredictability characterizing human motion, these assumption rarely hold when tracking head movements. While previous work, such as that of LaValle et al. [30], have attempted to track head movements with augmented reality (AR) headsets (e.g., Oculus), earables-based approaches have yet to be explored. To this end, the lack of commercially available earables equipped with a magnetometer represents a significant hurdle. Without the aid of a global reference, like that provided by a magnetometer, it is particularly challenging to recalibrate the inertial sensors, thus leading to poor accuracy when tracking head movements.

2.1.2 Magnetometer calibration

As we have discussed in Section 2.1.1, as of today, earables do not feature a magnetometer. Yet, to facilitate accurate in-ear IMU-based applications, such as head motion tracking, it is key have a properly calibrated magnetometer to rule out drift from accelerometers

and, most importantly, gyroscopes. However, as we will detail in the remainder of this section, calibrating a magnetometer is a non-trivial task, especially when bearing in mind the usability requirements of earables.

Although in-ear magnetometers are mostly affected by a static interfering component induced by permanent magnets in the earables' case, they are also impacted by a dynamic component caused by radio frequency (RF) communications and audio playback. A possible approach to rule out the dynamic interfering component, could be to isolate the magnetometer with some kind of special material, preventing magnetic disturbance. Sadly, this is not a viable option: without even considering its cost, it would also likely isolate the magnetometer from the magnetic field we wanted to measure in the first place [31]. Similarly, filtering approaches [32] do not work for perturbations generated by RF circuitry [31]. Besides, increasing the air gap between sources of interference and magnetometer [13, 31] is also not feasible: assuming these can be modeled as magnetic dipoles, the strength of the magnetic field they generate decreases with r^3, where r is the radius of a sphere with the magnetic dipole as its center [33]. Furthermore, this would be a questionable choice given the inherent design constraints and the miniaturization trend of commercial earables. While it might minimize the interference caused by the earable circuitry itself, this approach is still not enough as it does not account for those interfering components external to the earable's case (i.e., external magnetic disturbances) [34, 35]. Likewise, similar arguments hold for factory calibration [34].

Sensor calibration is a well studied topic with a rich body of literature. However, only a few works specifically tackle calibration for mobile devices [36], and, to the best of our knowledge, none investigate calibration strategies specifically for earables, which are affected by variable interference. The variable nature of interference affecting earables demands frequent calibration updates. For this reason, and the fact that they are often error prone, relying on user inputs (as most of traditional approaches do) is not a viable option. The aim of a magnetometer calibration is to find some parameters, *bias* and *scale factor*, such that it is possible to measure the Earth's magnetic field, and nothing else [37]. Thus, the sensor can be used, for example, as heading source.

A magnetometer calibration can be either static or dynamic [34]; attitude dependent or not [35].

Static calibrations are often performed with the aid of specialized equipment (e.g. a proton magnetometer [38], or a robot arm [39]), or by manual direction placement [40]. Historically, the most common approach is known as *compass swinging* [41]. Used to calibrate marine and aviation compasses, it consisted of leveling and rotating the ship/aircraft through different, known, orientations. This only works for 2D-magnetometers, and requires the user to be instructed to rotate the compass in specific orientations [42].

Dynamic calibrations are generally more practical for mobile devices. Usually, these calibration approaches rely on additional information from the system (e.g. IMU or GPS). Many are iterative and carried out at run-time, often trading accuracy for adaptation. For this reason, they are more practical for mobile devices. Examples are ellipsoid fitting [43], Kalman-filter-based iterative algorithms [34], and stochastic optimization approaches [35], with the most famous being the *figure 8* calibration: the user has to move the magnetometer along an 8-shaped trajectory, to collect enough data to run an ellipsoid fitting algorithm. However, these are cumbersome for the user, and often error-prone [36]. To cope with the performance and the usability requirements of inertial-based applications of earables, it is crucial to devise a completely user-transparent, adaptive, magnetometer calibration, which specifically targets earpieces, without requiring any specialized equipment.

2.2 Acoustic motion sensing

Motion sensing is a key area of interest in wearable and earable research. While IMUs are often the most commonly adopted sensor [2, 10, 11], researchers have also explored the potential of different modalities for this purpose. For instance, Ando et al. [44] observed that the human ear canal is deformed when subtle motion and facial expressions occur. Building on this observation, the authors proposed an in-ear barometer to recognize facial expressions, by capturing the pressure differentials caused by facial muscle movements in the ear canal. Differently, Matthies et al. [45] utilized the electrodes on earbuds to capture the electricity variations caused by muscle movements, also during facial expressions. Yet, to date, there has been very little work around in-ear facing microphones for motion sensing. Notably, in-ear microphones are a key component for active noise cancellation (ANC), and therefore are available in many off-the-shelf earables and in most hearing aids.

Thanks to the proliferation of microphones and speakers in Internet of Things (IoT) devices and wearable ecosystem, acoustic sensing has rapidly became crucial for a broad

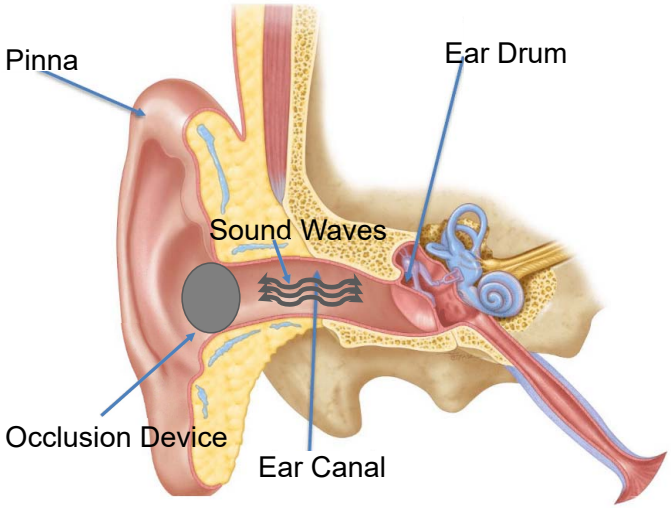


Fig. 2.2.1: Anatomy of the human ear and explanation of the occlusion effect. When the orifice of the ear canal is occluded, sounds are trapped inside the ear canal, resulting in the amplification of their low frequency components.

range of applications [46–49]. Various properties of audio signals, such as time-frequency features [50–52], time difference of arrival (TDoA) [53], Doppler shift [54], phase [55], and channel impulse response [56], have been exploited for motion sensing in the past. Based on whether a speaker is involved in actively generating a reference audio *beacon*, or not, these applications can be categorized into passive sensing and active sensing. In this section, we focus on passive acoustic sensing as it is more relevant to our work.

2.2.1 Facilitating in-ear acoustic sensing: the Occlusion Effect

Before surveying the existing efforts related to passive acoustic motion sensing, this section introduces a physiological phenomenon known by the name of *occlusion effect*. Practically, to facilitate in-ear acoustic sensing, in this dissertation, we leverage the natural low-frequency boost provided by the occlusion effect. As measured in [57], the occlusion effect can boost the sounds below $1000Hz$ by up to $40dB$ depending on the frequency.

Hence, the reminder of this section provides the reader with a basic understanding of what the occlusion effect is in practice, and how it can be exploited to facilitate a number of acoustic sensing applications.

From the physiological point of view, the occlusion effect can be defined as a dominance of the low-frequency components of a bone-propagated sound due to the loss of relevance of

the outer ear sound pathways whenever the ear canal orifice is sealed (i.e., occluded) [58]. For example, such phenomenon could be experienced in the form of echo-like/booming sounds of their own voice if a person is speaking and is obstructing her ear canals with a finger or an earplug. Essentially, sound is nothing but vibrations propagating as acoustic waves. Usually, it travels through bones and escapes the inner-ear via the ear canal orifice. However, if this opening is obstructed, the vibration waves are blocked inside the canal and are bounced back to the eardrum [59], as illustrated in Figure 2.2.1. As a result of that, the low-frequency-bone-conducted sounds are amplified [60]. A more precise definition of what the occlusion effect entails can be denoted by the ratio between the sound pressure inside the occluded ear canal and that in the open ear [61]. Specific to the applications considered in this dissertation, whenever a user touches their face, or while a person walks, the vibrations generated by the tap on the skull, or by a foot hitting the ground, as soon as a person takes a step (e.g., basic component of a gait cycle), propagate through the body via bone-conduction. Interestingly, these vibrations are amplified if the ear canal of the person is occluded by, for example, an earable, ultimately facilitating in-ear acoustic sensing.

2.2.2 Activity recognition

Passive acoustic sensing measures the sounds generated by target activities/motions with microphones. The mobile sensing literature presents several instances where passive acoustic sensing is leveraged for activity recognition. For instance, Wang et al. [50] presented a keystroke recognition system named UbiK using the smartphone microphone. Ubik calculates the amplitude spectrum density (ASD) of the clicking sound and identifies different keystrokes using a fingerprinting-based method. Employing a dual-microphone design on current smartphones, Liu et al. [53] proposed to discriminate keystrokes based on the TDoA of the keystroke sound at the two phone microphones. By recording the sound during sleeping, Ren et al. [51] detected human breathing rate with a correlation-based method and recognized different events (like cough and snore) with the Mel-frequency cepstral coefficients to estimate the quality of sleep. Chauhan et al. [52] proposed to authenticate users using the breathing sound. Yet, to date, no one has leveraged the sounds recorded from inside the ear canal for activity recognition.

2.2.3 Hand-to-face interactions

In addition to classifying medium to intense motions such as walking, running, chewing, and drinking, the natural low frequency boost enabled by the occlusion effect (Section 2.2.1) facilitates sensing light motions like hand-to-face gestures. The human face offers a large area for interactions with head-worn devices such as earables. In particular, hand-to-face gestures have been explored using various modalities. Examples are the work by Serrano et al. [62] where they proposed to detect hand-to-face gestures using a camera array. A similar effort is the work by Kikuchi et al. [63] who devised EarTouch. Kikuchi et al. built an earbud featuring four photo-reflective sensors to measure the deformation of the auricle (or ear rim) caused by touching the ear with the fingers. Similarly, although with a different form of device (head-mounted display), Yamashita et al. [64] also use photo-reflective sensors to measure the deformation of the skin when the user touches their cheeks. Moreover, Lee et al. [65] used an Electrooculography (EOG) sensor equipped on the nose pad of commercial eyeglasses to detect nose touching gestures.

Closer to the work presented in this dissertation (Chapter 4), Xu et al. [66] introduced a hand-to-face interaction system called EarBuddy using acoustic signals. However, EarBuddy recognizes face-touching gestures using audible air-conducted sounds picked up by an external-facing microphone on the main earbud body. Ultimately, this suffers from poor signal-to-noise ratio (SNR) due to the dramatic attenuation experienced by sound-waves propagating through the air. As a result, EarBuddy shows good performance only in recognizing gestures close to the ear. Further, it suffers from acoustic noise and can only work in quiet environments. In contrast, no one has yet taken advantage of the occlusion effect nor leveraged the characteristics of in-ear microphones to record bone-conducted sounds.

2.2.4 Gait tracking

The improved sensory capabilities of earables facilitated by the occlusion effect (Section 2.2.1), suggest the potential of ear-worn devices in counting steps and tracking *gait*. The walking style of a person is commonly known as their gait. Medical and physiological studies [67] suggest the human gait has 24 different components. The differences between the gait of distinct subjects are caused by the uniqueness in their muscular-skeletal structure. The human gait is regulated by precise bio-physical rules [68]. These, in turn, are dictated by the tension generated by the muscle activation and the consequent movement of the joints. As a result, the forces and moments linked to the movement of the joints

cause the movement of the skeletal links which, therefore, exert forces on the environment (e.g., the foot striking the ground). Hence, the human gait can be described as a generation of ground-reaction forces which are strongly correlated with the muscular-skeletal structure of each individual. In practice, differences in the body structure of individuals are among the factors that produce the interpersonal differences in walking patterns that enable gait-based identification.

Gait is a well-studied human biometric, proven to be unique for each individual, and, therefore, often used as an identification biometric [67,68]. Traditional approaches for gait recognition enumerate machine vision-based strategies [69], floor sensor-based techniques [70], wireless fingerprinting based methods [71], and wearable sensor-based systems [72,73]. All of these techniques have their own specific advantages (e.g., user-transparent and complete device-free) and disadvantages (e.g., high computation cost and privacy-related issues for vision-based method; need for deploying the wireless transceivers for wireless fingerprinting based method, etc.), thereby complementing each other under different scenarios.

Contrary to the more traditional techniques, there are two prior works using acoustic as the modality for gait recognition. Geiger et al. [74] exploited a microphone attached to the user's feet to record the walking sounds when the feet hit the ground. The main drawback of their approach is that the step sounds might change with different ground materials or footwear soles. In fact, in the extreme case, the microphone may not be able to pick up any noticeable sound when walking on a carpet barefoot. Instead, as we will detail in Chapter 5, the work presented in this dissertation leverages the occlusion effect to measure bone-conducted sounds (essentially vibrations) in the ear canal, guaranteeing robustness to both footwear and ground material. Wang et al. [75] proposed a fingerprinting-based system (named Acoustic-ID) for human gait detection using acoustic signals. Specifically, by deploying a pair of acoustic transmitter (ultrasound) and receiver, the gait pattern is extracted by measuring the reflected acoustic variations when the user is walking within the sensing range. Unlike Acoustic-ID, that requires to actively transmit ultrasounds, our approach is completely passive, and does not require the deployment of any additional hardware.

Instead of using gait, researchers have discovered other biometric traits that can be acquired from the human ear for user identification. Nakamura et al. [3] proposed to identify and authenticate users by leveraging in-ear electroencephalogram (EEG) measured using a customized earpiece. More recently, Gao et al. devised EarEcho [76]. Their system

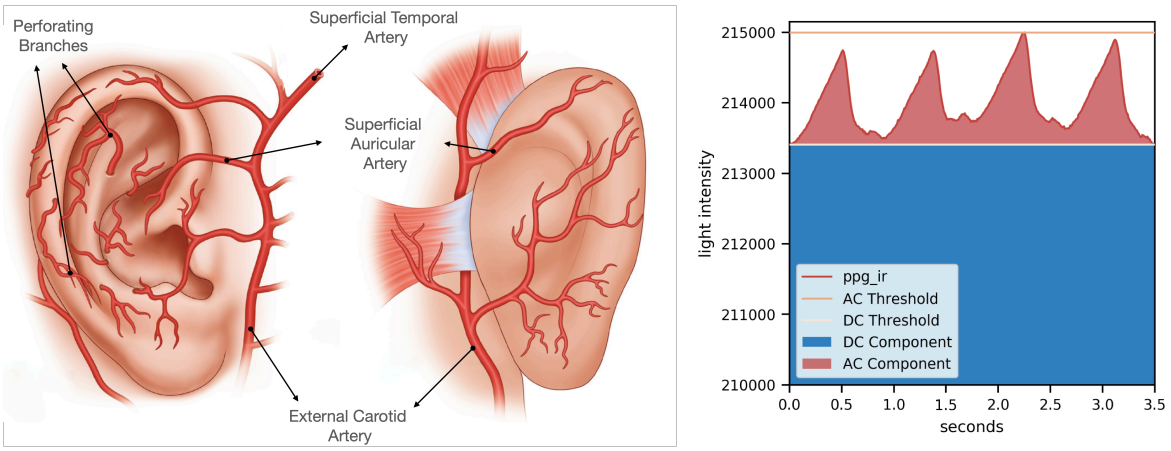
leverages the uniqueness of ear canal geometry to recognize and identify the users. However, as demonstrated in [77], the geometry of the ear canal changes during different facial expressions, which might impact the effectiveness of EarEcho in the daily life. Although a similar counter argument could be raised for gait, which also might change over time, the latter usually evolves over a larger time span. Moreover, the uniqueness of the ear canal geometry has never been tested over a large-scale population (featuring only 20 subjects involved in [76]). As of today, a reliable system to identify the wearer of a pair of earables, through passive measurements, is still missing.

2.3 Photoplethysmography sensing

Earables have the premises to revolutionize personal health and well-being. In this section we provide the reader with some basic notions on the anatomy of the human (Section 2.3.1) before introducing the underlying principle of photoplethysmography (Section 2.3.2) and some of its wearables applications (Section 2.3.3).

2.3.1 Ear anatomy

The ear is part of the human sensorium. It is the organ that enables hearing and balance. In mammals, the ear is generally composed of three parts: the outer ear (the visible part of the ear), the middle ear (where the sound waves, coming form the outer ear, are modulated), and the inner ear (where the modulated sound waves are finally transmitted to the brain). In this dissertation, we focus on the outer ear which is where earables are typically located. In fact, unlike the middle and the inner ear, placing electronic sensors around the outer ear does not require any invasive or uncomfortable interaction. Further, focusing on the outer ear, we can neglect those conditions (e.g., tympanic infections, eardrum and cochlea impairments) that might affect the middle and inner ear. Given the objective of this section is to brief the reader on in-ear photoplethysmography, it is also key to consider a part of the ear which is well supplied by blood vessels. The outer ear is composed by the pinna (or auricle, it is the visible part of the ear); the concha (it is the depression in the pinna leading to the ear canal orifice); and the ear canal itself. These three areas are all well supplied with blood by branches of the external carotid artery (Figure 2.3.1a), one of the most important arteries in the human body. Specifically, the pinna (on both sides), as well as the concha, are supplied by capillaries known by the name of *perforating branches*; the ear canal, on the other hand, is adjacent to another major blood vessel, the *superficial*



(a) **(b)**

Fig. 2.3.1: Blood vessels around the outer ear (Figure 2.3.1a). Pulsatile (AC) and non-pulsatile (DC) components of a typical PPG signal (infrared wavelength in this example). The DC component indicates the light absorption from the tissues, the bones, and the static blood in veins and capillaries; the AC component reflects the pulsations of the arterial blood caused by the cardiac activity (Figure 2.3.1b).

temporal artery.

2.3.2 Photoplethysmography primer

Photoplethysmography, more commonly known by its acronym PPG, is an optical measurement technique featuring a light emitting diode (LED) and a photodetector (PD). Often used as a more convenient (less expensive and easier to implement) alternative to electrocardiography (ECG), a PPG signal is the measurement of the reflected (back-scattered), or of the transmitted light through the region of tissues under examination. By looking at the intensity of the light at the PD, it is possible to detect variations in blood volume which occur with each heartbeat. While in clinical settings transmissive photoplethysmography (the light transmitted by the LED through the tissues is absorbed by the PD on the other end of, for instance, a finger) is often employed, for practical reasons, we focus on reflective PPG. The sensor setup of reflective PPG is analogous to that of its transmissive counter part, with the only exception being the placement of the PD. Indeed, rather than being on the other end of the extremity (e.g., finger, ear lobe), the photodetector is placed next to the LED. Reflective PPG is the de facto standard for heart rate (HR) monitoring in wearables. The wavelength of the light used in PPG sensors typically ranges between 500nm (green color) and 1100nm (infrared). Red and infrared (IR) lights are absorbed less

by the water present in the human tissues compared to green light. As a result, depending on the wavelength of the LED, the light is absorbed differently by the skin and, therefore, it is possible to achieve different depths in the tissue. It is well studied how, for example, green light penetrates the tissues less than red or infrared (IR) light [78]. This has crucial implications: by penetrating less, green light is less susceptible to motion artifacts, whilst, on the other hand, red and IR light provide a higher resolution of both the DC and AC components.

A typical PPG signal is depicted in Figure 2.3.1b. The pulsatile (AC) component reflects the pulsations from the cardiac cycle [79], while the non-pulsatile (DC) component comprises absorption from the tissue and bones, as well as static blood absorption (arterial, venous and in smaller amounts, capillary) [79]. From this signal a number of clinical information can be derived, for example heart rate and heart rate variability, respiratory rate and blood oxygen saturation. However, whether the photoplethysmograph can provide all of these insights, or not, depends on many factors. First and foremost, the PPG signal has to be clean, with peaks clearly distinguishable and with systolic and diastolic waves well visible. A PPG pulse (as those in Figure 2.3.1b) is composed of the anacrotic phase (the rising component, also known as systolic wave) which signifies the systole of the heart and of the catacrotic phase (the falling component, also referred to as diastolic component) which signifies the diastole and wave reflections in the periphery. However, whenever motion (or ambient light) is concerned, it is not always possible to observe them. As a result, PPG-extracted information often lose their clinical valence in presence of motion [80]. In the next section we will examine the typical applications enabled by a clean PPG signal and the corresponding features along with the impact of motion artifacts on the signal quality.

2.3.3 Wearable applications of PPG sensors

Most consumer devices such as fitness trackers[1] [2] and mobile phones[3] use PPG sensors for heart rate monitoring. In recent years, accurate and real-time sensing of the three main vital signs (HR [81], SpO_2 [82], and BP [83, 84]) has been demonstrated using wrist-worn wearables equipped with PPG sensors. In addition, explorations have been made to detect heart rate variability [85, 86], sleep apnea [87, 88], atrial fibrillation [89, 90], respiration

[1] https://www.apple.com/apple-watch-series-7/
[2] https://www.fitbit.com/
[3] https://www.samsung.com/us/heartratesensor/index.html

Table 2.3.1: Summary of PPG signal features essential for bio-markers as well as other health sensing applications.

Applications	AC Component	DC Component	Time domain signal features	Frequency domain signal features	First order derivative features	Second order derivative features
Vital sign sensing (HR [81], SpO_2 [82], BP [83])	✓	✓	✓	✓	✓	✓
Heart rate variability (HRV) [85, 86]	✓	×	✓	×	×	×
Respiration rate (RR) [91, 92]	✓	×	×	✓	×	×
Sleep apnea [87, 88]	✓	×	✓	×	×	×
Atrial Fibrillation [89, 90]	×	×	✓	✓	✓	✓
Arterial Stiffness [93, 94]	×	×	✓	✓	✓	✓
Energy expenditure [95]	✓	×	✓	✓	✓	✓
Dehydration [96, 97]	×	✓	×	×	×	×

rate [91, 92], arterial stiffness [93, 94], energy expenditure [95] and dehydration [96, 97]. In this section, we briefly summarize the most common signal processing techniques used with PPG signals for measuring the following main bio-markers:

1. **Heart rate:** Peaks are detected from the AC component of the PPG signal to obtain the number of beats per minute. Typically the raw PPG signal is band-pass filtered between [0.4Hz, 4Hz] to obtain the AC component corresponding to the heart rate.

2. **Oxygen saturation (SpO_2):** Oxygenated hemoglobin absorbs less red light whereas deoxygenated hemoglobin absorbs less infrared light. Thus, the ratio between red and infrared light intensities measured by the PPG sensor can be used to estimate SpO_2 (R) as follows:
$$R = \frac{R_{red}}{R_{infrared}} = \frac{AC_{red}/DC_{red}}{AC_{infarared}/DC_{infrared}} \tag{2.1}$$

3. **Heart rate variability (HRV):** Heart rate variability is measured as the time difference between adjacent peaks in a PPG signal.

4. **Respiration rate (RR):** A Synchrosqueezing transform (SST) [98] is applied on the raw PPG signals to extract the respiration component (0.1-0.9Hz). The number of peaks in the resulting respiration component of the PPG signals correspond to the respiration rate (breaths per minute). Besides, there are other techniques [99] using time domain and frequency domain features extracted from the PPG signal along with machine learning to estimate respiration rate.

5. **Blood pressure (BP):** Blood pressure is typically computed by placing PPG sensors at two locations on the same artery (say, finger and wrist) and then measuring the time taken by the pulse wave to travel from one PPG location to the other (pulse transit time). BP is inversely proportional to the pulse transit time obtained by

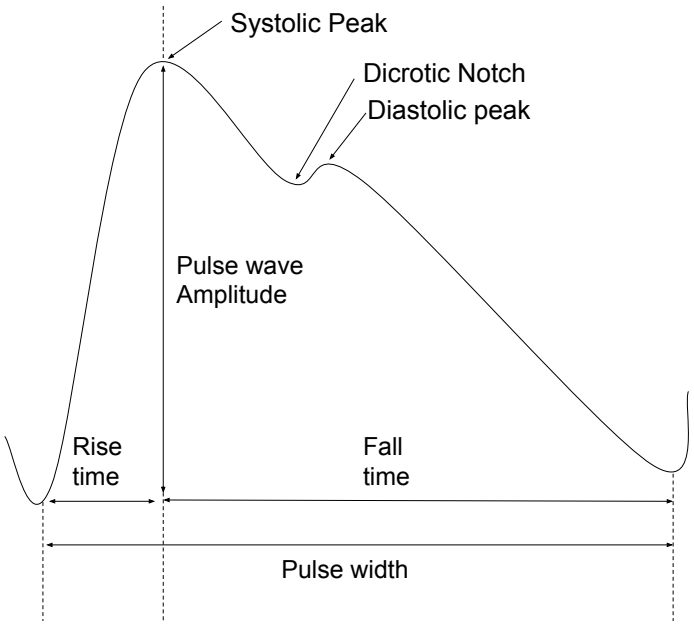


Fig. 2.3.2: Typical time domain signal features extracted from a PPG signal.

calculating the peak time shifts between the two PPG sensors. In recent years, many machine learning and deep learning techniques [83, 100] have also been proposed to estimate blood pressure from the extracted PPG signal features.

As seen from the above bio-markers, the time domain signal features from the PPG signal are essential to estimate heart rate, heart rate variability as well as blood pressure. Some of the frequency domain features help in differentiating a normal sinus rhythm from an arterial fibrillation (AF) signal or an abnormal heart signal. In addition to the above mentioned features, many techniques use features extracted from the first order derivatives and the second order derivatives of the PPG signal to compute arterial stiffness [93] and blood pressure [83]. The second order derivative of a PPG signal provides useful information such as dicortic notch, time at which the diastolic peak occurs which provides information regarding the blood flow dynamics (systolic and diastolic phases).

Table 2.3.1 shows the main feature categories required for several critical health sensing applications. In addition to the PPG signal features mentioned earlier, useful physiological features marked in Figure 2.3.2 can also be derived from the PPG signal. The following list describes in more detail these main features (which are also the ones we used in Chapter 6 to analyze how various head and facial expressions affect in-ear PPG signals):

1. **Systolic phase:** The Amplitude of the systolic peak and the time at which the systolic peak is located in the PPG signal.
2. **Diastolic phase:** The Amplitude of the diastolic peak and the time at which the diastolic peak is located in the PPG signal.
3. **Ratio between systolic and diastolic phase**: An indicator of the abnormalities in blood pressure. It is also referred to as Augmentation index or Reflection index.
4. **Pulse width**: The time between the beginning and end of a PPG pulse wave. It correlates with our heart's systemic vascular resistance.
5. **Rise time**: The time between the foot of the PPG pulse and the systolic peak.
6. **Perfusion index (PI)**: PI is the ratio of the pulsatile blood flow (AC component) to the non-pulsatile or static blood in peripheral tissue (DC component).
7. **Dominant frequency**: The dominant frequency of PPG signal can be useful to give insights concerning the presence of artifacts at a different frequency outside the heart rate frequency band [0.4,4Hz].
8. **Spectral Kurtosis**: Also known as Frequency Domain Kurtosis, describes the distribution of the observed PPG signal frequencies around the mean and is a very useful indicator of the PPG signal quality.
9. **Peak-to-peak magnitude variance**: The variance of the difference between the pulse wave amplitude between two adjacent pulse wave.
10. **Peak-Time interval variance**: The variance of the pulse width between peaks of two adjacent PPG waves.

2.3.4 Motion artifacts on in-ear PPG

For real-time and accurate estimation of the vitals we have mentioned above, the PPG signal needs to be free from any distortion or noise hindering the computation of the signal features. While many works in the literature have proposed filtering techniques for PPG sensors in wrist-worn devices [101, 102], noise and motion artifacts generated by devices worn in or around the ear have been so far ignored. Despite the high concentration of blood vessels in the outer ear naturally favors photoplethysmography, it also opens the door to multiple potential placements for a PPG sensor, which have never been investigated before.

As we will discuss in Chapter 6, collecting PPG data from inside the ear canal offers several advantages over other potential locations around the ear (and in the rest of the body): the natural darkness of the ear canal ensures ambient light shielding, a particularly desirable feature when it comes to optical sensing. Further, by placing a PPG sensor on an ear-tip, it is possible to achieve a stable positioning. In addition, the natural vibration damping offered by human musculoskeletal system, makes the head relatively motion-resilient than wrist [10]. On top of that, if carefully designed, a comfortable ear-tip could be worn for prolonged amount of time, paving the way to continuous upper-body sensing. Finally, even more importantly, in an earable panorama characterized by heterogeneous designs, shapes, and form factors, ear-tips do not bound ear-worn devices to set form factors unlike, for instance, if the sensor was placed behind the pinna. Nonetheless, in-ear PPG sensing is also susceptible to a new set of artifacts, never studied before, that could dramatically hinder the clinical valence of the vitals and bio-markers extracted from such devices.

Motion artifacts (MA) in PPG can be caused by both the mechanical sliding and rubbing of the PPG sensor over the skin, as well as by the *actual* acceleration of the human body (which changes the blood flow dynamics), and the compression/decompression of tissues and muscles. Unlike wrist-worn devices which are mostly affected by body and hand movements, in-ear PPG sensors will be affected the most by face and head motions. Facial expressions serve as the primary tool to convey emotions between individuals. Common facial expressions like smiles, disgust, or frowns, are used to express emotions such as joy, sadness, anger or disappointment. The facial muscles play a key role in making each and every facial expressions. The facial muscles are located around the mouth, eye, nose and ear. Out of these locations, the muscles in the auricular group (around the ear) are affected when facial expressions are being made – thereby causing distortions in the PPG signals acquired by in-ear PPG sensors. Unlike dominant full-body motions affecting PPG sensors in wrist-worn devices, the facial motions being made are very subtle and may possess different characteristics compared to full-body motions like walking, running, hand movements etc. For instance, a slight smile (lip puller) causes very noticeable distortions in the PPG signal recorded by an in-ear PPG sensor (as we will show in Chapter 6). Although the impact of full-body motion on PPG sensors has been widely studied in prior works [101, 102], the extent to which motion artifacts caused by facial and head movements affecting in-ear PPG sensors is largely unexplored at best.

Facial expressions, and facial movements in general, are taxonomized based on the Facial Action Coding System (FACS) [103]. First published in 1978 and updated in 2002, the

FACS encodes a number of individual facial muscles movements which leads to changes of expressions. These fundamental actions of individual muscles or groups of muscles are called Action Units (AUs). In designing a representative study, it is paramount to focus on a subset of action units, selected from the existing FACS, most of which are known to be picked up by in-ear motion sensors [104, 105].

2.4 Summary

This chapter has reviewed the state of the art in the area of personal-scale sensing with earables. Specifically, we outlined the limitations of current motion tracking earable systems (Section 2.1) highlighting the need for a magnetometer to enable more advanced applications. This dissertation fills these gaps by first assessing the performance of eSense, an exemplar of kinetic earables, in tracking head movements (without the aid of a global reference), and then by presenting a completely user-transparent, adaptive, magnetometer calibration, which specifically targets earables (Chapter 3).

Surveying acoustic motion sensing (Section 2.2), we found that the potential of in-ear microphone applications is greatly unexplored. To this end, in this dissertation we leverage the occlusion effect (Section 2.2.1) to boost the SNR of sounds recorded from inside the ear canal. Practically, this allows us to classify a number of different activities (e.g., walking, running, chewing, drinking, and a stationary baseline), hand-to-face interactions, count steps, and identify people from their gait (Chapter 4 and Chapter 5).

Finally, examining today's in-ear PPG sensing systems (Section 2.3) we identified the need for better understanding of the impact of motion artifacts (MA) on the PPG trace. By identifying the more robust location to sense PPG in/around the ear, and by thoroughly characterizing a number of head and facial MA, this dissertation represents a step forward towards facilitating in-ear PPG sensing (Chapter 6).

'If you can keep your head when all about you
Are losing theirs and blaming it on you,
If you can trust yourself when all men doubt you,
But make allowance for their doubting too;
...
Yours is the Earth and everything that's in it,
And—which is more—you'll be a Man, my son!'

Richard Kipling

Chapter 3

In-ear Inertial Sensing

3.1 Introduction

In Chapter 1 we have discussed the yet-to-be-unlocked potential of earables for personal-scale sensing and the limitations of today's earable research. In this chapter we explore the potential of kinetic earables (i.e., earables featuring inertial sensors such as accelerometers and gyroscopes) in tracking head movements [13]. Facilitating a better hearing experience for future hearing-aids, delivering augmented reality (AR) functionalities to earables, enabling spatial audio, as well as improving navigation, all these applications are all predicated to earables being able to track head movements accurately. In fact, acting as a proxy for visual attention, tracking head movements is a key enabler for a number of personal-scale applications. Moreover, in the second half of this chapter we go one step further and, for the first time, we investigate how to augment earables with a magnetometer to improve the motion tracking performance [5].

As we have discussed in Chapter 1, recent years have seen the rise of wearable technologies, both in the form of specialist devices such as pacemakers and in consumer devices, primarily smartwatches. A growing trend is the use of wireless earbuds that, while designed primarily for personal audio playback, offer a new sensing platform at an important site on the body. For instance, earables could be used both as sensors, collecting useful data like head movements, and actuators, enhancing the listening experience or providing feedback to the user. Ears represent an extremely good vantage point to track both gaze and head movements. Interestingly, sensing these behaviors can yield substantial improvements in

personal-scale applications such as navigation [22], driving safety [106], augmented experience [107], as well as assistance to elderly people [108]. Besides, previous studies from the medical community, have also highlighted the importance of leveraging the visual attention of the user [109, 110] for improving hearing-aids. For example, Favre-Félix et al. [110] use electrooculography data (EOG) to characterize the visual attention of the patients. However, because of the challenging signal processing required when dealing with such complex signals, the authors suggest that EOG might not be the best enabler for personal-scale applications leveraging visual attention. Notably, head movements are closely linked to eye-movements [110], and therefore they are considered a good proxy to sense *visual attention*, too.

Inertial motion tracking is a well known and studied problem. Yet, due to the lack of a reference point to re-calibrate the sensors, and to estimate the 3D orientation of the tracked object, tracking head movements with a device without a magnetometer represents a challenging task. To the best of our knowledge, because of the interference generated by the magnets of the speakers and in their cases, none of the earables in the market is equipped with a magnetometer. Indeed, like the eSense we are using in this work, the Apple AirPods[1], the Google Pixel Buds[2], and the Samsung Galaxy Buds[3] do not have a built-in magnetometer. *In this chapter, we focus on the evaluation of the eSense platform [2, 8] in tracking the head movements of a user concentrating on a specific spatial point.* To do so, we ran experiments with ten volunteers. We probed and stressed the robustness of the system by asking our volunteers to perform different activities, such as chewing and talking, while focusing on a series of targets placed at different spatial locations.

By tracking instantaneous head movements as a proxy to track visual attention, our study shows how a system, that relies only on accelerometer and gyroscope, can still provide useful insights on where the visual attention of a person is. Despite the fact that head movements are user dependent, we obtained estimations with an average error that ranges from 5.4° for short movements done by silent subjects, to 18.7° for longer movements carried out by subjects who are chewing. To better contextualize that, let us consider of a 4 lanes intersection (≈10 meters wide). For instance, considering a 4 lanes intersection , an error of 3° on the heading would entail a displacement offset of about $10 \times sin(5.4/2) \approx 0.47m$ when looking at short movements without any particular noise source. Yet, when

[1] https://www.ifixit.com/Teardown/AirPods+2+Teardown/121471
[2] https://medium.com/@justlv/google-pixel-buds-teardown-396183cbbc18
[3] https://root-nation.com/audio-en/headphones-en/en-samsung-galaxy-buds-review/

accounting for sustained, longer movements, under more challenging circumstances (e.g., user chewing), the errors ramp up to $\approx 1.62m$. This dissertation lays the foundations of a line of work aiming to sense and characterize human attention through earables, wearables that are neither socially-awkward, nor cumbersome, unusable or unrealistic (e.g., combining an hearing-aid with a pair of eye-tracking glasses). Lastly, it sheds light on how a magnetometer would behave if placed in an earable.

Building upon these preliminary results, in the second half of this chapter we focus on the feasibility and value of adding magnetometers to earables. As we have discussed, today's consumer earables already contain inertial sensors (IMU) like accelerometers and gyroscopes. In other mobile devices, these are paired with magnetometers to provide movement descriptors in a global, absolute frame of reference [14], which would be highly valuable at the head too. Although a reasonable argument could be using the magnetometer in the phone to provide the earbuds' IMU with the references needed to calibrate/re-calibrate them, in practice, this would not work. **Notably, the head does not always move according to the way the body does**; besides, they often do not face the same direction. Hence, relying only on the phone would inevitably provide descriptors that do not necessarily match those when moving the head. Beyond this, adding a magnetometer to an earable would allow for a variety of applications including inertial navigation; magnetic-field-based indoor localization [111]; driver monitoring systems [112]; and medical applications (i.e. intra-body localization [113, 114], speech-language therapy [115], trans-cranial stimulation [116]), to name a few. However, as we discussed in Chapter 2, the highly constrained earable form factor and practicalities around their calibration have so far prohibited the availability of magnetometers on earables. Common magnetometer calibration techniques are cumbersome and error-prone [36], and regular manual calibration of *both* earbuds (and any personal devices) is unrealistic. Instead, we leverage the heading of the user's phone – typically trustworthy, as we will discuss in Section 3.8 – to auto-calibrate the magnetometers in the earbuds when the devices are believed to be pointing in the same direction. We ensure the latter constraint by applying the algorithm only when the user is directly interacting with the phone.

Specifically, we first explore how magnetometer signals are affected by magnetic disturbances expected in an earable, highlighting the need for good calibration. Based on these findings, we propose a novel magnetometer calibration technique that leverages the user's phone sensors (typically well calibrated). Calibration routine that we devise can run in the background, without user intervention, providing a semi-continuous calibration. We report

a thorough evaluation of the performance of the calibration framework, both in terms of accuracy and system performance during a proof of concept study with a navigation application.. Besides, we theoretically and practically assess the computational and energy efficiency of our algorithm, showing our approach is **accurate** yet **computationally inexpensive**, allowing its regular execution on a constrained wearable. Our analysis shows that our approach is extremely lightweight consuming $\approx 2.9\%$ extra power over idle.

The techniques and the results presented in this chapter have been published in [13] and in [5].

3.2 Platform overview

In this section we introduce the eSense platform and the challenges of performing head motion tracking on a device that can not rely on the data from a magnetometer.

3.2.1 eSense Platform

The eSense platform consists in a pair of wireless earbuds which have been augmented with kinetic, audio and proximity sensing options [2, 8]. The left earbud has a 6-axis Inertial Measurement Unit (IMU) with accelerometer and gyroscope and a Bluetooth Low Energy (BLE) interface which is used to stream data and to send periodic beacons that can be used to detect proximity to nearby devices. Both earbuds are also equipped with microphones to record external sounds. The benefit of eSense, contrary to other commercial earbuds, is the availability of application programming interfaces (APIs) to access the raw data collected by the onboard sensors, together with the complete flexibility in the configuration of the sensors' parameters. In addition to serve as a well established and socially acceptable device, for example to listen to music and take phone calls, eSense allows to gather real-time sensor data, opening the door to novel sensing applications involving the head, a part of the body which has been mostly unexplored, so far.

3.2.2 Challenges of inertial tracking

The primary goal of this section is to understand the accuracy achievable by a device that solely relies on accelerometer and gyroscope to track user's head movements. The biggest challenges we had to face while investigating that, were related to the device itself. The

small form factor, together with the true "*wireless experience*" and the relatively short battery life altogether represent non-trivial constraints to deal with.

The presence of multiple magnets in the case compelled the hardware manufacturer to put a 6 degree of freedom (6 DoF) IMU, instead of a 9 DoF, more complete, sensor. Practically, as a result of this design choice, the platform ends up being bounded to the 3 DoF of the accelerometer (x_{acc}, y_{acc}, z_{acc}) and the 3 of the gyroscope (x_{gyro}, y_{gyro}, z_{gyro}), lacking the presence of a magnetometer.

Accelerometer and gyroscope provide relative movement estimates that are well known to drift over time [14]. Without the magnetic north as a reference, we could not rely on the state-of-the-art calibration and re-calibration techniques [14]. Besides, when tracking the motion of an object (or of the head of a person), it is crucial to initialize the system correctly, with the right 3D orientation of the object itself. Unfortunately, once again, the majority of the algorithms to solve the so called *3D orientation problem* rely on the absolute direction reference provided by the magnetometer (combined with gravity and instantaneous gyroscope readings) [14, 15]. For instance, in their work, Shen et al. [14] widely discuss the 3D orientation problem and present MUSE, a magnetometer-centric sensor fusion algorithm for orientation tracking. MUSE outperforms all the previous state-of-the-art orientation tracking approaches. Prior to their effort, A^3 [15] was heavily relying on the gravity to determine the object orientation in the space, using the magnetometer data mainly to re-calibrate the system. However, as reported by Shen et al. [14], most previous works rely on the following assumptions:

1. slow linear motion, with accelerometer data that have gravity as average;
2. slow rotational motion, with Gaussian errors that preserve the linearity of the system;
3. motion with frequent pauses, needed to reset the gravity estimation.

Yet, because of the unpredictability of human nature and motion, none of these assumption holds when tracking human movements.

3.3 Head tracking methodology

We leverage the eSense earbuds to collect accelerations (from the accelerometer) and rotational velocities (from the gyroscope). To get an idea of where a user is paying their visual attention to, we integrate accelerations and velocities to estimate positional displacements.

Combining multiple IMUs: Tracking head movements from only one ear might not give precise enough results, especially considering the absence of external reference points, such as a the magnetic north provided by a magnetometer, to aid the tracking and recalibrate the inertial sensors. As having more fine grained data points enhances the precision of the estimation, we combined the streams of inertial data coming from the IMU sensors in the two earbuds. Notably, since only the left eSense buds are equipped with the IMU sensors [9], we had to use two left buds. In doing so, we leverage the duality of earables relying on the assumption that the two different IMUs are recording, from different vantage points, the same rotations. Yet, fusing the data correctly becomes crucial. We do so by concatenating and averaging the accelerometer and gyroscope data over a window of 200ms. Prior to that, we resample and filter the readings from the IMU sensors. A further challenge comes from the orientation of the IMUs themselves. We obviate that by making our system independent from the human coordinate system. To do so, we only consider the intensity of the rotation, rather the motion components along the 3 axis of the accelerometer and the 3 of the gyroscope. In fact, we only leverage the motion components to tell whether the rotation is positive (to the right) or negative (to the left). As both the earbuds are left buds (only the left bud of the eSense platform is equipped with an IMU sensor) the x axis of one device is inverted compared to the other one.

Quaternions and complementary filter: Euler angles, better known by their components *yaw*, *pitch*, and *roll*, are the most common, and widely used, coordinate system to represents rotations. Despite their diffusion due to their ease of interpretability, they come with the problem known as *gimbal lock* [117] (Figure 3.3.1). To obviate the gimbal lock problem, it is common to switch to a better suited coordinate system: *quaternions* [30,118].

Integrated gyroscope measurements are subject to short-term drift, which may be more or less severe depending on the application. The situation worsen when we are integrating it over time, as the error grows faster. To mitigate that, a common approach is to fuse the gyroscope readings with the accelerometer ones, as the latter is known to be more stable than the gyroscope, in the short-time. Therefore, when estimating the head's motion only, leveraging gyroscope might yield extremely poor performances.

Our angular estimation is based on a complementary filter, which allows us to fuse gyroscope and accelerometer data, and follows the approach proposed by LaValle et al. [30]. We derive our estimation as follows:

$$q_{gyro} = \cos(\frac{\theta}{2}) + i\omega_x \sin(\frac{\theta}{2}) + j\omega_y \sin(\frac{\theta}{2}) + k\omega_z \sin(\frac{\theta}{2}) \tag{3.1}$$

where:

$$\omega = (\omega_x, \omega_y, \omega_z) = (\frac{gyro_x}{||\omega||}, \frac{gyro_y}{||\omega||}, \frac{gyro_z}{||\omega||})$$

, and

$$\theta = ||\omega||dt$$

with q_{gyro} being the quaternion that describes the instantaneous rotation of the head, based on the gyroscope data. Notably, if we were only using the gyroscope data to update the position estimation we would have stopped by doing:

$$pos[t] = pos[t-1] * q_{gyro} \tag{3.2}$$

Instead, we partially account for the gyroscope drift by adding the 3D orientation estimation and tilt correction of the head. To do that properly, we should rely on a combination of magnetometer and accelerometer data [14]. Instead, because of the limitations of today's earables which, like eSense, do not have a 9 DoF IMU, we could only rely on the gravity as an external reference to perform a rough orientation estimation and tilt correction.

$$\begin{aligned} q_{acc\_body} &= 0 + i\ acc_x + j\ acc_y + k\ acc_z \\ q_{acc\_world} &= pos[t] * q_{acc\_body} * pos[t]^{-1} \\ q_{cf} &= \cos(\tfrac{\phi}{2}) + i\tfrac{n_x}{||n||}\sin(\tfrac{\phi}{2}) + j\tfrac{n_y}{||n||}\sin(\tfrac{\phi}{2}) + k\tfrac{n_z}{||n||}\sin(\tfrac{\phi}{2}) \end{aligned} \tag{3.3}$$

where q_{cf} is the implementation of the complementary filter in the space of quaternions, and where:

$$\begin{aligned} \phi &= (1-\alpha)\arccos(\tfrac{q_{acc\_world_y}}{||q_{acc\_world}||}) \\ v &= (\tfrac{q_{acc\_world_x}}{||q_{acc\_world}||}, \tfrac{q_{acc\_world_y}}{||q_{acc\_world}||}, \tfrac{q_{acc\_world_z}}{||q_{acc\_world}||}) \\ n &= v * (0,1,0) \end{aligned}$$

We can now estimate the final rotation by doing:

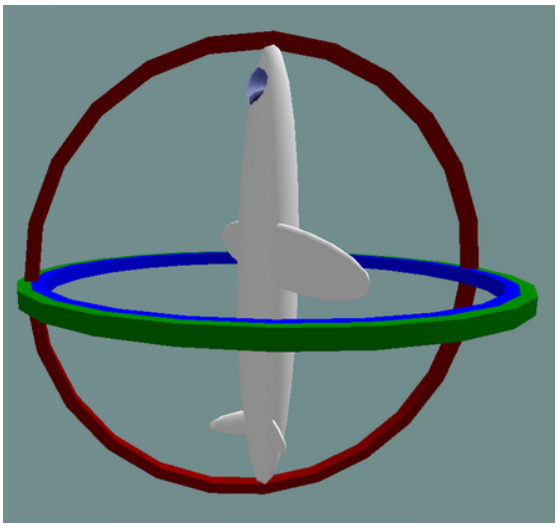


Fig. 3.3.1: Example of the Gimbal Lock: two out of three degrees of freedom collide on the same plane (graphic credits: MathsPoetry, CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0, via Wikimedia Commons).

$$final\_position = q_{cf} * pos[t] \tag{3.4}$$

Notice that for each earbud, we account for the factory offset of both accelerometer and gyroscope by using the techniques described by Kok et al. in their work [119]. In addition, because of the absence of the magnetometer, we only focus on relative rotations (which we refer to as *delta motions*).

3.4 User study and head motion tracking performance

In this section we describe the protocol we followed to investigate head motion tracking through earables. We detail how we collected the data, running a small (10 people) user study, and the results we obtained by tracking the head movements of our volunteers. Ethical approval was obtained to conduct the user study.

3.4.1 Data collection methodology

We recruited 10 volunteers to take part in our data collection campaign. Each individual was wearing two earbuds (both eSense left bud, as discussed in Section 3.3) connected

via BLE (Bluetooth Low Energy) to an Android application running on a smartphone[4] provided by us, the investigators. The eSense earables sampled data, and streamed to the smartphone, at $100Hz$. For the sake of reproducibility we report the configuration of the two earables. Noticeably, both earables run the same configuration:

- `AccelerometerRange` = $\pm$ `2g`
- `GyroscopeRange` = $\pm$ `500 degrees/second`
- `AccelerometerLowPassFilter = 5Hz`
- `GyroscopeLowPassFilter = 5Hz`

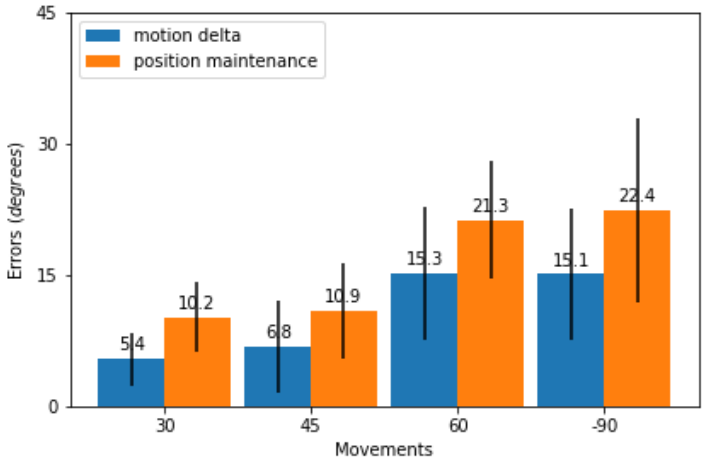
The experimental set up consisted of 4 targets (red cross) attached to the walls in an empty room. The angle between the position of the volunteer and the targets was known and represented ground truth. The targets were placed respectively at 30°, 45°, 60°, and −90°. We chose 30° as the smallest angle we investigate assuming that for smaller angles people would mostly move their eyes, barely moving their heads. With −90°, we wanted to show how our system could capture both clockwise and counterclockwise rotations of the head. Ultimately, we asked our volunteers to perform the following actions:

- Standing in silence and looking at different targets, and keep facing them, according to the instructions of the investigators;
- Standing, chewing a piece of chewing gum, and looking at different targets, and keep facing them, according to the instructions of the investigators;
- Standing, conversing with one of the researchers, and looking at different targets, and keep facing them, according to the instructions of the investigators.

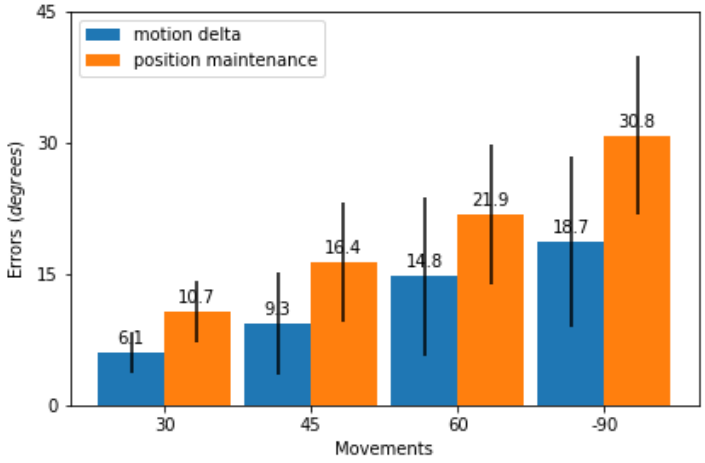
3.4.2 Baseline: silent subject

We now present the results of what we consider our baseline. In this experiment, the volunteers were standing in silence, looking at the different targets according to a set of instructions provided by us. For each target, the volunteers started facing an initial reference, placed at 0°. They then rotated their heads towards the given target (*delta motion*). Once there, we asked them to keep their head turned towards the target for about $5seconds$ (*position maintenance*). We estimated the movements done by the volunteers

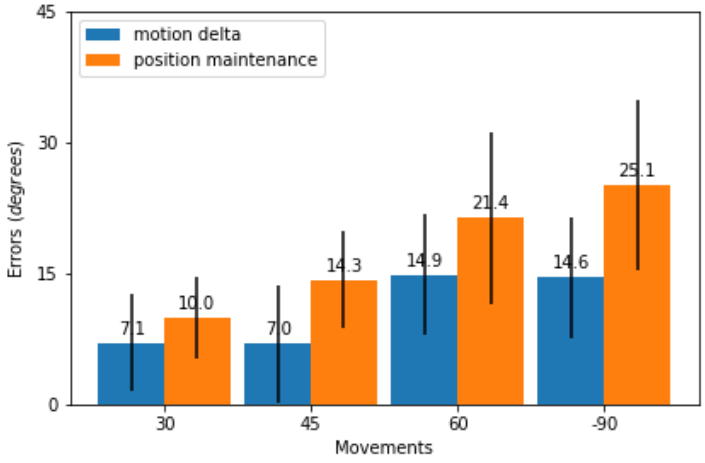
[4]Google Pixel 2, https://en.wikipedia.org/wiki/Pixel_2



(a)



(b)



(c)

Fig. 3.4.1: 3.4.1a Mean error and standard deviation of the head movements estimation of 10 silent volunteers. 3.4.1b Impact of chewing activity on the mean error and standard deviation of the head movements estimation of 10 volunteers. 3.4.1c Impact of speech on the mean error and standard deviation of the head movements estimation of 10 volunteers.

processing the readings of accelerometer and gyroscope according to what described in Section 3.3. Figure 3.4.1a respectively reports the mean errors of the motion delta and of the position maintenance estimations when the users were rotating their heads clockwise by 30°, 45°, and 60°, and counterclockwise towards the target at −90°. From the bar-plot in Figure 3.4.1a, we can immediately appreciate how the position maintenance errors, in orange, are greater than the motion delta ones, in blue. This is due to the inertial sensors' drift that heavily affects the integration. Because of our long term application, i.e., tracking visual attention, we are interested in instantaneous movements. Therefore, we mostly care about motion delta errors. Another interesting observation is how the errors grow for longer movements (i.e., greater angles), indicating a higher precision of the system in tracking shorter movements (i.e., small angles). Moreover, the high standard deviation that characterizes the mean errors denotes a strong user dependency of the motion estimation. In fact, for some volunteers, we even registered sub-degree motion delta accuracy in some movements.

3.4.3 Impact of chewing activity

Once we evaluated our system on the simplest case, we started testing the robustness of our motion estimation. We gave a chewing-gum to our volunteers, and asked them to repeat the same sequence of movements performed for the baseline case. The chewing activity generates spurious vibrations that are inevitably picked up by the inertial sensors in the earbuds. To make our system more robust to this kind of noise, we tuned the parameter α of our complementary filter, aiming for the best performance in all the three types of experiments. Figure 3.4.1b depicts the mean errors of the estimation. If compared to our baseline (Figure 3.4.1a), as expected, the errors are slightly higher. As we observed in the previous case, because of the drift, the motion delta mean errors are smaller than the position maintenance ones. Additionally, once again, the high standard deviation highlight the user dependency of the experiment.

3.4.4 Impact of Speech

Lastly, we assessed how speech impacts our estimations. To do so, we asked our volunteers to talk about a subject matter of their choice while performing the same set of head rotations described in Section 3.3. As for the chewing experiment, speaking generates unwanted vibrations and micro-movements that are captured by the sensors, and could

ultimately hinder the robustness of our head motion tracking system. Figure 3.4.1c shows how our system behaves with talking subjects. Notably, the mean errors of the estimation in this third case are sightly higher than the ones for silent users. As in both the previous cases, the motion delta accuracy is higher and decreases for longer movements. Comparing Figure 3.4.1b and Figure 3.4.1c, we can observe that the chewing activity seems to have a comparable impact on our motion estimation system as speech does. Finally, the high standard deviation of the mean errors further remarks the findings we got from the previous experiments.

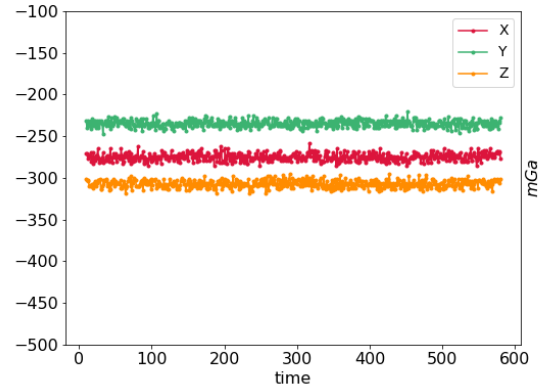
Summary of the results: Evaluating the performance of the eSense platform, our work investigates the potential of ears as a vantage point to sense visual attention through head movements. Overall, we achieve estimations with an average error that ranges from 5.4° for short movements in the least challenging situation, up to 18.7° for longer movements, under noisier circumstances. In the reminder of this chapter, we try to further improve the accuracy of our system, investigating whether it would be feasible add a magnetometer, thus gaining an external absolute reference point. As we will see in Section 3.6, featuring a magnetometer in an earable would allow us to recalibrate the IMUs and track absolute and incremental movements, in addition to relative motion.

3.5 Why is the Magnetometer Missing?

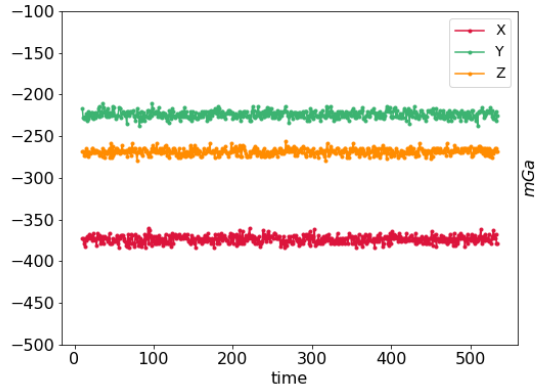
In order to asses if we could use state-of-the-art motion tracking approaches to further improve the precision of our motion estimation, we tried to push the boundaries of the platform we are investigating, and more in general, of all the existing commercial earables, by adding a magnetometer. Practically, we studied how the magnets used to hold the earbuds into the case and the magnet in the speaker affect the readings of a magnetometer. For this purpose, we initially placed a STEVAL-STLCS01V1 sensor tile[5] at different distances from one earbud, and we plotted the data captured by the magnetometer in Figure 3.5.1.

To start off, we put down the magnetometer at a distance of $5cm$ from the earbud (Figure 3.5.1a). We proceeded by keeping the earbud still in the initial position, while moving the sensor tile closer to the earable. Figure 3.5.1b shows the magnetometer readings when $3cm$ apart. We can immediately appreciate how the magnets start affecting the data, intro-

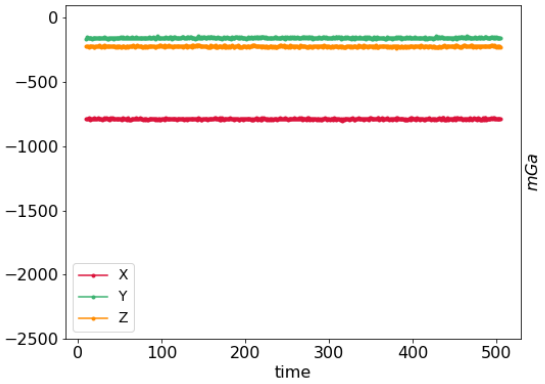
[5]https://www.st.com/en/evaluation-tools/steval-stlcs01v1.html



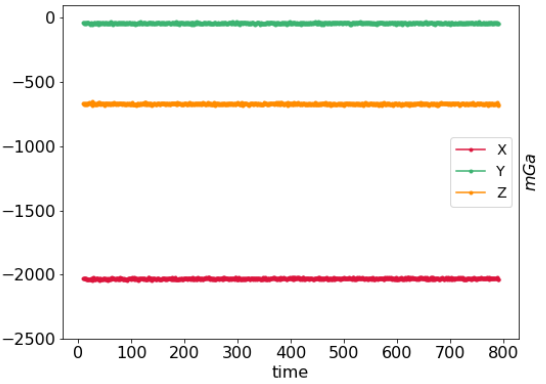
(a) 5cm from the earbud.



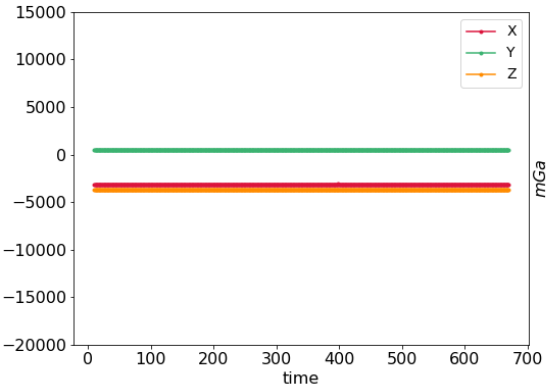
(b) 3cm from the earbud.



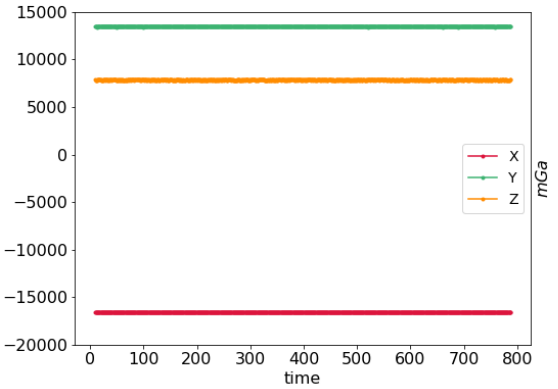
(c) 1.5cm from the earbud.



(d) 1.25cm from the earbud.

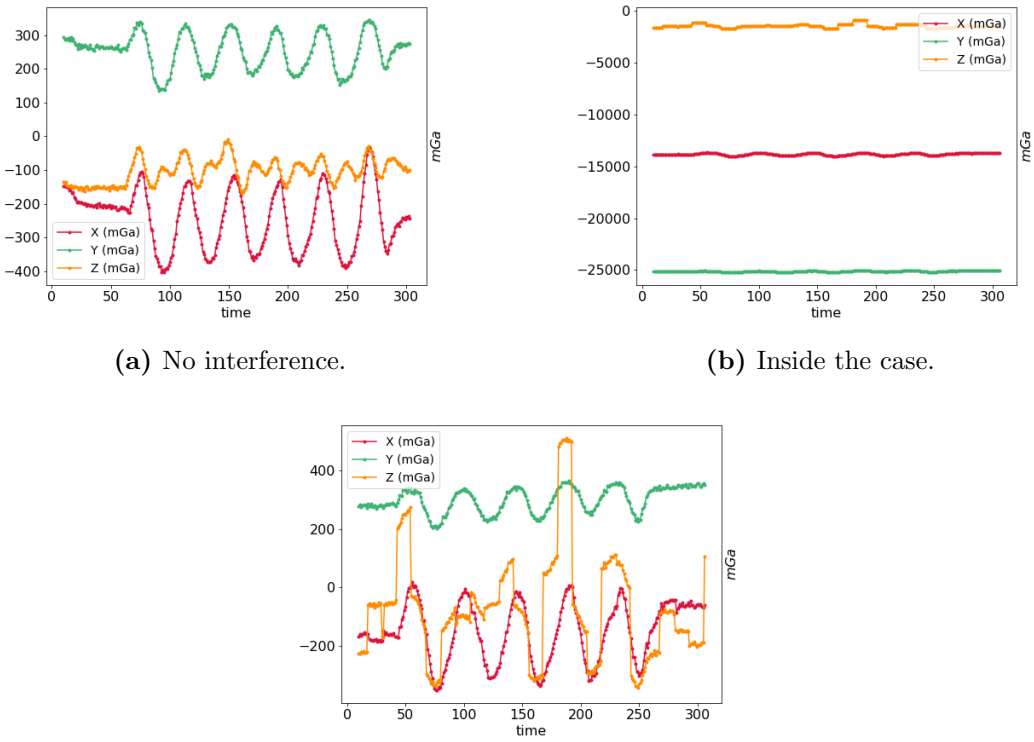


(e) 1cm from the earbud.



(f) Inside the earbud case.

Fig. 3.5.1: Magnetometer readings from a STEVAL-STLCS01V1 device at different distances from one eSense earbud. Notice how the scale of the plots changes dramatically from top to bottom.



(a) No interference.

(b) Inside the case.

(c) Distortion correction.

Fig. 3.5.2: We run two experiments where a volunteer was asked to shake his head while having the a STEVAL-STLCS01V1 close to the ear. (a) We first collected data without magnetic interference. (b) We then asked our volunteer to repeated the same movements with the STEVAL-STLCS01V1 placed inside the eSense case. (c) This way we could observe how correcting the offset introduced by the magnets, the magnetometer placed in the earbud is still able to record motion data.

ducing an offset. This offset is likely caused by hard iron distortions (as we will see discuss more in depth in Section 3.10). It is worth noticing how the offset does not fluctuate when the magnetometer is fixed. As expected, the readings change at different distances from the earbud. The closer we get to the earbud, the higher the influence of the magnets is. We further moved the STEVAL-STLCS01V1 at a distance of $1.50cm$, $1.25cm$, and $1cm$, as respectively depicted in Figure 3.5.1c, Figure 3.5.1d, and Figure 3.5.1e. Here, we can clearly notice how the magnetic field generated by the magnets in the earbud overtakes the Earth's one, flattening all the readings. Eventually, we put the sensor tile inside the case (Figure 3.5.1f). The readings skyrockets, as the earbud's magnetic field adds a significant offset to the Earth's.

Because of the constant offset at different locations, we decided to delve deeper into the behavior of the magnetometer, collecting more data samples while moving the device. We repeated the same set of movements twice. At first, we moved the STEVAL-STLCS01V1 alone, without any direct external interference caused by either the earbud or by the vicinity of a metallic source (Figure 3.5.2a). We then recorded the data, performing the very same movements, but placing the STEVAL-STLCS01V1 inside the case of the earbud, swapping it with the eSense's existing PCB (printed circuit board) (Figure 3.5.2b). While in the first instance, where there was not interference, the magnetometer was able to record the motion events, from Figure 3.5.2b we are unable to observe any motion-related data. However, the constant trend of the offset allows us apply a standard calibration technique to get rid of the hard iron distortion. As a result, we managed to recover most of the motion related information, especially along x and y (Figure 3.5.2c). These preliminary results provide an initial indication about the possibility of integrating a magnetometer even with the presence of strong magnets in the sensor's vicinity. Yet, they also fuels the need for an earable specific, hassle-free, calibration, in order to preserve the information carried by the magnetometer readings. However, experiments with different conditions (e.g., when there is music playback) and a more detailed analysis of the resulting data are needed to confirm our initial exploration. Hence, in the reminder of this chapter, we further characterize the magnetic interference experienced by a magnetometer in an earable (Section 3.7) and we devise an user-transparent calibration routine to counteract it (Section 3.8).

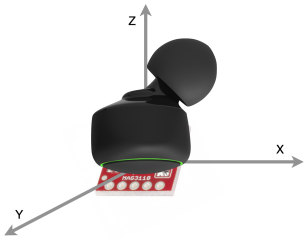
3.6 Enabling in-ear magnetic sensing: motivation

After evaluating eSense as an earable platform to perform in-ear head motion tracking we have observed that the accuracy of our estimation decreases for longer movements. To improve the performance of our system, we build up on our preliminary study on the magnetometer which, we believe, could represents an interesting avenue to take into account in the development of future earables. Hence, in this section we first enumerate the reasons why IMU data are often coupled with magnetometer readings. We then provide the reader with a basic understanding on the possible sources of interference that may affect a magnetometer in an earable.

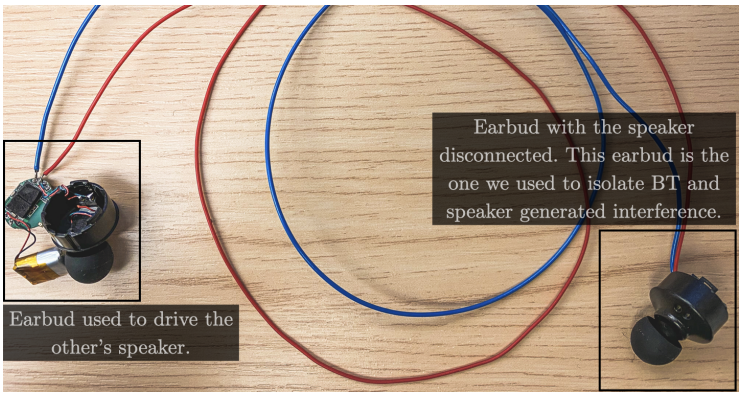
Magnetometer and inertial sensing: Historically, the presence of a magnetometer has been key to improve the accuracy of inertial based applications. Inertial sensors drift significantly when integrated over time [120]. For this reason, IMUs are often paired

with a magnetometer: while the former measures relative motions (i.e., linear acceleration and rotational velocity), magnetometers sense the Earth Magnetic Field, and are used to find the (absolute) direction of the *Magnetic North* in a global reference frame. This constitutes an absolute anchor to be constantly re-calibrate IMUs. Further, magnetometers are also coupled with IMUs for 3D-motion tracking: without a magnetometer, it becomes extremely hard to have knowledge of the tracked object's heading in a global reference frame [14] (used to correctly initialize the tracking system). Unfortunately, neither IMU calibration, nor 3D-motion tracking (e.g., of the head) are feasible with today's earables, which lack a magnetometer. Fusing the user's smartphone magnetometer data and IMU readings from their earables would naturally result in a wrong estimation, given how user's head and phone often face in different directions. There are many compelling use-cases that can be unlocked from a magnetometer in earables, provided the magnetometer is accurate enough. Concretely, an in-earable magnetometer could enable acoustic AR [47, 121,122] providing precise navigation thanks to spatial audio based on head orientation. For instance, considering a 4 lanes intersection (≈10 meters wide), an error of 3° on the heading would entail an offset of $10 \times sin(1.5) \approx 0.26m$. Further, by leveraging head rotations to compute the angle of arrival (AoR) of incoming sounds with higher accuracy [123], it could provide improved noise-cancellation. To be effective, AoR estimation errors should be $< 20°$ [124]. This requirement becomes more stringent when using the earable for immersive audio or speaker isolation, especially in multi-source conversations. In fact, to vouch for real-time spatial audio features that could be used for immersive audio or speaker isolation if the earable is being used as a hearing aid require tight tolerances in the in the attitude estimation. Concretely, at a $2meters$ distance the estimation difference between 3° and 10° error is more than $1meter$, thus reducing the usability of such features, especially in multi-source conversation setting.

Magnetic interference: There are various sources of magnetic interference which would impact a magnetometer in an earable, all being implicitly linked to the user's patterns. Firstly, earbuds usually have one (to drive the speaker) or more magnets (for docking purposes). These, being in close proximity to the magnetometer due to the earable's form factor, can interfere with its readings [33], making inference tasks unreliable at best [36]. Further, earables mostly communicate via Bluetooth (BT). Radio frequencies (RF) communications, like BT, require substantial electrical current which, flowing in the circuitry, generates an electromagnetic field, interfering with the magnetometer. Specifically with earables, RF communications are intense while streaming and just a few beacons otherwise.



(a)



(b)

Fig. 3.6.1: Basic system setup (Figure 3.6.1a) and setup used to isolate the interference generated by BT streaming and speaker (Figure 3.6.1b).

Practically, BT requires variable current, thus generating variable magnetic fields [33], entailing a hard-to-model electromagnetic interference, highly dependent on the users' patterns. Similarly, when playing music the speaker coil vibrations that generate sound also result in magnetic interference. Like sound, the interference depends on the vibration patterns, and therefore on whether the user is playing music, and what music they are playing. To gain a deeper understanding of how this practically affects a magnetometer in an earable, we present a thorough analysis of these interfering phenomena (Section 3.7). We look at RF communications (i.e., music streaming form a smartphone to the earbuds), as well as music playback. Besides the popularity of the task, music playback was selected since we could reproduce a wide spectrum of tonal patterns playing different music genres, allowing us to quantify interference in the presence of contrasting and unique tonal patterns. We analyze the impact of voice calls, too. Contrary to the music playback, modeling orthogonal interference patterns is limited due to the inherent frequency range similarity of the human voice. Interestingly, these patterns, while present, are eclipsed by the interference produced by the RF circuitry (Section 3.7.2).

3.7 Magnetometer Calibration and Interference

Proper calibration is key to obtain accurate sensor readings as we have seen in Section 3.5. This is especially true for continuous magnetic sensing: whether we are looking for the heading of an object [37, 42], or tracking magnetic bodies [112–114], reliable sensor data is crucial.

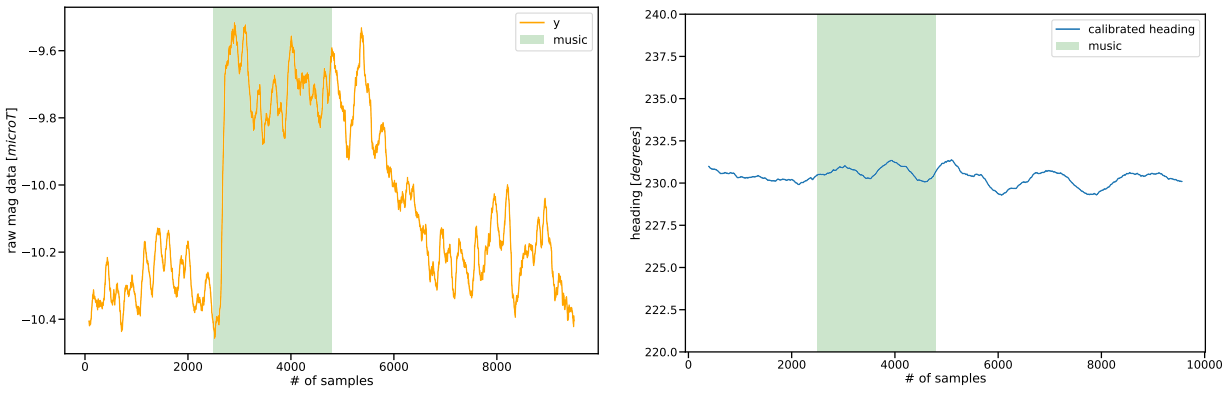
3.7.1 Magnetometer Calibration

To calibrate a magnetometer, common approaches seek to estimate the *bias* and *scale factor* for each axes. Magnetometer calibrations can be grouped in static or dynamic [34] and whether they rely on attitude information or not [35]. However, one major obstacle in calibration is interference from other magnetic fields. Concretely, this presents two challenges during calibration, i) interference can rarely be detected thus resulting in incorrect calibration parameters and ii) during use, where interference can result in incorrect bearing estimations.

3.7.2 Interference Characterization

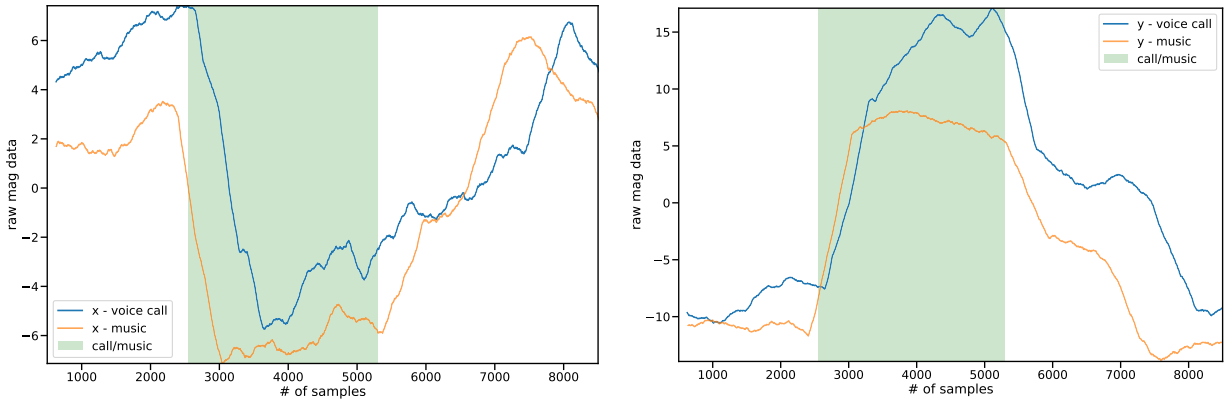
Testing device: As previously (Section 3.5), an example of earable, we chose eSense [2]. Contrary to other commercial earbuds, eSense permits access to the raw data streamed by the IMU and BLE chips, which is used to stream data and to send periodic beacons that can be used to detect proximity to nearby devices. Since eSense was not equipped with a 3-axis magnetometer, we attached an external one (Freescale MAG3110 [125]) on top of the earbud (Figure 3.6.1a) to mimic a realistic position where the sensor might be placed. We sampled the external magnetometer at $80Hz$ using an Arduino Uno. Similarly to what we did with the STEVAL-STLCS01V1, this setup allowed us to collect data while the earbud was performing operations that could interfere with the sensor (i.e., music streaming or phone calls). Notably, we opted for the MAG3110 over the STEVAL-STLCS01V1 because of the smaller form factor which allowed us to further experiment with it. In the remainder of the section we describe the conditions we have explored.

Results: We initially looked at the effect of streaming audio to the earbud on the magnetometer. This requires both active speaker movement and significant electrical current associated with the earbud circuitry. Figure 3.7.1a shows the effect on the y axis of a pre-calibrated magnetometer before, during (green shaded zone), and after playing music. Playing audio introduced an offset of 10 μT. This is a fraction of the Earth's typical magnetic field, hence the effect on the heading estimate was minimal Figure 3.7.1b. Nonetheless, there is a clear audio-induced change. Analogous trend was also evident in the magnetometer signals during voice calls. Figure 3.7.2 illustrates the changes in magnetometer x and y signals measured during a voice call (green shaded region), overlaid with the changes associated with playing some music in post-processing. Both signals show a clear audio-related change. We can further observe that all of the disturbances were re-



(a) (b)

Fig. 3.7.1: Impact of audio streaming and consequent music playback on the magnetometers readings.



(a) (b)

Fig. 3.7.2: Impact of voice calls and music playback on the magnetometer readings.

versed when the audio was stopped, but that this was *not* immediate. Rather, we observed a gradual return to the pre-audio values.

From this we can conclude that audio playback in an earable affects its magnetometer signal. While we observed a relatively small change, it is enough that a one-off calibration procedure during audio playback is to be avoided. Besides, it may also hinder applications that do not use the magnetometer for heading (e.g., magnetic map matching). Furthermore, magnetic fields dissipate quickly with distance and a magnetometer soldered to the earbud board (beyond the scope of this work) might experience greater disruption and heading errors.

To better understand the source of the interference, we examined the frequency spectrum

of the magnetometer when playing pure tones. We observed that tones played at (even very) low volume produced noise across all frequencies in the magnetometer. Increasing the volume resulted in additional spikes corresponding to the tone frequency: e.g., a $20Hz$ tone produces a corresponding spike at $20Hz$ (Figure 3.7.3). From this we infer that the interference comes partially from the audio playback (speaker/driver circuit produces the spikes) and partly from some non-specific part of the circuitry (giving the general noise). The slow reversal of the audio-induced changes when the audio stops is further evidence that the interference is not solely due to the sound reproduction.

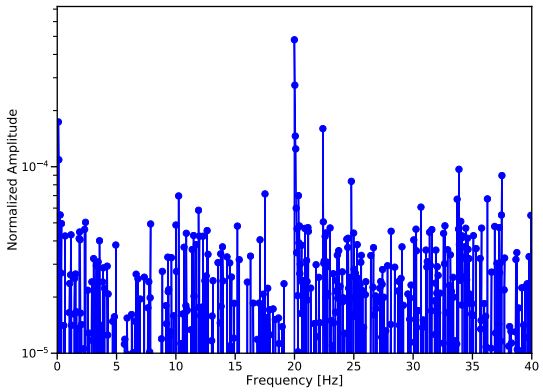


Fig. 3.7.3

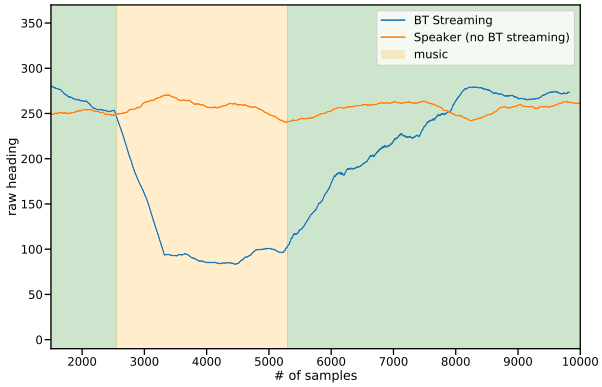


Fig. 3.7.4

Fig. 3.7.5: 3.7.3 `FFT` of the magnetometer traces while playing (on high volume) pure tones at $20Hz$. 3.7.4 Impact of the interference generated by BT streaming and by the speaker on the heading estimated by the raw magnetometer traces. Although the magnitude of the interference on calibrated traces is smaller, for clarity in this figure we use raw readings.

We sought to isolate the core circuitry from the speaker. We used two earbuds, A and B. First, we unsoldered and removed the speaker from A. When streaming audio to A, we observed wideband noise in the magnetometer, persisting beyond the end of the music (Figure 3.7.4). In this mode, the main operational circuitry is the BT module and the audio decoder. Since the latter is not used after the music stops, we focused on the BT module. We used a packet sniffer to establish when BT radio was in use, finding that BT packets continued to be sent for a period *after* the audio was manually stopped. This period corresponded directly to the magnetometer recovery phase. We can therefore attribute a substantial part of the interference to the Bluetooth circuitry being active. We then used two wires to connect the speaker terminals on A to the speaker terminals on B. In this way we could assess the impact caused by the speaker alone without interference introduced by the circuitry (Figure 3.6.1b). We powered both earbuds on, and then streamed music

to A. This caused music to be played on B, where only the core circuitry was active (not BT or other components). The magnetometer on B exhibited a small deviation when music was played, but significantly smaller than that observed when isolating the BT hardware. While the interference induced by the BT is almost identical for different songs, it changes between music players (e.g., Spotify, Apple Music, YouTube) as they adopt different protocols.

Findings Summary: Streaming audio to a modified eSense earbud resulted in a local magnetic field that appeared as interference in the magnetometer readings. The magnitude of the interference was small, but we cannot rule out a larger effect for a magnetometer fully integrated onto the earbud. The interference came primarily from the BT circuitry being active, with a smaller component due to the speaker. Therefore, static magnetometer calibrations in an earbud should not be carried out while the BT radio is active.

3.8 A Novel Calibration Algorithm

The previous section established that internal components of an earbud introduce magnetic interference during RF usage, but that it is possible to incorporate a magnetometer such that the heading estimate is minimally affected. However, this is contingent on the magnetometer being either correctly calibrated prior to the interference, or having the calibration dynamically updated, something that is impractical due to the manual calibration procedures conventionally used. In this section we describe a technique to provide calibration that does not need manual intervention and can be updated dynamically.

3.8.1 Overview

Performing user-transparent magnetometer calibration is an extremely challenging task [120]. Rather than trying to calibrate earbuds independently, we propose leveraging the user's smartphone to assist in the calibration. Our technique assumes the phone has itself a calibrated magnetometer and, therefore, can provide reliable global heading. In practice, this assumption is justified since today's phones are able to maintain a calibrated heading by fusing the array of sensors, from IMU to GPS [126]. Besides, whenever a phone does not have a calibrated magnetometer, it is typically capable to notice it and will request manual intervention to the user. Our key observation is that when people interact with their smartphones, e.g., unlocking it, their head is almost certainly facing the phone. In this

case the smartphone and the earbuds are aligned and should report the same bearing, if correctly calibrated (Figure 3.8.1). Our approach is to use trusted bearing of the phone to estimate the calibration parameters for the earbuds. Any calibration routine must be done quickly, while the user is looking at their phone. Empirically, we find that unlocking an *iPhone* using the *FaceID* usually constitutes the perfect user head-smartphone positional relationship. This work considers unlock interactions, but the technique could be applied whenever phone and head align.

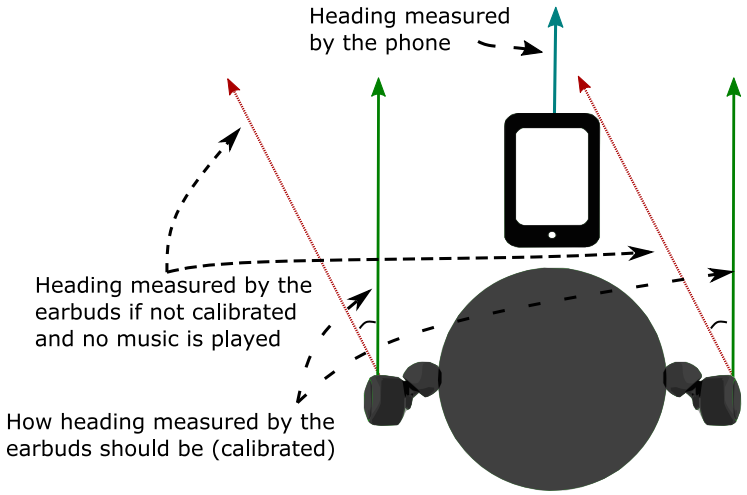


Fig. 3.8.1: Intuition behind the proposed calibration technique.

3.8.2 Calibration Approximation

We estimate the heading (the angle between the direction of facing and the Magnetic North) by doing:

$$heading = \text{atan2}(mag_y, mag_x)\frac{180}{\pi}. \tag{3.5}$$

This is the standard way of estimating the heading given mag_x, mag_y – the leveled (with respect to the ground) readings of the magnetometer respectively along the x and y axis. The sign, as well as the order, of mag_y and mag_x change depending on the magnetometer orientation. Several commercial earables ensure the earbuds remains in a still, standard position in the users' ears. Hence, the orientation of each earbud is likely to remain unaltered for most of the time. We leverage this, and the fact that most of the substantial rotations and changes of heading happen along a plane parallel to the ground, to avoid continuously

accounting for tilt compensation. Having defined how we estimate the heading, we can lay the foundations of our approach. Firstly, we apply a standard sensor model for mag_x and mag_y respectively as:

$$mag_x = S_x(x_{earbud\_raw} - x_{earbud\_offset})$$

and

$$mag_y = S_y(y_{earbud\_raw} - y_{earbud\_offset}),$$

where x_{earbud_raw} and y_{earbud_raw} are the raw, uncalibrated magnetometer readings in the xy plane, S_x and S_y are scaling factors and x_{earbud_offset} and y_{earbud_offset} are constant offsets. According to our key assumption (Figure 3.8.1), we can re-write the way we compute the heading as:

$$h_{phone} = \text{ atan}\left(R\frac{(y_{earbud\_raw} - y_{earbud\_offset})}{(x_{earbud\_raw} - x_{earbud\_offset})}\right)\frac{180}{\pi}, \tag{3.6}$$

where h_phone is the phone's heading estimate and $R = S_y/S_x$ the scale factors ratio. Our goal is to collect multiple h_{phone} values to solve for the unknowns in this equation. We make the assumption that $R \approx 1$. We expect sensors to be factory calibrated, that should ensure this approximate relationship (more in Section 3.10). Small perturbations from 1 have minimal effect on the heading ($\arctan(x + \delta x) \approx \arctan(x)$ for small δx). The assumption allows us to reduce the unknowns to two (y_{earbud_offset} and x_{earbud_offset}), requiring as few as $k = 2$ phone interactions to estimate a calibration. In practice, we gather as many phone headings (k) as needed to ensure good calibration quality:

$$\begin{aligned}
h_{phone_1} &= \arctan\left(\frac{y_{earbud_raw_1} - y_{earbud_offset}}{x_{earbud_raw_1} - x_{earbud_offset}}\right)\frac{180}{\pi} \\
h_{phone_2} &= \arctan\left(\frac{y_{earbud_raw_2} - y_{earbud_offset}}{x_{earbud_raw_2} - x_{earbud_offset}}\right)\frac{180}{\pi} \\
h_{phone_3} &= \arctan\left(\frac{y_{earbud_raw_3} - y_{earbud_offset}}{x_{earbud_raw_3} - x_{earbud_offset}}\right)\frac{180}{\pi} \\
&\dots \\
h_{phone_k} &= \text{ atan}\left(\frac{y_{earbud_raw_k} - y_{earbud_offset}}{x_{earbud_raw_k} - x_{earbud_offset}}\right)\frac{180}{\pi}
\end{aligned} \tag{3.7}$$

By solving the over-determined systems of equations constituted by the k-th phone's headings (Equation (3.7)), we derive x_{earbud_offset} and y_{earbud_offset}, continuously updated at every interaction. Collecting these *reference* measurements from the phone can occur in the background, without any intervention from the user. Phone readings have to be reliable. In a smartphone the magnetometers are assumed to be calibrated, as both Android and Apple devices fuse the magnetometer readings with GPS (if available) [126]. A new reference is found every time users interact with their phone (i.e., *FaceID* unlock), provided the phone measures an undisturbed magnetic field (i.e., trustworthy heading). This is further borne out by the high average number of daily interaction people have with their mobile phones [127]. While more references are good as they should lead to a better model fit, calibrating with fewer references is valuable. A good fit depends on there being sufficiently distinct h_{phone_k} values, and we do not have control over the users behavioral patterns. Hence, we favor calibrating as soon as possible and refining the calibration with extra measurements later on, without explicit user interaction.

3.8.3 Calibration Algorithm

The overall functioning of the our calibration procedure is depicted in Figure 3.8.2. At any given time, we monitor for events (e.g. phone-pickup/unlocking) that can potentially provide measurements suitable for calibration. Although not being limited to unlock events only, often the suitability coincides with, for example, facial unlocking, as the phone is aligned with the earbuds. These events occur fairly frequently: to be truly user-transparent, our calibration methodology fully leverages people's smartphone usage patterns [127]. Given the variability of the interference and the consequent volatility characterizing magnetometer calibrations [120], rather than aiming at the perfect calibration routine, instead we strive for the best possible approximation of it. In this way, we can afford to calibrate as soon as possible and continuously monitor the status of the calibration, updating it when necessary. In addition, by using the phone's heading as a reference, there is no need to further process the heading estimated from the magnetometer data in the earbuds by calculating the magnetic declination [128].

Phone pickups and data check: Once a phone-pickup is detected, we perform a data check to ensure the smartphone's data are suitable for the calibration: first, we ensure the magnitude of the phone's readings matches, at least in the order of magnitude, that of the Earth's field at the current location. This is a standard way to check the magnetometer

readings trustworthiness. Once secured we are not in a magnetic anomaly, we make sure the phone is in portrait mode. Although this is true for most of the interactions, to calibrate we can not afford using a reference off by 90° (i.e., phone in landscape mode). Lastly, we verify no sharp head movements occurred, which may result in misalignment in the phone's and earbuds' heading, or if the user moved the head but not the phone. These steps are key to ensure the reference heading is truthful and of good quality.

Calibration Execution: If the data pass this check, we store them and wait until we have enough references to perform our calibration. At least two equations are needed to linearly solve a system of equations with two unknowns (the offsets). Hence, we need at least two reference headings (Section 3.9). Once satisfied the number entries to perform the calibration, we execute it by applying a least squares fitting (eq. (3.7)). Before finalizing the calibration, we perform a sanity check to ensure there was not any significant interference skewing the fit: we compare, on-the-fly, the instantaneous heading recorded by the phone and the average bearing of the earbuds. The calibration is only committed after the freshly-calibrated earbud magnetometer successfully passes this additional step.

Calibration check and update: In an ideal scenario, we would know that after a certain amount of time Δt, or a certain displacement in space Δx, there is the need for re-calibrate our sensors. However, in reality, this is not possible. Unfortunately, accurately modeling the *life time* of a magnetometer calibration (i.e., how long the calibration will last before the sensor readings will start being off) is extremely difficult. A number of factors can invalid the calibration of a magnetometer, such as the environment, the temperature, and the number of people in a room [129]. We avoid faulty calibration models by regularly checking the validity of the calibration. This is an inexpensive operation we carry out in two ways, depending whether there is a phone unlock event, or not. If the positional relationship between the user's head and their smartphone is the same that satisfies our pre-calibration checks, we compare the bearing of the phone and that reported by the earbuds. If their difference is under a certain threshold, we assume our calibration is still valid, otherwise we drop the existing calibration. The value of the threshold depends on the application and the desired accuracy. Alternatively, if no phone-interaction is detected, we cannot assume the phone's heading is the same of the earbuds'. In this eventuality, we compare the earbud's magnetometer bearing with the earbud's IMU. If there are sharp changes in the magnetometer data, but no rotations or linear accelerations are registered by the IMU (and vice-versa), the existing calibration is likely off. Notably, we look for the magnitude of the motion recorded by the gyroscope and we compare it with the change in

bearing reported by the magnetometer. Even if uncalibrated, gyroscopes are fairly precise in measuring relative motion, while they fail in tracking sustained motion.

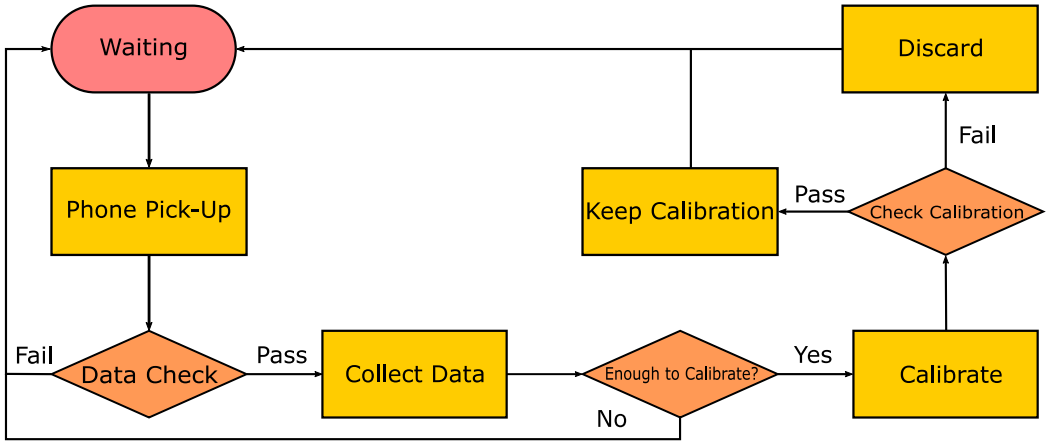


Fig. 3.8.2: Auto-Calibration procedure: the calibration watcher checks if the phone is picked up (i.e., during an unlock event). If such event occurs, and the phone is unlocked in a position suitable for calibration, then the data is collected and stored. Once we have enough data to calibrate (i.e., at least two), then we perform the calibration procedure as described in Equation (3.7). If successful, before applying the updated calibration values we perform a sanity check to eliminate any potential interference or bad measurements; which, if found present, the newly calibrated parameters are discarded.

3.9 Calibration evaluation

In this section we provide a detailed evaluation of the proposed calibration procedure. We evaluate our system in controlled conditions by looking at how accurate the calibration is with respect to static angles as well as tracking angular movement. We look at a series of variables that may affect the accuracy of the calibration. Further, we assess the performance of our system in-the-wild with a case study. We conclude by presenting some theoretical considerations on the proposed algorithm complexity, supporting this analysis with power consumption experimental results of our calibration routine.

3.9.1 Micro benchmarks

We start our evaluation with a list of micro benchmarks. Figure 3.9.1a reports the setup we used to benchmark our calibration. Once assessed how the calibration removes the interference caused by both BT and music playback using the hardware described in Section 3.7, we focus on benchmarking the calibration technique by using the magnetometer embedded

(a) **(b)**

Fig. 3.9.1: Setup used to benchmark the proposed calibration technique (3.9.1a) and volunteer wearing the Arduino as if they were earbuds (3.9.1b). This is the setup used for our in-the-wild use test.

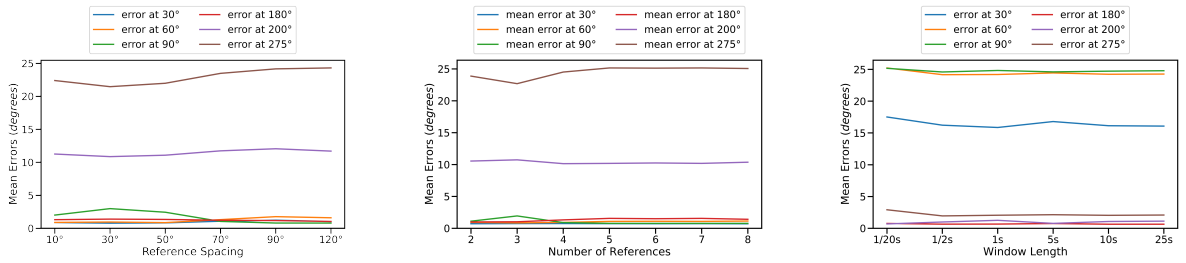
in an Arduino Nano 33BLE [130]. For reproducibility, we build a stand (Figure 3.9.1a) to simulate the positional relationship between the magnetometer and the smartphone (iPhone 8Plus).

Static case

We begin assessing how different factors may affect the calibration accuracy. We look at the impact of spacing between references, the number of references, and the number of data points fed into the calibration algorithm. We record ground truth data, as well as raw readings from a magnetometer, for a set of angles (30°, 60°, 90°, 180°, 200°, and 275°). All the measurements are done in static conditions with the set up described above (Figure 3.9.1a).

Difference in heading between references: The bearing of a person (i.e., the direction in which the subject is facing with respect to the Magnetic North) at two phone pickups is likely to be different. This difference in heading is also likely to vary, e.g., the difference between the first 2 pickups could be 20°, with $h_{phone_1} = 10°$ and $h_{phone_2} = 30°$, whereas could be 50° between the second and the third pickup, with $h_{phone_3} = 80°$. We will refer to this difference in heading as *spacing* between reference headings (i.e., $spacing(i, i+1) = |h_{phone_i} - h_{phone_{i+1}}|$). In this set of experiments, we look at how different reference headings spacing conditions impact calibration accuracy. We set $h_{phone_1} = 10$ and we vary h_{phone_2}. Our results are reported in Figure 3.9.2a: we plot the average errors (aggregated per testing angle) between the reliable smartphone heading and that computed applying our calibration to the uncalibrated magnetometer traces, while changing the spacing between two references. We do not observe any significant gain with larger spacing of the reference headings. Heading spacing has far less impact than we originally thought; however, the choice of the references affects the overall accuracy which is high for small angles (using *small* references), yet decreases for bigger ones.

Number of heading references: Similarly, we evaluate the impact of multiple heading references on the accuracy of our calibration technique. We pick $h_{phone_1} = 30$ and we add references spaced by 20° (Figure 3.9.2b). We noticed that we do not require more than $2 - 3$ reference headings, without getting into the territory of diminishing returns – i.e., paying more in terms of energy consumption (increasing the number of reference headings inevitably results in greater complexity – Section 3.9.3) without getting a similar accuracy boost. This desirable property provides usability benefits, allowing us to perform the



(a) **(b)** **(c)**

Fig. 3.9.2: Errors aggregated per testing angle when changing: 3.9.2a the spacing between two reference headings; 3.9.2b increasing the number of references; 3.9.2c the number of data-points fed to our algorithm.

calibration more often, as very few reference headings are required for each run, resulting in increased stability over time and more robustness to interference variability.

Number of magnetometer data points: Of particular importance is the different trend in Figure 3.9.2c when compared to Figure 3.9.2b as we observe higher angles, (200°, 270°) to be *more* precise; this is intentional as for this particular experiment we use higher reference headings (250° and 300°).Notably, we observe that the need for only a few data points, together with the fact we require $2-3$ reference headings to perform a calibration precise up to few degrees (sometimes even sub-degree) makes our user-transparent calibration extremely appealing. In fact, these are key indicators of how inexpensive and easy to perform our calibration is: concretely, calibrating becomes just a matter of solving a system of $2-3$ equations with 2 unknowns. This being an extremely easy task for today's processors. Because of that, we can easily afford to calibrate more often increasing the overall accuracy.

Dynamic case: angular movement

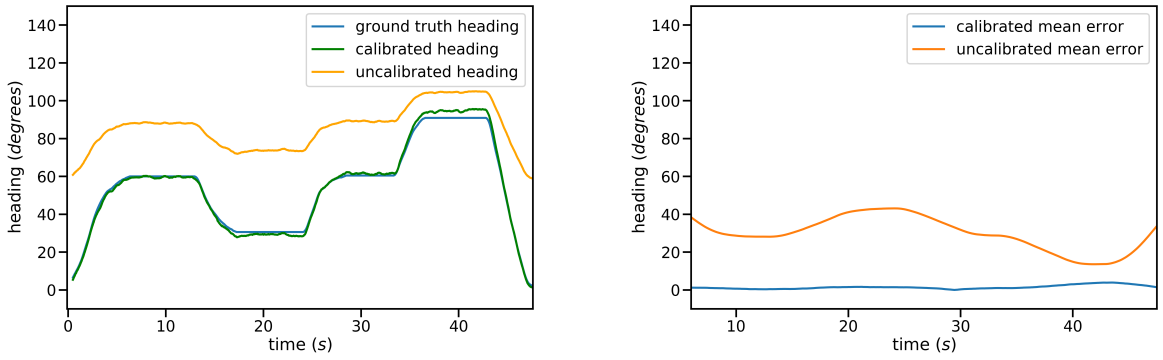
After testing our calibration technique under static conditions, we look at impact of rotational movements. Key requirement for a calibration is to work regardless the device is still, facing in a single, static, direction, or in motion. For instance, if we consider a person rotating their head, their bearing would change of an angle equal to that of head rotation. A calibration must remain valid for every direction faced during the movement. Therefore, we evaluate whether our calibration can properly keep track of motion. We do that by using the stand in Figure 3.9.1a. This way, we can constantly record both the magnetometer data at any given orientation, and the ground truth from the phone. Concretely, we look

at whether the calibrated heading diverge from the ground truth whenever there is motion. We perform movements of different amplitudes, simulating head rotations, and plot the heading and the average errors. For all the experiments, we use only two references, collected over windows of $1/20s$.

Practically, we begin by looking at small movements of a few degrees (both positive and negative) and outline our findings in Figure 3.9.3a. The uncalibrated error is significant, averaging 30° over the duration of our experiment (mean rolling error in Figure 3.9.3b), being at best is 10° off ground truth and at worst 45°. Besides, the magnitude of the motions seems not to reflect those of the real movements. Conversely, we can observe that the calibrated heading trace is very close to the ground truth heading averaging an error smaller than 5° at any given time, resulting in a very stable heading trace in terms of its error, providing a significant improvement overall. Analogously, Figure 3.9.4a reports the performance of our technique for larger movements. We expect that sharp or large movement to induce greater errors in the magnetometer heading than subtle, smaller ones – precisely depicted by the mean error over time (Figure 3.9.4b). The error of the uncalibrated heading trace is at one point in excess of over 90°, providing unusable data for most applications. On the other hand, again, we observe that the calibrated heading trace closely follows the ground truth, with an overall mean error of just a few degrees, with the error always *smaller* than 5°.

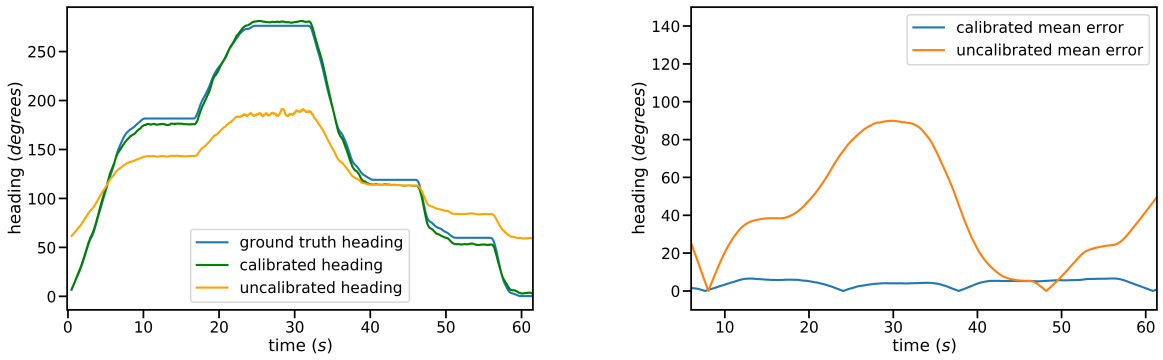
3.9.2 In-the-wild example case-study: navigation

We believe, in the case of navigation, earables could be better suited than smartphones. The rationale behind that lays in the ability of earbuds to track head movements: a desirable feature at complex intersections. Besides, contrary to smartphones which when in the pocket often mess up with the person's direction (the phone moving in the pocket keeps changing orientation), earables, once in the ears, are in a relatively stable position, less prone to sharp changes – unless the user is moving. Further, earables can provide extra robustness by recording 2 independent measurements of the same heading. This can be exploited by aggregating them to enhance accuracy or scheduling them to maximize battery life. Notice that an end-to-end earable-based navigation system is out of the scope of this work, instead, we use this toy example to evaluate, in-the-wild, our calibration. In this work we focus on evaluating our calibration on a case study, rather than building a complete end-to-end earable-based navigation system.



(a) **(b)**

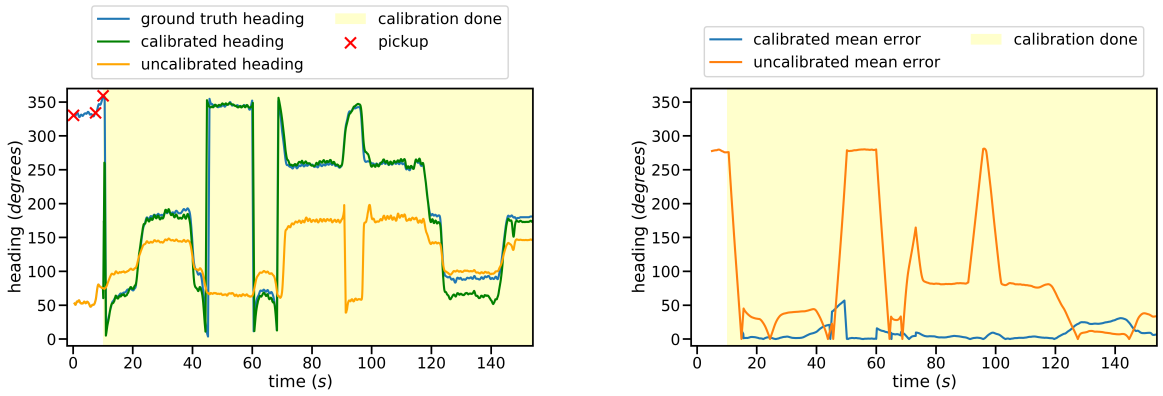
Fig. 3.9.3: Heading estimation (3.9.3a) and mean errors (3.9.3b) for small angles.



(a) **(b)**

Fig. 3.9.4: Heading estimation (3.9.4a) and mean errors (3.9.4b) for large angles.

Assessing the goodness of our calibration in-the-wild, we deal with both rotational movements and human motion (i.e., linear acceleration). Concretely, a user (ethics approval grated by the departmental ethics board) *wore* two Arduino (with a build-in magnetometer) as earbuds (Figure 3.9.1b), while holding the phone in their hands. The volunteer was told to walk as desired in two distinct locations the first one being indoor (in a house) while the latter outdoor (over a block). We had no control over the potential source of interference in the environment. This experiment showcases how the proposed calibration is capable of enabling earable-based in-the-wild navigation, without constraining nor bounding the user, both indoor and outdoor. Figure 3.9.5 reports our results when estimating the heading with calibrated and uncalibrated magnetometer traces and their average errors. Through our calibration, we are able to achieve accuracy up to few degrees (below 3° for most of the time) for the whole duration of the experiment. Notably, considering the 4



(a) **(b)**

Fig. 3.9.5: In-the-wild heading estimation (3.9.5a) and mean errors (3.9.5b).

lanes intersection example, our $\approx 3^\circ$ error on the heading would lead to a $\approx \pm 0.26m$ error. We believe $\pm 0.26m$ is an acceptable tolerance for pedestrian navigation.

3.9.3 Computational and power consumption considerations

A calibration has to be accurate, yet computationally inexpensive; however, all of the established calibration techniques require to perform a regression to fit a model based on the observations received. Especially, online schemes do the fitting every time the calibration is performed; in our case, as dictated from Figure 3.8.2, we check if a calibration is needed at every suitable phone pickup. Normally, such techniques use a form of least-squares fitting having a formal complexity of $\mathcal{O}(c^2 n)$, where c the number of features and n the number of vectors in $\mathbb{R}^c$ used when performing the fitting. However, in our case, we require very a few vectors $n < 10$, all in $\mathbb{R}^2$, representing the readings of the magnetometer on the xy plane. Hence, the amortised complexity of our procedure ends up being even more affordable in practice, even if we have to run the calibration procedure often. We evaluated this claim with a power consumption experiment with a Raspberry Pi Zero to gauge the potential overhead of our calibration procedure. Set side by side to the overhead of radio communications over idle ($\approx 1.7\%$ TX and $\approx 2.4\%$ RX), our scheme imposes a comparable low overhead, only consuming an additional $\approx 2.9\%$ over idle, with a total of $456.38mW$ (idle: $442.58mW$).

3.10 Discussion and limitations

This dissertation proposes a method to calibrate in-ear magnetometers in a user-transparent and efficient way and shows the feasibility of having magnetometers in earables. This section discusses the limitations of our proposed calibration technique, further motivates the rationale behind our choices, and sheds a light onto possible future directions.

Presence of a phone: Our technique requires the earbuds user to carry a phone. Phones are ubiquitous and carried almost everywhere. Moreover today's earables are not stand-alone, requiring a companion device they are connected to – usually a phone. While this could also be a laptop or even a smartwatch, in practice, people tend to pair their earbuds to the phone they carry with them. With the advancements in capabilities and resources that earables are experiencing, the hope is, in the future, they will become stand alone. We leave as a future work devising a user-transparent magnetometer calibration for a future stand-alone earbud.

Tilt compensation: Magnetometers usually provide a 2 degrees of freedom orientation parallel to the ground [14]. If the sensor is not leveled, tilt compensations is needed. Common strategies usually exploit gravity measurements from an accelerometer to compute the sensor tilt and map it back in the correct reference frame [14]. While earables experience less significant rotations than other devices by virtue of their attachment to the head, tilt compensation is still required. In our work, although not explicitly stated, we always make sure the x and y we are using are leveled with respect to the xy plane. Further, in our system, head dips of α° can be modeled as rotations about the y axis, resulting in $(mag_x \cos(\alpha), mag_y)$. $\cos(\alpha) \approx 1$ for small values of α, we assume little head dips (like those when normally interacting with a phone) only marginally affect our system.

Scaling factors: Our approach does not estimate the sensor scale factors. Instead, we assume the ratio of scale factors would be 1. This allowed us to reduce the number of distinct headings to estimate the calibration, favoring usability and low complexity. Our justifications for the assumption are:

(i) scale factors are used to address soft iron distortions which, although having a non-negligible effect when looking at the intensity of the Earth's magnetic field, are substantially less significant than hard iron distortion when looking at the heading;

(ii) we expect earbud manufacturers to perform a factory calibration of their sensors, which would build in compensation for any soft or hard iron biases internal to the

earbuds themselves. Calibration parameters do of course change with environmental factors, necessitating in-field calibrations. However, the changes are likely perturbations around the factory calibration. Therefore the *ratio* of observed scale factors – on which the heading computation depends – would be expected to be approximately 1. Notably, this was the case experimentally for all of the magnetometers we tested. Perturbations to this ratio have a limited effect on the estimated heading, so our assumption has minimal effect on the error, whilst reducing the complexity of our calibration procedure;

(iii) assuming $R \approx 1$, we do not estimate the true scale factors. While this have minimal effect on the heading accuracy, it does mean we do not obtain a reliable estimate of the overall magnetic field magnitude, implying we cannot use the field magnitude to discard readings when in a magnetic anomaly. Nonetheless, we are able to accurately compute the earbuds heading, key enabler to many applications, starting from IMU calibration. In the future we hope to extend our model to scale factors, too. In the meantime, we note that anomalies may be detected through large rotation discrepancies between the gyroscope and the magnetometer.

Interactions with the phone: As a proof of concept, in this work, we rely on *FaceID* unlocks to ensure the positional relationship phone-earbuds is what we require for our calibration to work properly (Figure 3.8.1). However, unlocks are not the only events leading to this specific positional relationship. By reliably detecting more of such occurrences (even during a single interaction) would be possible to further increase the granularity of our measurements, which might be suitable for some application which need more frequent references.

3.11 Conclusion

The ever increasing number of personal-scale applications relying on accurate head motion tracking has fuelled our research efforts towards tracking head movements from an ear-piece. Specifically, in this chapter, we first evaluated eSense as an earable device to perform in-ear head motion tracking. Our technique combines multiple streams of data, and, despite the absence of a magnetometer in the inertial sensor equipped in eSense, achieves results precise up to a few degrees, also under realistic situations (e.g., with the subjects speaking or chewing). However, we found the accuracy of our estimation decreases

for longer movements. A preliminary study suggests this can be attributed to the lack of a magnetometer to recalibrate the earables' IMUs. Hence, we went one step forward and, for the first time, we investigated how to augment earables with a magnetometer to aid even more personal-scale applications.

Our research shows how embedding magnetometers in earables is a challenging task, as these rely heavily on radio (mostly Bluetooth) communication (RF) and contain magnets for magnetic-driven speakers and docking. At first, this chapter presented a comprehensive study of the magnetic interference impacting the magnetometer when placed in an earable, showing how by RF (music streaming and voice calls) communications account for the better part of it. We find that appropriate calibration of the magnetometer removes the offsets induced by the magnets, the speaker, and the variable interference due to BT. Therefore, we devise and present an automatic, user-transparent adaptive calibration that obviates the need for alternative, expensive, and error-prone manual, or robotics, calibration procedures. Our evaluation shows how our calibration approach performs under different conditions, achieving convincing results with errors below 3° for the majority of the experiments.

Last, regarding the application area, this chapter focused on facilitating accurate personal-scale inertial sensing applications by (i) proposing an head tracking system that leverages the duality of earables to track motion, and (ii) paving the way to magnetometer-enabled earables to allow for IMU calibration. Having shown the goodness of earables as a sensing platform, the following chapter seeks to investigate the potential of another family of commodity sensors present in earables. To this end, Chapter 4 and Chapter 5 study how to leverage in-ear microphones for general motion sensing and identification purposes, respectively.

Chapter 4

In-ear Microphone Sensing: OESense

Do or do not. There is no try.
–Yoda

4.1 Introduction

In Chapter 3 we have explored the potential of kinetic earables in tracking head movements and, for the first time, we investigated how to augment them with a magnetometer to improve the motion tracking performance. On the contrary, in this chapter, as well as in Chapter 5, we investigate motion sensing with a different modality: *in-ear microphones*. Indeed, historically, researchers used inertial measurement units (IMU) to sense motion – accelerometers in particular. Some example applications are human activity recognition [131], eating habits monitoring [132], smoking gesture recognition [133], and gait analysis [134]. However, as we will show in Section 4.2.1, while in-ear accelerometers can detect intense motions (e.g., walking and running) reliably, the signals recorded under light motions (e.g., chewing and tapping gestures) are often compromised whenever head movements are present. Besides accelerometers, more traditional external-facing microphones have also been adopted to detect motion events (e.g., gesture recognition [66]). However, traditional microphone-based methods suffer from low signal-to-noise ratio (SNR) due to the strong attenuation of sound in the air, thereby significantly limiting the sensing range. Further, such methods are also extremely vulnerable to acoustic interference in the environment (as shown in Figure 4.2.2).

To achieve *reliable* detection of both intense and light human motions with earables, we present OESense, a novel acoustic-based in-ear human motion sensing system. OEsense performs robust motion sensing based on two critical design choices. First, to tackle environmental noise, OESense relies on an *inward-facing* microphone to record motion-induced sounds from *inside* the ear canal. As a result, most of the environmental noise is naturally suppressed. Further, acoustic signals are inherently immune to motion artifacts like those caused by head movements. Second, to cope with the poor SNR of traditional acoustic approaches based on external-facing microphones, OESense exploits the physiological phenomenon known as the *occlusion effect* to enable the detection of both intense and light motions in human ear canal. Concretely, as we discussed in Section 2.2.1, when a motion stimulus is applied to the human body, the occlusion effect boosts low-frequency bone-conducted sounds when the ear canal orifice is occluded. Given most human motions are of the order of a few Hertz, the occlusion effect yields a significant SNR gain.

In summary, we proposed the joint use of the occlusion effect and in-ear microphones for general human motion sensing, which is robust to motion/acoustic interference and capable of reliably sensing intense and light motion occurrences. After having prototyped OESense using a Raspberry Pi [135] and a pair of wired earbuds, we developed the sensing pipelines for three typical applications: step counting, activity recognition, and hand-to-face tapping gestures recognition. To ensure the occlusion effect is present at any given time when we are collecting acoustic-motion data, we proposed a software-based fit test to measure the sealing quality of earbuds. With data collected from 31 subjects, we comprehensively and thoroughly evaluated the performance of OESense under various realistic conditions. Our results show that OESense achieves 99.3% step counting recall, 98.3% recognition recall for 5 activities, and 97.0% recall for five tapping gestures on human face, respectively. Finally, we assessed the system performance of OESense. We measured the system power consumption and latency, showing that for gesture recognition OESense consumes $746mW$ power and the response latency is $40.85ms$. Ultimately, this offers an initial indication of the feasibility of the idea which could be further optimized on more energy-efficient hardware platforms.

4.2 Motivation

With this chapter, we aim at developing a general earable sensing system for human motion detection. The system should be able to accurately and reliably detect both intense

and light (e.g., body surface vibrations) human motions. Specifically, we select three applications as the representatives of intense, light, and mixed motion detection tasks.

- **Step counting (intense)**: human walking involves large scale movements of the whole body and can be detected at different body positions (like foot, hip, waist, and head).
- **Human activity recognition (mixed)**: we select five activities including walking, running, being still, chewing, and drinking, which combines both intense body motions and weak surface vibrations.
- **Hand-to-face gesture interaction (light)**: vibrations generated by tapping different parts of the human face propagate to the ear via different paths. The received signals present distinctive patterns, enabling the recognition of different tapping gestures.

Next, we investigate the feasibility and robustness of two commonly used sensors, i.e., accelerometer and microphone, for these three applications.

4.2.1 Preliminary exploration: accelerometers

Accelerometers have been widely adopted for motion sensing applications due to their capability of sensing both intensive and weak vibrations. As we have seen in Chapter 3, accelerometers have been successfully integrated in many contemporary earbuds [9, 136]. However, as the human head has high degrees of freedom to move and rotate, and it does not always do so accordingly with the rest of body, accelerometers on earables are inevitably affected by head movements. To explore the severity of such interference, we embed an accelerometer (MPU6050 [137]) into an earbud and record its readings when a subject is performing eight activities within the scope of the three applications we consider.

Figure 4.2.1 compares the raw acceleration signals for the eight activities, with (lower row) and without (upper row) head movements. We can observe that: (i) overall, the accelerometer can detect most intense (walk and run) and light (chew and taps) activities, but fails to capture extremely weak signals like drink-induced vibrations. (ii) head movements have minor impacts on intense activities (walk and run). The head movement only produces small variations on the signals whereas the overall patterns remain unchanged, i.e., each step is clearly observable. (iii) the movement of the head completely obfuscates the accelerometer readings of light activities, as the magnitude of head movement is substantially

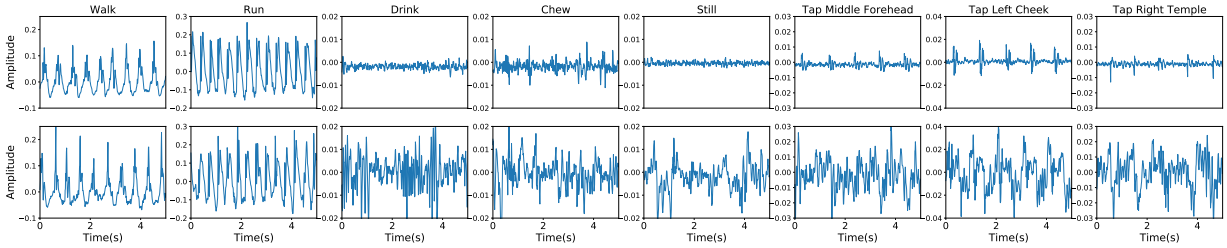


Fig. 4.2.1: Comparison of signals from the accelerometer without (upper row) and with head movements (lower row).

larger. Thus, due to the fact that interference from head movements is unavoidable, we can safely state that accelerometer-based approaches for earables can only reliably apply to intense motion detection.

4.2.2 Preliminary exploration: traditional external microphones

Microphones are the most widely available sensors in earbuds, originally used to capture human speech. Although microphones can measure motion-induced sounds and thereby could be used to infer activities, traditional external microphones are vulnerable to environmental noise. To validate this, we record microphone (external facing) data from an earbud when a subject performs the same activities mentioned above. Figure 4.2.2 compares the raw microphone signals under the eight activities, with (lower row) and without (upper row) background noise (music playing). We can observe that: (i) compared to the accelerometer recordings, the data collected with the external-facing microphone shows less potential for motion detection (only run and two tapping gestures can be reliably detected). The reason is that external microphones measure the air-conducted sounds, which suffer from strong attenuation. Hence, only motions producing relatively high volume can be detected. For walking, the step sound also depends on the material of the ground and shoes. (ii) the plots in the bottom row indicate that the sensing signals are completely hidden in the background music. Given that motion sounds and background music are both audible and share most of the frequency spectrum, it would be very challenging to filter out such interference with signal processing techniques. We will present a more detailed performance comparison between the accelerometer-based approach and the proposed OESense in Section 4.6.1.

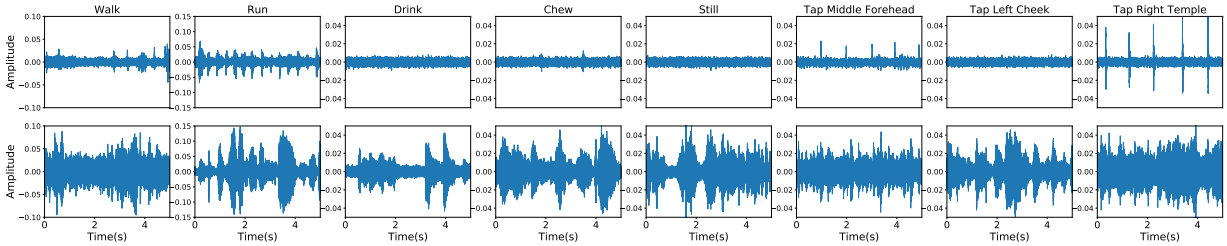


Fig. 4.2.2: Comparison of signals from the external facing microphone without (upper row) and with background noise (lower row).

4.3 OESense: system design

In this section we discuss the working principles leveraged by OESense, we present a preliminary exploration of the feasibility of the system, and we describe in details the sensing pipelines we designed.

4.3.1 Overview

As discussed, accelerometer and external microphone based methods cannot be applied to general motion sensing due to the impact of head movements and background noise, respectively. To achieve our aim, we propose OESense, which makes joint use of the occlusion effect and an in-ear microphone to sense human motions with earbuds. When wearing the OESense earbuds, vibrations/sounds generated by motion stimuli applied to human body will propagate to the ear canal through bone-conduction and be captured by in-ear microphones. With signal processing and machine learning, the signals can be used to infer the applied human motions.

4.3.2 Impact of the occlusion effect

In this section we cover how we envisage to leverage the occlusion effect for human motion sensing. Notably, a more detailed description of the occlusion effect is reported in Section 2.2.1. When vibratory stimuli are applied on the human body, the generated sound will propagate to other parts of the body through bone conduction. Ordinarily, bone-conducted sounds induce vibrations in the ear canal wall, which then *escape* through the opening of the ear canal (ear canal orifice). However, when the ear canal is occluded, the sounds are trapped and reflected back to the eardrum [59] (Section 2.2.1). As a result, the acoustic impedance of the ear canal opening at low frequencies increases [58, 60].

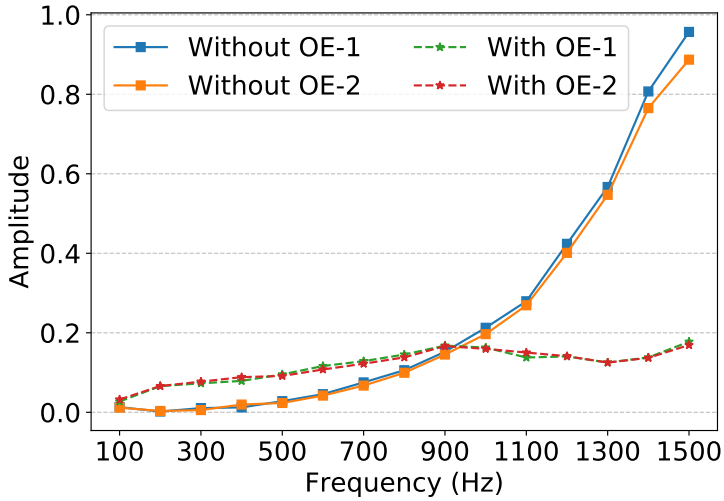


Fig. 4.3.1: Impact of occlusion effect on the frequency response.

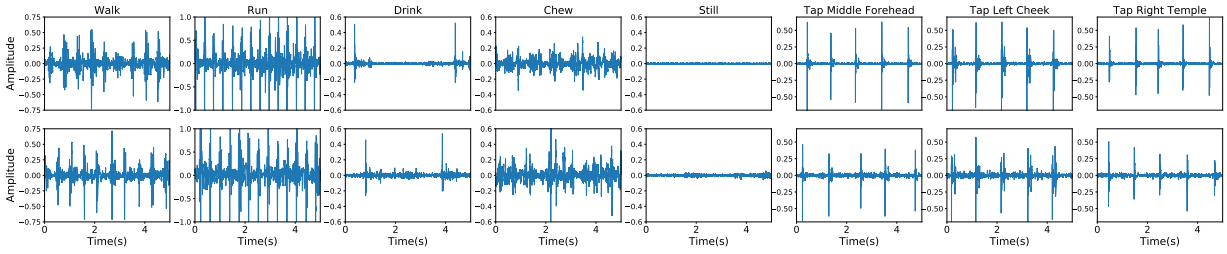


Fig. 4.3.2: Comparison of signals from the inward-facing microphone without (upper row) and with head movements (lower row).

Similarly to what is done by Carillo et al. [57], we also measure the impact of occlusion effect on the ear canal frequency response. This allows us to gauge the extent of the sound boost. We do so by using the earbud speaker to transmit a single tone between $100 - 1500Hz$ (with a $100Hz$ step) and record the reflected sound with an inward-facing microphone. Figure 4.3.1 compares the frequency response with and without the occlusion effect (i.e. the complete blocking of the ear canal opening). We can immediately appreciate how the blocked ear canal produces a stronger response at frequencies below $900Hz$, whilst the open ear canal gains much higher response at higher frequencies. We repeated the measurements twice (removing the earbud and wearing it again) and the response is highly consistent.

Leveraging the occlusion effect for human-related sensing promises three advantages.

(i) First, due to the occlusion of ear canal orifice, the inward-facing microphone mainly captures the bone-conducted sound in the ear canal and is less susceptible to envi-

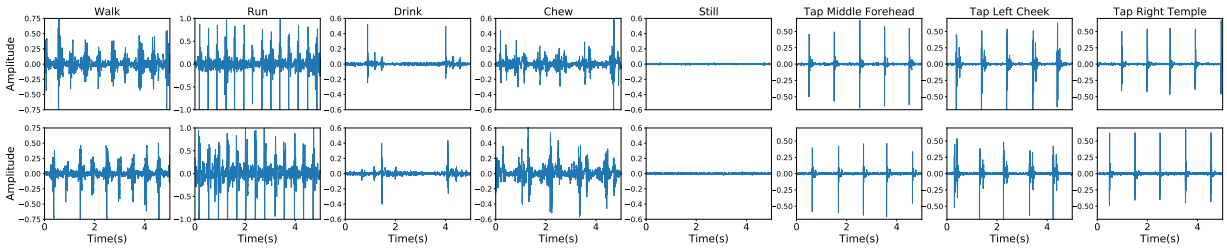


Fig. 4.3.3: Comparison of signals from the inward-facing microphone without (upper row) and with background noise (lower row).

ronmental noises including traffic sounds and human speech.

(ii) Second, given that most human-produced motions are in relatively low frequencies (a few Hertz), the amplification gain provided by the occlusion effect can improve the SNR of the sensing signal.

(iii) Third, although earbuds are mainly used for delivery of sounds (e.g., music or phone calls) to the human ear, these sounds are usually in higher frequencies so sound delivery and human sensing (under 50Hz) can coexist without mutual interference.

4.3.3 Initial exploration

To demonstrate the feasibility of OESense for general motion sensing, we repeat the experiments conducted in Section 4.2.1 and Section 4.2.2 with the inward-facing microphone. Figure 4.3.2 illustrates the recordings for the eight activities, with and without head movements. We observe that the signals for the same activity show similar patterns under both cases, indicating head movements have a negligible impact on the sensing signals. Notably, the slightly higher noise levels under head movements are due to the friction of the earbuds wires, which could be eliminated when a wireless earbud is used. Besides, as shown in Figure 4.3.3, background noise has no impact on the sensing signals as it is naturally suppressed with the occlusion of the human ear. Moreover, since the target motion signals are in frequencies below $50Hz$, any audible background noise can be easily removed with a low-pass filter. A comprehensive evaluation of the sensing performance will be presented in Section 4.6.

4.3.4 Sensing Pipelines

After demonstrating feasibility, we develop a light-weight sensing pipeline for each application. Specifically, we propose a robust step counting algorithm based on envelop extraction and peak detection. For activity and gesture recognition, we propose an audio-based feature set and adopt machine learning based classification.

Step Counting

Human steps create a periodic sinusoidal pattern on the IMU signal, so traditional step counting algorithms aim to calculate the frequency of the sinusoidal signal or match with a sinusoidal template [10]. However, as shown in Figure 4.3.4, the step signals recorded with microphone exhibit a completely different pattern. Specifically, looking at the audio recording, each step is composed of a small chunk of spikes (corresponding to the foot strike stage) and a relatively silent period (corresponding to the foot swing stage). Based on this observation, we propose the following step counting algorithm for in-ear audio recorded signals.

Given the vibrations generated by foot strikes are at a low frequency, we first apply a low-pass filter with cut-off frequency at $50Hz$ on the collected audio signal to eliminate environmental noise and human speech. Then, we feed the filtered signal $f(t)$ to the proposed algorithm. As described in Algorithm 1, we first apply a Hilbert transform on $f(t)$. This outputs the upper envelop (up_evlp) and lower envelop (low_evlp) of $f(t)$, as shown in Figure 4.3.4. We then apply another low-pass filter (cut-off $5Hz$) on the two envelops to smooth them. Afterward, we run peak detection on the smoothed envelops. This outputs the time index ($peak.x$) and amplitude ($peak.y$) for each peak. To avoid over-counting (i.e., false positives), we further propose two strategies to filter the detected peaks: (1) the minimum peak interval (θ_intvl) between adjacent peaks is set to $0.3s$ as the normal human walking frequency is lower than $3.3Hz$; (2) the minimum peak amplitude ($\theta\_amplitude$) is set to $0.3\times$ the average amplitude of all detected peaks. Any peak that fails to satisfy either one of the conditions will be culled. Lastly, to combat the sporadic noise that induces either only an upper peak or a lower peak, we count a step only when a pair of upper peak and lower peak is aligned, i.e., the time lag (refers to maximum alignment interval δ) between them is shorter than $0.2s$.

Algorithm 1: Step counting algorithm.

Input: Low-pass filtered signal $f(t)$, t = 1,2,...,T; minimum peak interval θ_intvl; maximum alignment interval δ

Output: Step counts N

$N \leftarrow 0$ `/* initialize N */`

`/* obtain upper and lower envelope */`

$up\_evlp(t), low\_evlp(t) \leftarrow$ HilbertTransform($f(t)$)

`/* smooth envelop with low-pass filter */`

$up\_evlp'(t) \leftarrow$ LowpassFilter ($up\_evlp(t)$)

`/* peak detection, peak={(x,y); x:time index, y:amplitude} */`

$peak\_up \leftarrow$ PeakDetection($up\_evlp'(t)$)

`/* filter peak_up and peak_low */`

$\theta\_amplitude \leftarrow 0.3 * average(peak\_up.y)$

for $i = 0$; $i < len(peak\_up) - 1$; $i = i + 1$ **do**

  if *peak_up(i+1).x - peak_up(i).x* $< \theta\_intvl$ **then**

    delete $peak\_up(i + 1)$

  if *peak_upper(i).y* $< \theta\_amplitude$ **then**

    delete $peak\_up(i)$

`/* Repeat lines 3-10 for low_evlp(t) */`

$peak\_low \leftarrow low\_evlp(t)$

`/* Align peak_up and peak_low */`

for *upper in peak_up* **do**

  for *lower in peak_low* **do**

    if $|upper.x - lower.x| < \delta$ **then**

      $N \leftarrow N + 1$

return N

Human activity recognition

To perform human activity recognition (HAR) on the in-ear recorded audio traces, we first apply a low-pass filter with cut-off frequency of $50Hz$ on the original signal to eliminate environmental and human sounds. We then proceed by dividing the recorded audio stream into small segments using a sliding window technique. The window size is fixed at $1s$ with

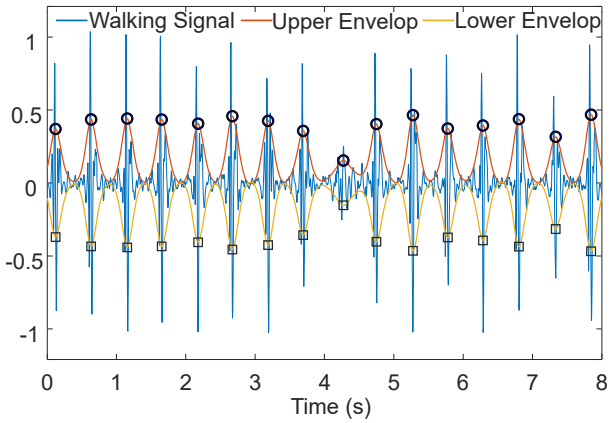
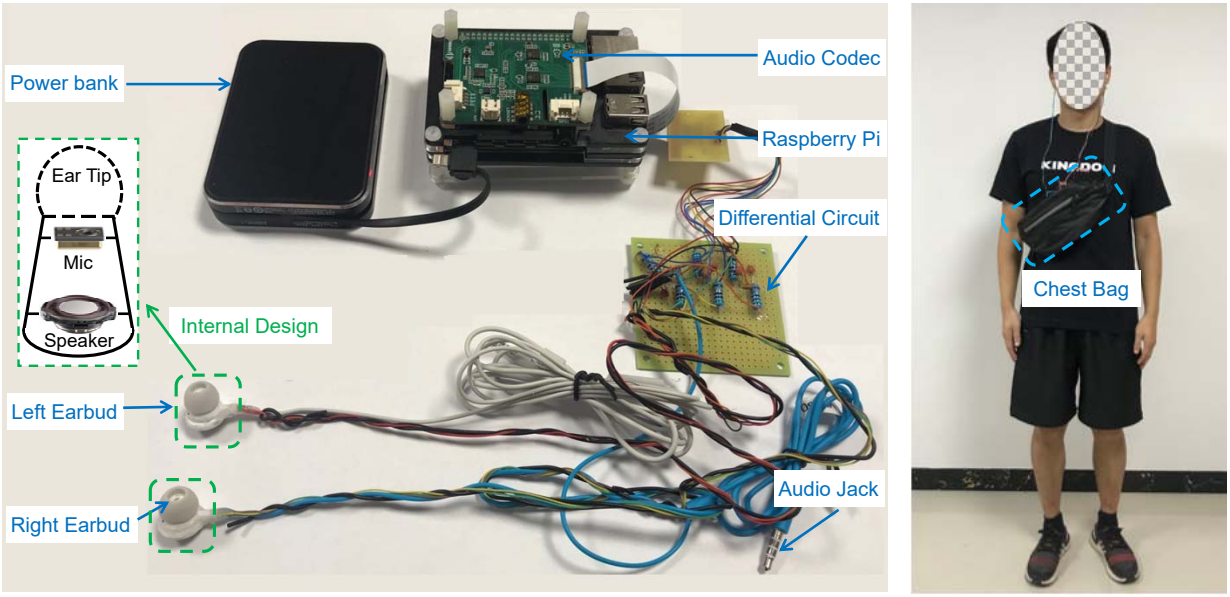


Fig. 4.3.4: A walking segment showing the performance of proposed step counting algorithm.

an overlapping ratio of 50%. Afterward, we leverage *librosa* [138], a widely-adopted python package, for audio features extraction. Specifically, inspired by [139], for each instance (samples in a window), we extract 187 features that cover frequency-based, structural, statistical, and temporal attributes. These features are Mel-frequency cepstral coefficients (MFCC) (40 features), first-order derivative of MFCC (40), second-order derivative of MFCC (40), mel spectrogram (40), chroma of short-time Fourier transform (STFT) (12), contrast of STFT (7), tonnetz (6), RMSE (Root Mean Square Error) (1), and onsets (1). A more detailed description of these features can be found in [139].

Finally, we perform classification with five typical machine learning classifiers: Logistic Regression (LR), Support Vector Machine (SVM), K Nearest Neighbours (KNN), Decision Tree (DT), and Random Forest (RF). We report the results for LR and SVM only as they achieve the best performance. We run the experiment for data from the left earbud and right earbud individually. Notably, we also create a *fused* signal by concatenating the features extracted from the two earbuds. The reason to concatenate features instead of raw signals is to retain the authentic sequential information in the data. Specifically, since the left and right earbud records concurrently, concatenating raw signals (and then performing feature extraction) introduces additional sequential information, while concatenating features avoid this issue as there is no sequential information among features. The code makes use of the Python scikit-learn package [140].



(a) (b)

Fig. 4.4.1: The developed data recording prototype (4.4.1a), and a participant wearing the device (4.4.1b).

Hand-to-face gesture recognition

Similarly to what done for HAR, in order to perform hand-to-face gesture recognition we first apply a low-pass filter (cut-off $50Hz$) for denoising. We then apply the same envelope-based peak detector described in Algorithm 1 for gesture extraction. In detail, we obtain the envelope of each data trace and detect its peaks, where each peak corresponds to a tapping gesture. Then, the gesture start and end is derived by shifting $0.15s$ backward and $0.25s$ forward from the peak, respectively. Consequently, each gesture is composed of all samples within the $0.4s$. To extract the features and run the classification, we followed the same methodology described in Section 4.3.4.

4.4 Implementation

In this section we describe the OESense prototype as well as the fit test we designed to guarantee the presence of the occlusion effect.

4.4.1 Hardware prototyping

First, we present the hardware design and the prototyping of OESense. Although inward-facing microphones have already been integrated into existing commercially available wireless earbuds [136, 141], they are designed for noise cancellation and developers do not have access the raw audio data. Thus, to prove our concept, we prototype OESense by adding an inward-facing microphone to pair of commercial (wired) earphones. We choose the MINISO Marvel earphones [142] as the base earbuds based on two criteria: (i) their internal body is large enough to accommodate a MEMS (Micro-Electro-Mechanical Systems) microphone as well as the original speaker; and (ii) they are equipped with silicone ear tips (removable and interchangeable) that can serve as the occlusion device, providing good sealing quality. To measure sounds from the ear canal, we opted for an analog MEMS microphone (SPU1410LR5H-QB [143]) due to its wide and flat frequency response between $20Hz$ and $20kHz$. As shown in Figure 4.4.1 (within the green dashed box), we embedded the microphone at the front-end of the earbud and moved the original speaker towards the back-end. Such design optimizes the SNR of the microphone, although modifies the internal structure of the earbud. We assess whether this would affect the audio quality of music playback through user perceptions in Section 4.7.

The earbuds have been modified to collect audio signals from both ears. To minimize noise, each microphone is connected to a differential circuit before being sampled by an audio codec (Figure 4.4.1). We choose a ReSpeaker Voice Accessory HAT [144] as the audio codec. The ReSpeaker is controlled by a Python program running on a Raspberry Pi 4B. We sample the microphone data at $48kHz$. The speaker is connected through the original wires using a $3.5mm$ audio jack. The whole prototype is powered with a power bank to facilitate collecting data remotely whenever the participants are moving. To avoid affecting the subjects' walking style, all the components are enclosed in a comfortable chest bag, as shown in Figure 4.4.1b.

4.4.2 Earbud fit test

As we have discussed, the key working principle of OESense is the presence of the occlusion effect. Ultimately, that requires that the ear canal orifice is completely blocked (i.e., occluded), sealing the ear canal cavity. However, through empirical assessment, we observed the earbud might be loosely attached during the experiments, especially whenever the subjects are moving, thus hindering the natural sound boost provided by occlusion effect. To

this end, we propose a *fit test* to check whether the earbuds are properly sealing the ear canal. We do so by ensuring the behavior induced by the occlusion effect is present at any given time (Figure 4.3.1).

The test works as follows: (i) before the subject wears the earbuds in their ears, we first use the speaker to transmit two single tones (with duration of $100ms$) at $300Hz$ and $1500Hz$, respectively. We proceed by measuring the amplitude of the microphone measurements (A_300_{base} and A_1500_{base}). (ii) Once the subject wears the earbuds, we repeat the single tone transmission and calculate the microphone amplitude (A_300_{test} and A_1500_{test}). (iii) Finally, if the amplitude ratio $A\_300_{test}/A\_300_{base} > 5$ and $A\_1500_{test}/A\_500_{base} < 0.2$, the sealing quality is good as the occlusion effect is observed. Otherwise, the user is requested to adjust the positioning of the earbuds and perform the fit test again.

4.5 Data collection

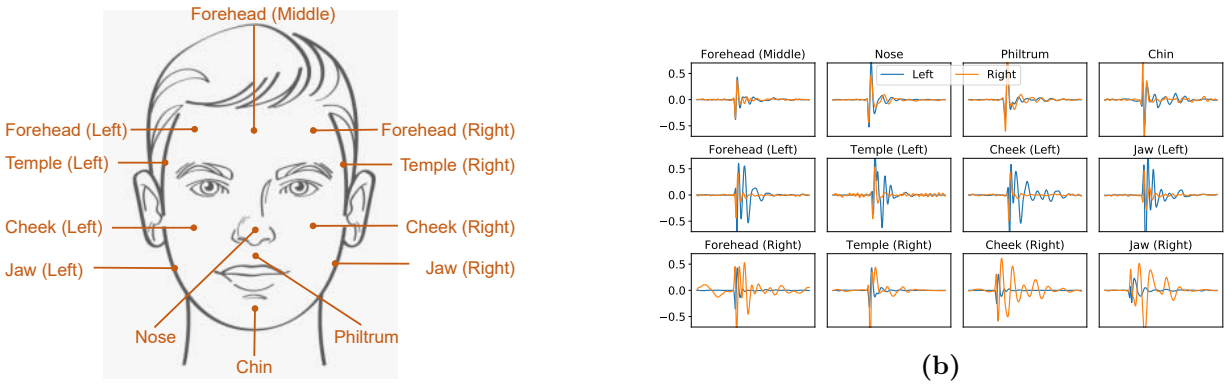
To experimentally compare the performance of OESense with accelerometer-based and external microphone-based approaches, we first asked one subject to collect data for each application under different conditions (with and without motion and acoustic interference). The microphone and accelerometer data were sampled simultaneously at $48kHz$ and $100Hz$, respectively.

We then proceeded by recruiting 31 participants (including 16 males and 15 females, with an age of 26.6 ± 5.8) for larger-scale data collection campaign. During this second data collection, we only recorded inward-facing microphone data [1]. In the reminder of this section we detail the procedures followed to collect the data needed to validate OESense for step counting, activity recognition, and hand-to-face gesture recognition.

4.5.1 Step counting

Participants were asked to walk in a quiet meeting room (around $30dB$ noise level) at their normal walking style and speed. The room size was $12 \times 6sqm$, and the participants were instructed to walk in circles along the walls. To explore the robustness of step counting under various practical conditions, we consider 2 different ground materials (i.e., brick and carpet), and 5 different walking scenarios (i.e., barefoot walking, walking with slippers,

[1] Ethical approval for carrying out all the studies has been granted by the corresponding institution.



(a) (b)

Fig. 4.5.1: Illustration of the designed tapping gestures (4.5.1a), gesture waveform from both earbuds of Subject 1 (4.5.1b).

walking with sneakers, walking while speaking, and walking while chewing a gum). For every ground material and scenario, each participant walked continuously for 1.5min, during which they manually counted the number of steps (serving as the ground truth). Notably, at normal walking speed, all the subjects walked between 156-176 steps within 1.5min. In summary, each participant performed 10 sessions and all the subjects walked 52047 steps in total.

4.5.2 Human activity recognition

The activity experiments were conducted in the same environment we described previously for step counting. Being still, drinking, and chewing gum were performed while the subjects sat on a chair. For drinking, participants held a bottle of water and kept swallowing as much as they could. For walking and running, participants were asked to move in the same room along the walls. During each session, the participants wore the prototyped earbuds (as per Figure 4.4.1b) and performed each of the activities continuously for 1.5min. Notably, every activity was repeated twice (2 sessions), between which a 30s break was set to avoid user fatigue. In total, we recorded $31 \times 5 \times (2 \times 1.5) = 465$ minutes of activity data.

4.5.3 Hand-to-face gesture recognition

As depicted in Figure 4.5.1a, we selected 12 different positions on the face: left forehead, middle forehead, right forehead, left temple, right temple, left cheek, right cheek, left jaw, right jaw, nose, philtrum, and chin. These are the locations on the human face we choose as

the interaction spots for the tapping gestures. Accordingly, we crafted twelve hand-to-face gestures by finger tapping (one-time) on each of the chosen position. Given the unique structure of the bones and the composition of tissues, the vibration paths between the ear canal cavity and the tapping spots are very distinct, ultimately acting as the foundation to recognize different gestures. Figure 4.5.1b illustrates the waveforms of the 12 gestures collected from left and right earbuds of Subject 1. Notably, we can observe how the same finger taps on different spots indeed result in distinctive patterns. Further, the two earbuds can be regarded as independent sensing channels.

The participants were instructed to perform each gesture for 60 times with a tapping interval of $1s$. To assist the participants in maintaining the tapping interval, a cyclic one-second countdown timer was displayed on a laptop screen in front of them. All the gestures were performed with the right hand for fair comparison, as it was the dominant hand for all the participants. The data for Subject 2 and Subject 6 were omitted as they were corrupted [2]. In total, we collected $29 \times 12 \times 60 = 20,880$ gestures.

4.6 Evaluation

In this section, we first compare the performance of OESense against that of the accelerometer and microphone-based approaches (with and without interference). Then, using the data collected during our large-scale data collection, we assess the performance of OESense under various conditions for each of the three applications considered. Finally, we discuss the impact of music playback as well as the power consumption and latency of OESense.

4.6.1 Baselines benchmarking

As shown in Section 4.2, conventional approaches, i.e. accelerometer (Acc) and external microphone (eMic) either fail to detect light motions or suffer from motion and/or acoustic interference. With data collected from one subject (single ear), we run the developed sensing pipelines for each application to compare the final sensing accuracy (recall), as presented in Table 4.6.1. For accelerometer data, we extract around 130 statistical and spectral features using the TSFEL Python library [145] and use a logistic regression classifier.

[2]The ear tips were loosened but subjects did not report, so the occlusion effect disappeared and no gestures was detected.

Table 4.6.1: Performance comparison of OESense with accelerometer (Acc) and external microphone (eMic) based methods. SC-Step Counting, AR-Activity Recognition, GR-Gesture Recognition. The values reported for SC are the number of steps counted (ground truth is 300), whereas for AR and GR are the recognition recall.

		SC	AR	GR
Acc	w/o head movements	300	72.76%	59.75%
	with head movements	299	53.01%	29.55%
eMic	w/o music	109	67.68%	46.53%
	with music	208	27.59%	72.76%
OESense	w/o head movements	300	91.26%	83.09%
	with head movements	299	88.74%	79.62%
	w/o music	300	91.26%	83.09%
	with music	300	90.99%	81.15%

For step counting, the subject walks 300 steps in each session. We can see that accelerometer precisely counts the steps even under more challenging situations with interference. This is because acceleration produced by the head movements is negligible if compared to that of walking and running. Conversely, the external microphone severely under counts the steps in both cases, as the air-conducted step sounds are very weak. OESense achieves great performance in all cases due to its natural immunity to motion and background noise.

For activity recognition and gesture recognition, we can observe that (1) without interference, Acc and eMic have lower recall than OEsense as they cannot detect some of the activities/gestures. (2) With interference, the recognition recall of Acc and eMic decreases significantly. Overall, OESense achieves comparable, if not superior, performance in all cases. The relatively big accuracy drop with head movements is actually caused by the fraction of earbud wires and could be resolved when a wireless earbud is developed.

4.6.2 Step counting

Figure 4.6.1 depicts the step counting performance of OESense under various conditions. Precision and recall respectively reflect the over-counting and under-counting behavior of OESense. Overall, we can observe that for step counting, both the precision and the recall of the system are higher than 97.5% regardless of the ground material and the walking

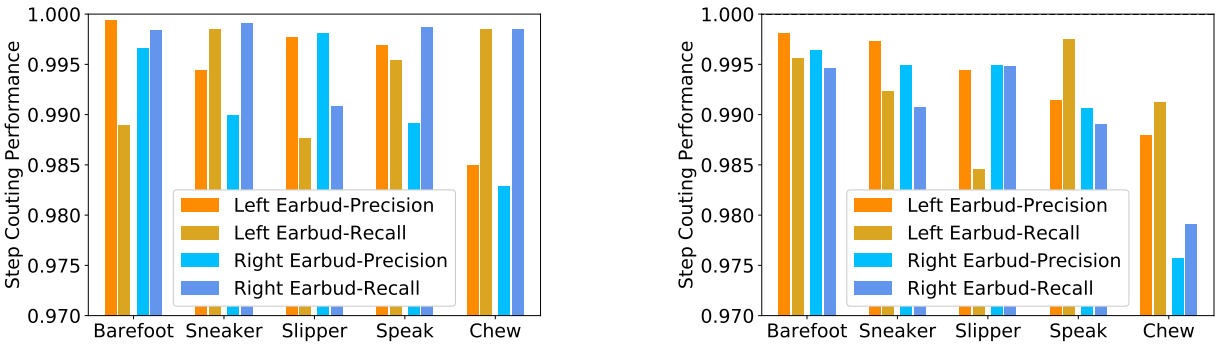


Fig. 4.6.1: Step counting performance for walking on brick (4.6.1a) and on carpet (4.6.1b).

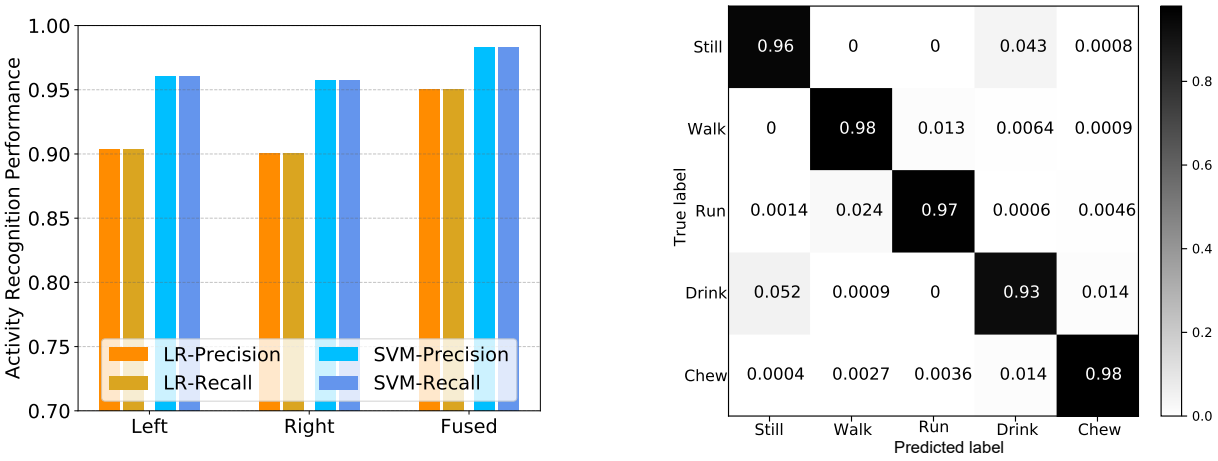
scenario, demonstrating the superior performance of OESense on the step counting task.

Notably, different walking scenarios have distinct impacts on step counting performance. For instance, when walking with slippers, the foot contact area whenever the feet hit the ground increases, thereby weakening the strength of the generated vibrations [146]. Ultimately, this leads to the algorithm missing some steps (false negatives), resulting in a lower step counting recall. Conversely, when people are chewing, the vibrations generated by the jaw movements propagate to the ear canal through bone conduction. Consequently, we can expect to observe more spikes on the audio signal, thus leading to over-counting. Hence the precision under the chewing scenario results to be the lowest.

For different ground materials, we can observe that brick surfaces generally yield higher recall compared to carpet. This is because soft carpet will dampen part of the vibrations, thus resulting in missing steps. In addition, the audio signal collected when walking on a carpet has lower amplitude (i.e., lower SNR). As a consequence, the signal is more easily polluted to other body movements. Remarkably, both the left and the right earbuds, when considered individually, achieve very high accuracy, indicating the proposed step counting approach can work well even with a single earbud. Overall, OESense achieves an average step counting recall of 99.32% and a precision of 99.26%, which dramatically outperforms the industrial standard for pedometers (for example, ±3% counting error set by the Japanese Ministry of Economy Trade and Industry [147]).

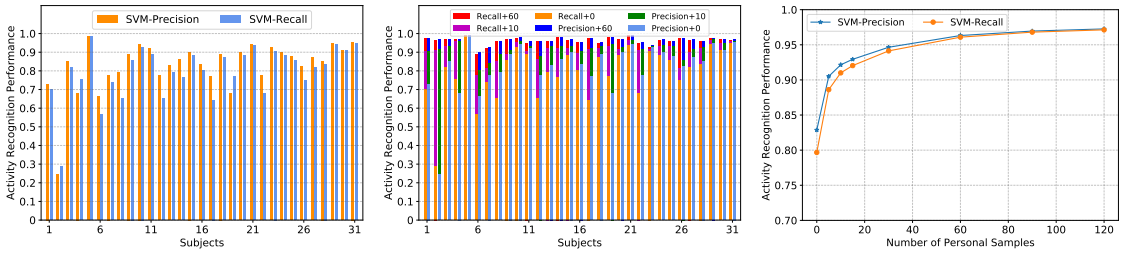
4.6.3 Human activity recognition

Overall performance: After aggregating data from all the 31 subjects, we compare the



(a) (b)

Fig. 4.6.2: Overall activity recognition performance (4.6.2a) and confusion matrix (4.6.2b).



(a) (b) (c)

Fig. 4.6.3: Activity recognition performance for leave-one-out test (4.6.3a), individual performance with model personalization (4.6.3b), and average recognition performance with different amounts of personal data (4.6.3c).

performance of three datasets: left earbud, right, and fused. For each task, we split the data into training set (80%) and testing set (20%). We then proceed by training the models with 5-folds cross-validation. The results are presented in Figure 4.6.2a. We can observe that: (i) SVM always achieves better performance than LR; (ii) the results of the classification run on the left and right earbud (individually) are similar; (iii) the fused dataset yields the highest recognition precision and recall both of around 98.3%. Such improvement might arise from the fact that fused dataset benefits from two sensing channels and, therefore, it is more resilient to the signal distortions taking place when one of the ear tips is loose. For these reasons, in the reminder of this section, we only present the results for SVM and with fused dataset.

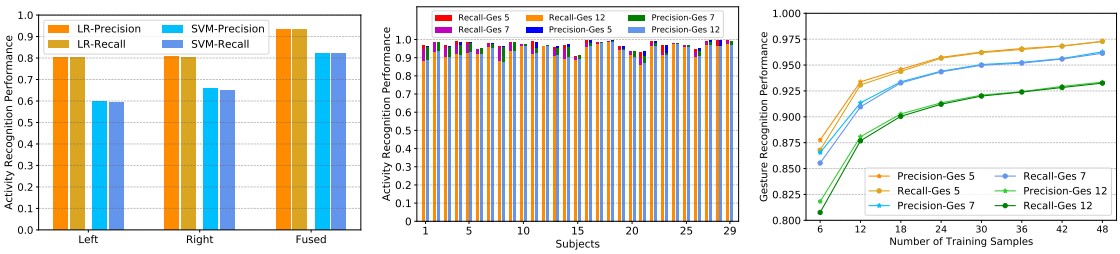
Figure 4.6.2b reports the confusion matrix for the five activities considered using SVM.

We can appreciate how the trained model recognizes walking, running, and chewing reasonably well; whilst the main accuracy loss comes from the baseline still and the drinking activity being confused. This stems from the fact that when collecting drinking data the participants were unable to continuously swallow water for prolonged periods. As a result, the traces are characterized by resting periods between swallowing episodes. Notably, if the resting period is longer than $1s$ (length of the window size), the recorded segment would be similar to that of someone sitting still as no action is performed, resulting in a confusion between drinking and being still.

Leave-one-out test: To justify how a pre-trained model can be generalized to a new user, we perform the leave-one-out (LOO) test on the fused dataset. Practically, we iteratively select one subject for testing and we train the SVM classifier using the data from the other subjects (30). Figure 4.6.3a shows the recognition precision and recall for the leave-one-out test. Notably, the results vary significantly among different subjects. The pre-trained model generalizes well to some subjects, while it appears to suffer from significant performance degradation for other subjects. For example, when testing on Subject 5, the model achieves approximately 98.4% recognition recall, while for Subject 2 the recall is only 24.4%. The reason for this might arise from the fact that people perform the five activities differently. For instance, the walking style (i.e, *gait*) of each person is unique. Besides, people chew in very distinct ways, such as slow/fast chewing and gentle/ravenous chewing. Thus, if the dataset used for pre-training the classifier does not include the style of a new user, the model will perform poorly.

Model personalization: A simple way to address the model generalization issue is by collecting activity data from as many as subjects as possible (e.g., hundreds or thousands). This way, the pre-trained model covers large variations of activity styles. However, this comes with a tremendous burden related to the data collection. In this dissertation, we explore an alternative technique that leverages a user-specific model (*model personalization*), by re-training the general model with personal data. Notably, the results show that our system only requires a limited amount of user data to benefit from the personalization performance boost. For each iteration in the leave-one-out tests, we include a different amount of data from the testing subject for training and testing on the rest.

Figure 4.6.3b shows the recognition performance when 0, 10, and 60 samples from each subject are used to re-train the model. We can observe that with 10 samples both the recognition precision and recall are significantly improved. This is particularly true espe-



(a) **(b)** **(c)**

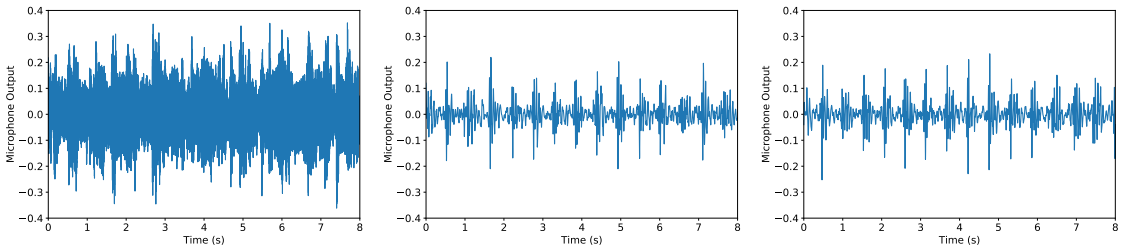
Fig. 4.6.4: Overall gesture recognition performance on 12 gestures (4.6.4a), individual gesture recognition performance with different gestures (4.6.4b), and impact of training data size averaged over 29 subjects (4.6.4c).

cially for those subjects for whom the performance were lower to begin with (e.g., Subject 2 improves from 24.4% to 91.6%). As expected, with 60 samples, the performance can be even further enhanced.

Figure 4.6.3c compares the average value with [5, 10, 15, 30, 60, 90, 120] personal samples added for model re-training. The results indicate that higher accuracy improvement can be achieved when more personal data is provided. The average precision reaches up to 92.1% (increasing by 9.2%) with 10 personal samples and 96.3% (benefiting from a 13.4% enhancement) when 60 personal samples are added. Targeting 90% precision, the users only need to provide 10 samples for each activity, which can be be easily collected within $25s$ ($5s$ each) with a 50% window overlapping. The need for user intervention (providing personal data) is socially accepted for different applications, such as user authentication (profile registration using face or fingerprint) and IMU-based motion tracking (user-assisted sensor calibration with magnetometers [40]).

4.6.4 Hand-to-face gesture recognition

Overall performance: Figure 4.6.4a compares the gesture recognition performance (averaged across 29 subjects) among the two classifiers for the three datasets. We can see that LR consistently achieves better results, which might be due to the fact that features from different gestures are likely to be separated linearly. As expected, the fused dataset (93.2% recall) outperforms the two individual datasets (80.1% and 80.5% recall for left and right, respectively) on the 12 gestures, demonstrating the benefits of sensing with both earbuds. We then limit the classifier to LR and, as before, we only consider the fused dataset in the reminder of the evaluation.



(a) **(b)** **(c)**

Fig. 4.6.5: The original signal (4.6.5a), the low-pass filtered signal for participant walking during music playback (4.6.5b), the low-pass filtered signal for the same participant walking without music playback (4.6.5c).

Gesture set optimization: Next, we perform gesture set optimization aiming to (i) further improve the recognition accuracy and (2) reduce the overhead for users to memorize gestures. Informed by the confusion matrix obtained when the users were performing 12 gestures, we limit the number of hand-to-face gestures to 7 and to 5 by removing those gestures with large confusion errors and close spatial proximity. The selected 7 gestures are {*Forehead Middle, Nose, Chin, Jaw Left, Jaw Right, Cheek Left, Cheek Right*}. Conversely, the optimal 5 gestures are {*Forehead Middle, Nose, Chin, Jaw Left, Jaw Right*}. We then run the experiment with LR classifier on the two new gesture sets and compare the results for each individual in Figure 4.6.4b. Although the recognition performance varies among subjects, most of them achieve $> 90\%$ precision and recall with 12 gestures. Notably, the performance improves even further optimizing the gesture set. Overall, the average recall for 12, 7, and 5 gestures is 93.2%, 96.0%, and 97.0%, respectively.

Impact of training size: The proposed hand-to-face gesture interaction is founded on the fact that vibrations from different tapping spots experience distinct paths to the ears: the model is actually trained to recognize these paths. Given that people have different head size and bone structure, it is expected that the model trained on one subject cannot be fitted to others. To confirm this, we run the leave-one-out test and obtain an average recognition recall of 20%. As a consequence of this finding, it is clear that users should train a personalized model with their own gesture data. To investigate the user burden for training data acquisition, we re-train the model with different number of gesture samples. As shown in Figure 4.6.4c, the accuracies of the three gesture sets have comparable behaviors increasing the number of training samples. With 12 samples from each gesture, when the model is trained and tested on the 5 gestures set, it achieves a recognition recall of 93.5%. Based on the proposed data collection protocol (one gesture per second), only

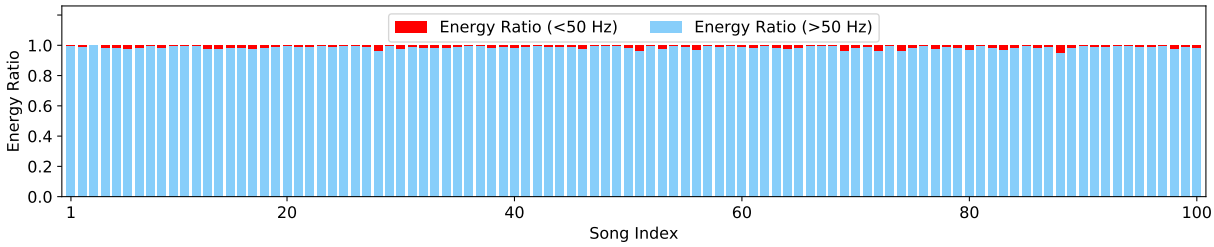


Fig. 4.6.6: Spectrum energy analysis of 100 songs.

$1min$ of training data is required. Practically, providing user data to train the personal model is acceptable in other input/gesture interaction systems [148], as long as the burden on the user remains small.

4.6.5 Impact of music playback

Given that the original functionality of earbuds is to play back sounds (e.g., music and phone calls), a fair concern would be whether these sounds (usually much higher in volume) might pollute the audio sensing signals. We sought to investigate this by asking one participant to walk while the earbuds are playing a song – with an appropriate volume level. Figure 4.6.5a illustrates the original signal collected from the left earbud and, as we can see, it appears to be dominated by the music. Conversely, Figure 4.6.5b shows the low-pass filtered ($< 50Hz$) version of the signal, where the steps are clearly noticeable. We further instructed the subject to walk without music playing, in the same condition. The low-pass filtered signal is plotted in Figure 4.6.5c. Visually, we can immediately see that the two filtered versions have high similarity and the step counts can be easily derived. We also quantify the similarity of signals from frequency domain using the structural similarity (SSIM, a well-known metric to compare similarity between two images [149]). Specifically, we first obtain the spectrogram of each signal using short-time Fourier transform (STFT), and then calculate the SSIM index between two spectrograms (images). The SSIM index ranges from 0 to 1. The higher the value, the greater the similarity between the images. Our results show that the SSIM index for {Figure 4.6.5a, Figure 4.6.5b}, {Figure 4.6.5a, Figure 4.6.5c}, and {Figure 4.6.5b, Figure 4.6.5c} is 0.35, 0.33, and 0.95, respectively, suggesting that music playback has an extremely limited impact.

To further confirm that OESense is robust against different types music, we analyze the spectrum of the *All-time Top 100 Songs* launched by Billboard [150]. For each song, we perform the fast Fourier transform (FFT) analysis and obtain the signal energy at

Table 4.6.2: Power consumption of OESense for gesture recognition.

Operation	Power (mW)
RasPi(Baseline)	2,340
RasPi+MicRecd	2,459
RasPi+MicRecd+GesRecg	3,086(LR), 3,106(SVM)

each frequency band. Then, we sum up the energy with frequencies below and above 50 Hz and calculate the energy ratios of the two frequency ranges, respectively. As shown in Figure 4.6.6, the energy of all the songs are dominated by frequencies higher than $50Hz$. The average energy ratio of $< 50Hz$ signal is only 1.5%, indicating that the impact of music playback is indeed negligible. In terms of phone calls, the frequency range of human voice over telephony transmission is within $300-3400Hz$ [151], hence it can be completely removed after applying a low-pass filter.

4.6.6 Power and latency measurement

To the best of our knowledge, there is no open platform to instrument a stand-alone earable sensing system. As a matter of fact, most of today's earables are directly offloading their data either to a phone or via the network. Instead, here, we wanted to explore the system level performance of OESense as a stand-alone system, as we envisage some of the future earables might become stand-alone. We opted for a Raspberry Pi 4B as a reference platform. Taking hand-to-face gesture recognition as an example, we evaluate the power consumption and latency of our system. We train the machine learning models (LR and SVM) on a laptop and implement the gesture recognition pipeline (including low-pass filtering, feature extraction, and inference) on the Raspberry Pi. As shown in Table 4.6.2, compared to baseline (idle) power consumption ($2,340mW$), powering and recording microphone data (MicRecd) consumes an additional $119mW$. This value is dramatically higher than the power draw of the microphone ($360\mu W$) reported in the data-sheet [143], which indicates that most of the power is consumed by the Raspberry Pi for data sampling. When the gesture recognition classifier is running, the Raspberry Pi consumes $746mW$ and $766mW$ overall for LR and SVM, respectively. Although at a first sight these figures may seem substantial, it is worth noticing that: (i) unlike low-power micro-processors (especially those designed for audio processing, like the Apple H1 chip embedded on the AirPods Pro), the Raspberry Pi is known to be power-hungry without energy efficiency optimization; (ii)

Table 4.6.3: Latency of OESense for gesture recognition.

Operation	Latency (ms)
Low-pass Filtering	1.54
Feature Extraction	38.97
Inference	0.34(LR), 0.95(SVM)

recognition (feature extraction and inference) is run only whenever a gesture is detected and such operation time is typically very short, so the actual energy consumption would be much lower. Table 4.6.3 lists the latencey of the various steps of the recognition pipeline. The majority of the time is due to feature extraction ($38.97ms$), while inference time is almost negligible ($0.34ms$ for LR and $0.95ms$ for SVM), granting OESense quasi-real-time performance.

To further ground our study with practical considerations, we estimated the overhead of running gesture recognition over any other earbud functionality. We consider the possible worst case scenario of a user performing a gesture ($0.4s$ long) every $2s$, a very aggressive assumption. The average energy consumption per second would be $E = 119mW \times 1s + 627mW \times 40.85ms \times 0.5 = 131.8mJ$. Considering a wireless earbud with a battery like that of an AirPod Pro ($81mAh$), OESense could operate for a time $T = \frac{81mAh \times 5V}{131.8mJ} = 3.07h$. Although these are ball park figures, on a non-optimized off-the-shelf device, they give an indication of the actual feasibility of OESense in practice. We believe the overhead would be significantly smaller on high-performance audio chips.

4.7 Discussion and limitations

While prototyping OESense, we changed the position of the speaker to embed the microphone. Such design optimizes the SNR of the microphone, yet modifies the internal structure of the earbud. Thus, we assessed whether the audio quality of music playback is affected through a user study. Each participant was instructed to listen to a music segment with unmodified and modified earbuds respectively and rated their perception towards the audio quality. 29/31 subjects reported that no difference was perceived between the two earbuds and 2 subjects even reported that the modified earbud has slightly better audio quality. Hence, we can conclude the add-on sensing capability does not hinder the audio playback quality. However, there are also several limitations.

First, the concept of OESense is applicable to in-ear headphones only due to the requirement of occluding the ear canal opening. Such physical occlusion might lead to impaired awareness of the surrounding environment (e.g., traffic sounds) and incur potential safety issues. A viable solution would be to replicate the transparency mode on AirPods Pro devised by Apple [136]. Practically, the external microphone can measure the outside sounds and replay the meaningful parts (like sirens, between 725-1600 Hz [152]) through the on-board speakers.

Second, we implemented the concept of OESense on a Raspberry Pi based data collection and processing system, which is cumbersome and impractical for mobile scenarios. Besides, the Raspberry Pi consumes substantial power (Section 4.6.6) as it is not optimized for low-power applications. Thus further efforts to implement OESense in an energy-efficient manner are required. In addition, current OESense prototype is built upon a pair of wired earbuds, where the connection wires swing during movements and produce additional noise. This impact is expected to disappear with wireless earbuds.

4.8 Conclusion

In this chapter we presented OESense, a novel human sensing system based on audio signals recorded inside the ear. Leveraging the occlusion effect (described in detail in Section 2.2.1), OESense shows great sensing potential for both intense and light human activities. We demonstrated three sensing applications (i.e., step counting, activity recognition, and hand-to-face gesture interaction) with the developed OESense prototype. All applications achieved good performance (average recall of 99.3%, 98.3%, and 97.0%, respectively). Further, our system analysis suggests that OESense can achieve quasi-real-time performance with acceptable power consumption. Given the ear contains abundant information about human vital signs and motions, OESense has the potential to be extended to other personal-scale sensing applications like heartbeat detection, jaw movement detection, facial expression recognition, and so on. Building up on these considerations, in Chapter 5 we devise EarGate, an acoustic-gait-based identification system which leverages a similar underlying principle to that presented in this chapter.

'Your eyes can deceive you; don't trust them.'

Obi-Wan Kenobi

Chapter 5

In-ear Microphone Sensing: EarGate

5.1 Introduction

In Chapter 4 we investigated the potential of OESense, an earable system equipped with an in-ear microphone to sense human motion and gestures. Having seen how the occlusion effect (Section 2.2.1) facilitates in-ear audio sensing, in this chapter we build up our observations and propose EarGate, a novel earable-based *acoustic-gait* identification system. Similarly to OESense, EarGate, is built around a cheap in-ear facing microphone that is already available on most earbuds and hearing aids (e.g., for noise cancellation purposes).

Human gait has been shown to be unique across individuals and hard to mimic [67, 68]. As such, there have been a variety of attempts to use gait as a biometric for user recognition and identification. While computer vision-based gait bio-metrics are widely spread, wearable-based gait tracking approaches are particularly attractive for continuous identification and, potentially, authentication. Wearable gait-based identification is an enabler for various applications including: keeping mobile devices unlocked and ready for interaction when on the owner's person; identification of the device owner as the policy owner when rewarding healthy habits sensed via the wearable for health insurers; automated entry systems for home, work or vehicles; automated ticket payment/validation for public transport [153, 154]; etc. There have even been a number of instances of using wearable gait data to generate a secure key to pair devices worn on the same body [155].

Concretely, wearable-based gait tracking methodologies leverage sensor data collected from wearable devices worn by the user to capture their motion dynamics. Typically these are

facilitated by accelerometer analysis [72], or step sounds [74, 156]. To date, the focus has been on smartphones or smartwatches as the current leaders in the mass-market wearables. In this dissertation, instead, we look at the use of ear-based sensing (via so-called *earables*) for this task. The importance of the approach we present is further highlighted by the fact that, in the near-future, earables are likely to become stand-alone devices [5,19]. Hence, the need for earable-based identification schemes becomes more and more crucial [27]. Being able to seamlessly identify the earable wearer can act as an authentication accelerator for earables or mobile devices, bypassing traditional bio-metrics such as fingerprint (requires the integration of capacitive sensing pads on earables with limited size) and face recognition (impossible to capture front face image from an earbud). Occasionally, if the identity is mistakenly rejected, a request for a secondary authentication can be triggered. The secondary authentication method could be implemented on a companion device, such as smartphone/smartwatch, that has full access to fingerprint and human face. Additionally, once the user has been successfully identified by the earable, the earable itself can act as a hub to authenticate the user for access control (e.g. opening their office door, validating their ticket at the train station, etc.). Furthermore, with the increased sensing capabilities of earables, successfully identifying the earable's wearer becomes crucial in order to associate the sensitive bio-medical information collected by the earable to the right user. As we have discussed in Chapter 3, today's earables broadly fall into two categories: advanced playback devices (e.g., Apple AirPods) or assistive devices (e.g., advanced hearing aids). Notably, although being useful for all earables, such continuous identification systems suit hearing aids particularly well, as they are continuously worn throughout the day.

Here we investigate acoustic-gait as a convenient alternative to inertial-based gait tracking. Specifically, we look at the possibility of gait-based identification from the *sounds* caused by the physical act of walking and transmitted internally via the musculoskeletal system. Similarly to OESense (presented in Chapter 4), our novel earable-based *acoustic-gait* identification system, EarGate, is built around a cheap in-ear facing microphone. In Chapter 4 we have shown that an earable equipped with a microphone *inside* the ear canal can not only detect motion signals, but that those are also amplified by the combination of two biological phenomena: *bone conduction* and the *occlusion effect*. Building up our findings from Chapter 4, we conducted a second user study, with the same 31 subjects of mixed gender, demonstrating how the acoustic-gait thus collected is a good candidate for a privacy-preserving identification system (benefit from the in-ear facing microphone). Further, we investigated the implications of running the identification-framework entirely

on-device, as well as offloading the computation (either to the Cloud or to a companion device) to show the versatility of the approach.

In summary, *our approach leverages sensors (microphones) already inherently embedded in earables for gait identification.* As mentioned, previous gait identification/recognition approaches have often relied on inertial sensors [72]. However, inertial sensors have made reasonable penetration into the high-end leisure devices market but not as much into cheaper earables, or the hearing aid market: the addition of inertial sensors to these devices would entail more complex system design and form factor [157].As a consequence, inertial-based earable solutions are likely to result in increased cost and delays in reaching the market. This dissertation offers an investigation into an alternative to inertial sensors through the use of microphones. Notably, while we acknowledge the merits of inertial-based gait recognition approaches, there is great value in showing the potential of a lesser explored modality, such as in-ear microphones. Particularly given what is suggested by research and market trends, in the near future, miniaturized form factor will have a key role, especially as the distinction between hearing aids and earables is likely to become less marked. Further, in-ear microphone offer complementary advantages as a sensor to enable noise cancellation, a must-have feature for both high-end leisure earables and in hearing aids. While in this dissertation we focus on acoustic-gait recognition, we note this is only one of the possible use cases for in-ear bone-conducted sounds. For instance, these could indeed be used for activity recognition (as discussed in Chapter 4), as well as physiological sensing [23]. Ultimately, in this chapter we describe a novel earable-based gait identification system, EarGate, consisting of a hardware prototype and a software pipeline. EarGate not only leverages a novel type of signal, in-ear bone-conducted sounds, to identify users based on their gait, but also is accurate, robust, and has very low burden on the user (only a few steps are sufficient to identify the user, without the need for them to be continuously walking). We designed an end-to-end signal processing pipeline an techniques to guarantee the reliable presence of the occlusion effect. We evaluate the identification performance of EarGate under various practical scenarios, showing we can achieve up to 97.26% Balanced Accuracy (BAC) with very low False Acceptance Rate (FAR) and False Rejection Rate (FRR) of 3.23% and 2.25%, respectively. Furthermore, we demonstrated that EarGate is robust to high-frequency internal (human speech) and external (music playback and phone calls) noises. Finally, we assessed the system performance of EarGate by measuring the power consumption and latency. We find EarGate can work in real-time ($74.25ms$ on-device identification latency) consuming acceptable energy ($167.27mJ$ for one-time on-

device identification). This confirms that EarGate could be deployed in new generation earables which will likely be standalone from the system perspective.

5.2 Preliminaries

In this section, we first brief the reader on the rationale driving our work and then provide evidence of the feasibility of our approach. As done when devising OESense (Chapter 4), when conceiving EarGate, we leverage the occlusion effect (Section 2.2.1) to boost the in-ear recorded audio. However, unlike in OESense where we detect and count steps, EarGate aims at identifying users from their gait (i.e., the walking style of a person, Section 2.2.4).

5.2.1 Out-ear microphone vs. in-ear microphone

Compared to traditional approaches leveraging external microphones (out-ear mic) for gait recognition [74, 156], the use of occlusion effect and an inward-facing microphone brings the following advantages: (i) given an occluded ear canal, an inward-facing microphone, which mostly records bone-conducted sounds, is less susceptible to external sounds and, consequently, environmental noise. This not only means our system is more robust to noise but, practically, it makes our approach more appealing from a privacy perspective: potentially sensitive external sounds, such as human speech, are hardly audible from our in-ear facing microphone. (ii) another direct consequence of the occluded ear canal is that the body-sounds we are interested in – which are relatively low-frequency – greatly benefit from the low-frequency amplification boost induced by the occlusion effect, thus resulting in an improved signal-to-noise-ratio (SNR) of the desired signal (the gait of the user).

Figure 5.2.1 plots the raw signals and the corresponding spectrograms collected with the out-ear microphone and in-ear microphone when a subject is walking, with and without environmental noise. From these figures, we can clearly observe how the in-ear microphone overcomes the drawbacks of traditional microphones, which capture air-conducted sounds. Firstly, air-conduction results in large attenuation, while bone-conduction, properly coupled with the occlusion effect, guarantees an excellent SNR. Secondly, the walking sounds measured by the external-facing microphone reside in a higher frequency band (audible) and end up being completely mixed with environmental noise (e.g., music and human speech). Consequently, they can not be separated from each other, even after a low-pass filter.

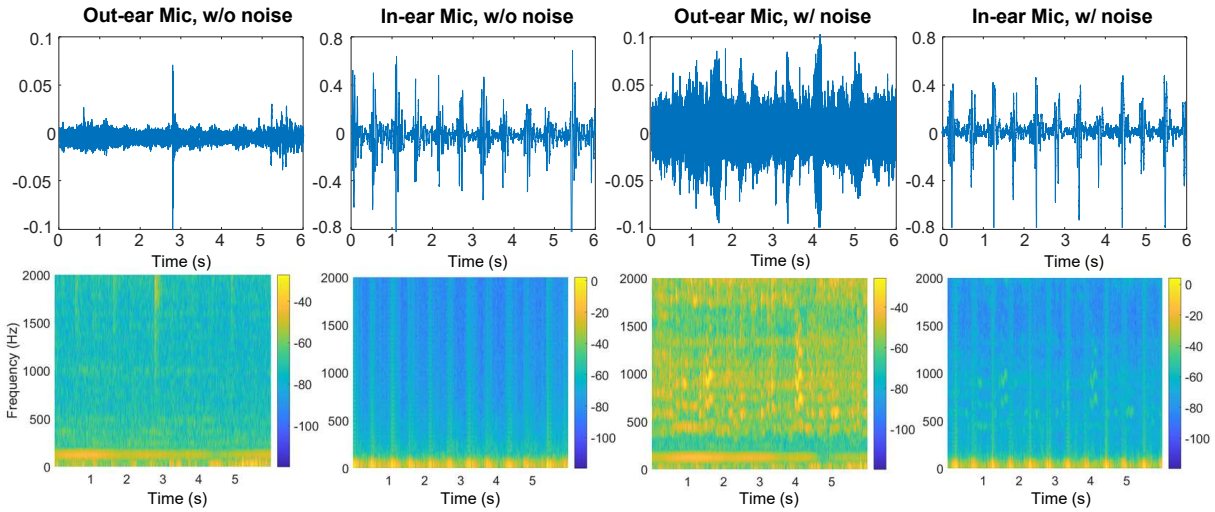
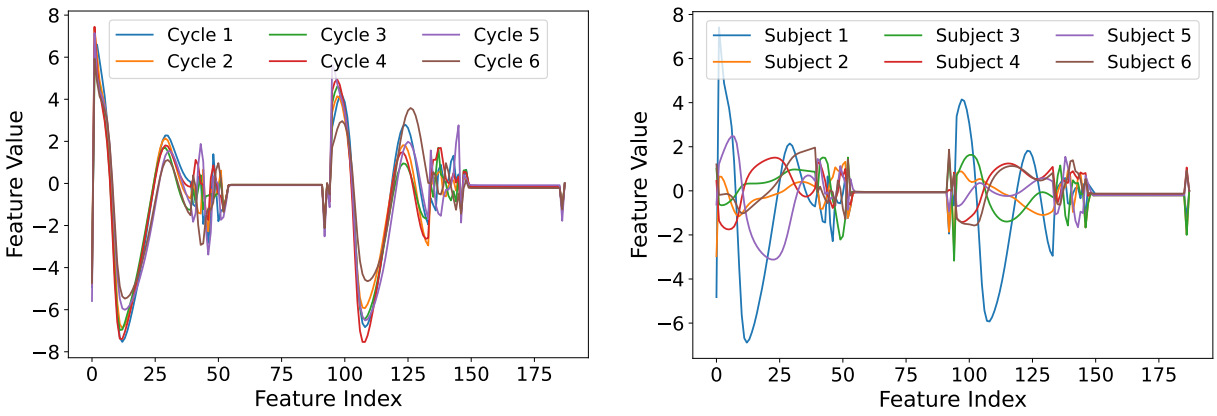


Fig. 5.2.1: Walking signals collected with a traditional outward facing microphone and an in-ear microphone under different conditions. Notably, thanks to the occlusion effect, the in-ear microphone data shows higher gain at low frequencies ($< 50Hz$) and more resilience towards environmental noise.



(a) (b)

Fig. 5.2.2: 5.2.2a Extracted feature vectors for four different gait cycles from the same subject, and 5.2.2b gait cycles from four different subjects. Figure 5.2.2a clearly shows how the gait cycles captured by EarGate are consistent for each individual (i.e. high intra-class similarity). Figure 5.2.2b, on the other hand, shows how the gait cycles are distinguishable among subjects (i.e. high inter-class difference).

5.2.2 Feasibility exploration

As we have seen in Section 2.2.4, the walking style of a person is commonly known as their gait. Medical and physiological studies [67] suggest that the human gait shows 24 different components. The differences between the gait of distinct subjects are caused by the uniqueness in their muscular-skeletal structure. In fact, the human gait is regulated by precise bio-physical rules [68], which in turn, are dictated by the tension generated by the muscle activation and the consequent movement of the joints. As a result, the forces and moments linked to the movement of the joints cause the movement of the skeletal links which, therefore, exert forces on the environment (e.g. the foot striking the ground). Hence, the human gait can be described as a generation of ground-reaction forces which are strongly correlated with the muscular-skeletal structure of each individual. In practice, differences in the body structure of individuals are among the factors that produce the interpersonal differences in walking patterns that enable gait-based identification.

Although the in-ear microphone is capable of detecting human steps, whether it is possible to leverage acoustic gait to differentiate people remains unclear. To demonstrate the feasibility of acoustic-based gait identification, we need to prove that (1) gait cycles belonging to the same individual are consistent with each other (i.e., intra-class similarity); and (2) gait cycles belonging to different subjects show significantly different patterns (i.e., inter-class dissimilarity). Thus, we collected data from four subjects and we proceeded extracting a number of features (see Section 5.3.3) to represent the user's gait. As shown in Figure 5.2.2, the extracted features exhibit high intra-class similarity and high inter-class difference. Therefore, we can conclude it is feasible to identify people leveraging the acoustic signals measured with the in-ear microphone.

5.3 System design

This section presents an overview of EarGate, illustrating its functionalities and shaping the proposed gait-based identification pipeline, comprising of signal processing, feature extraction, and gait authentication methodology.

5.3.1 EarGate: system at a glance

The overall functioning of EarGate is reported in Figure 5.3.1.

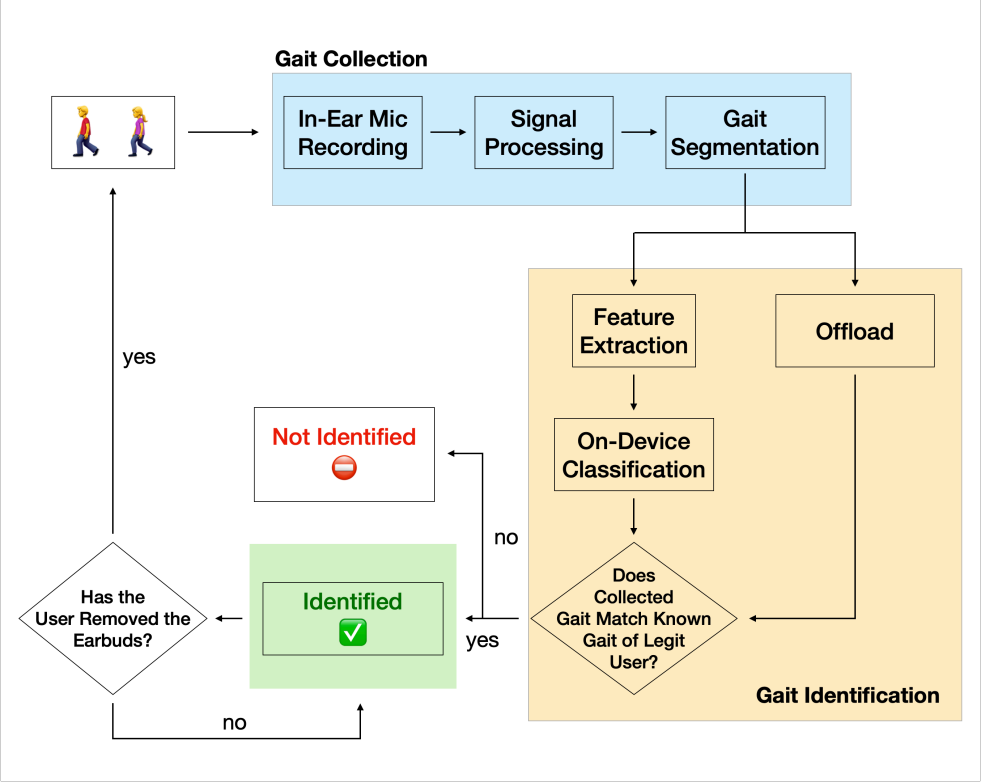


Fig. 5.3.1: EarGate functioning.

Initially, the user has to take part in an enrollment phase, a required stage during which the system acquires the data and process them (pre-processing and feature extraction) before training the model to recognize the acoustic-gait of the legitimate user. Notably, as we will discuss in Section 5.5.9, a small number of steps and, therefore, little time, is sufficient to achieve good identification performance. Once the enrollment phase is over, the system is ready to operate. As shown in Figure 5.3.1, EarGate silently collects, and pre-processes on-device, acoustic-gait data. The system is provisioned to either execute the (*limited*) computation required to identify the user (or reject them) on-device, or to offload it, e.g., to the cloud, a smartphone, or a remote server (Section 5.6).

Notably, there is NO need for the user to be *constantly* and *instantly* walking to be identified. Instead, EarGate can perform a one-time identification once the user wears the earbuds and walks a couple of steps (e.g., to grab a coffee, to go to the restroom, etc.). The identification result (either recognized as legitimate user, or not) is then considered valid until the user removes their earbuds (i.e., the only chance the wearer has of switching identity). Detecting such occurrence is, in practice, very simple, as the natural properties of the occlusion effect will suddenly disappear from the in-ear microphone recordings whenever the user removes the earbuds. Besides, commercially available earbuds, like Apple AirPods, already do that (i.e., to automatically stop the music whenever the user removes their earbuds). Such a scheme significantly relieves the user burden of *"walk now to be identified*" and saves system energy as one identification could be valid for a longer time.

5.3.2 Signal processing

Our prototype records the microphone outputs in $48kHz$, while the generated step sounds are at relatively low frequencies. Notably, in this chapter, the term *frequency* refers to the *pitch* of the step sound, rather than to the cadence/speed of walking. To minimize the computation overhead during processing, we first down-sample the recorded data to $4KHz$. Then, a low-pass filter with a cut-off frequency of $50Hz$ is applied to eliminate the high-frequency noise. The choice of a $50Hz$ cut-off frequency is to guarantee a good signal-to-noise-ratio of the walking signal, whilst retaining the robustness to most environmental noise (typically higher than $50Hz$). Moreover, due to the large attenuation of sound propagating through the air, as well as the blockage of the ear canal opening, the majority of the noise is suppressed or canceled directly in the ear canal. Figure 5.3.2 shows the low-pass filtered signal from the left earbud when one participant is walking on tiles with

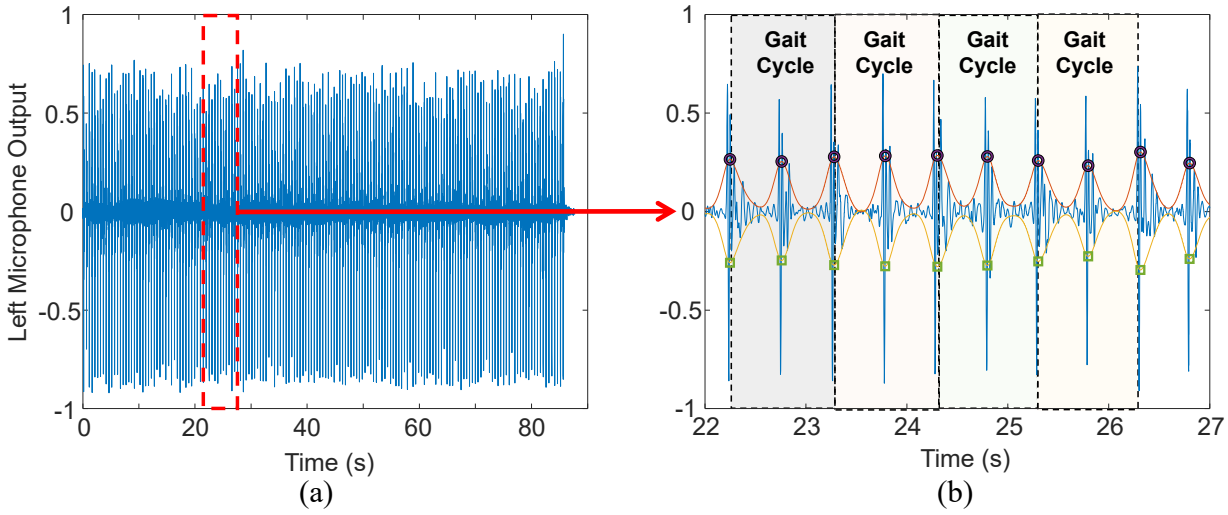


Fig. 5.3.2: (a) The low-pass filtered signal collected when one participant is walking, (b) a segment showing the performance of proposed gait segmentation algorithm.

sneakers. We can observe that an acoustic gait cycle is composed of two spikes (happening when the foot hit the floor in the *strike phase*) and two relatively flat (silent) periods (denoting the *swing phase*). This is different from sinusoidal-like patterns recorded by the IMU [10], therefore, existing gait segmentation approaches adopted for IMU data are not applicable. Hence, we propose a peak detection based algorithm to segment the audio signals and extract the acoustic gait cycles.

Specifically, we first use a Hilbert transform to extract the envelopes of the filtered signal. We proceed by applying a low-pass filter (with a cut-off frequency of $3Hz$) on the envelops to smooth them, as illustrated by the red (upper envelop) and yellow (lower envelop) curves in Figure 5.3.2b. Then, we perform peak detection on the filtered envelops and regard the peaks as the points when the human foot hits the ground. Whenever a pair of upper peak-lower peak is aligned, we treat it as a step. We select the gait cycle starting points by skipping one between every two peaks, as each gait cycle consists of two steps. Finally, we consider a gait cycle the samples between every two consecutive starting points. Most of the extracted gait cycles last for around one second. Therefore, we interpolate them into the same length of 4000 samples using spline interpolation.

5.3.3 Feature extraction

To differentiate between users, EarGate leverages a set of features that, we believe, could reliably represent the characteristics of user gait from each cycle. For the feature extraction, we rely on *librosa* [138], a Python package specific for audio processing. Specifically, the features we look at, covering the frequency, structural, statistical, and temporal characteristics of the data, are:

- Mel-Frequency Cepstral Coefficients (MFCC): obtained from the short-term power spectrum, MFCC certainly is one of the most common and known features in audio processing [158];
- Chroma of Short-Time Fourier Transform (STFT);
- Mel Spectrogram: the signal spectrogram in the Mel-scale;
- Root-Mean-Squared Energy (RMSE): the Root-Mean-Squared (RMS) of the STFT of the signal, which provides information about the power of the signal;
- Onsets: the peaks from an onset strength envelope, result of a summation of positive first-order differences of every Mel-band.

Rather than only using the data recorded by either the left-earbud-microphone or those coming from the right-earbud-microphone, separately, inspired by the positive results achieved with OESense (Section 4.6) we instead fuse them concatenating the features we extracted from each. As we have discussed, unlike other wearables (e.g., smartwatches) earables inherently come in a pair and, therefore, it is possible to leverage, and fuse, their two independent measurements of the same phenomenon.

5.3.4 Identification methodology

Like most identification systems (e.g., FaceID and TouchID), before being able to run the online identification, EarGate requires an enrollment phase. It is during this phase that the legitimate users provide samples of their gait to system, thus training a model to classify the users as either legitimate users, or impostors. Notably, all the users that have not been seen by the model in the enrollment phase will be regarded as impostors. In this work, we consider two enrollment schemes: **(i)** with and **(ii)** without impostor data. The former leverages a pre-trained model with benchmark impostor data, together with some data from the legitimate user. However, given it is not always possible to assume

the availability of benchmark impostor data, especially shortly after the release of the system, we also look at a model solely trained on the legitimate user data. In either cases, the system needs some walking data samples from the legitimate user and, therefore, will instruct the user to follow an enrollment protocol (basically walking for a few steps in order to train a classifier). For the online-identification framework we adopt: **(i)** a two-class Support Vector Machine (SVM) classifier (if benchmark impostor data are available); or **(ii)** a one-class SVM classifier (when we only have the data of the legitimate user), due to its high computational efficiency and low complexity [76].

Prior work unequivocally showed that gait is a unique user fingerprint, and that it is actually very hard for an impostor to impersonate another person's gait [159, 160]. Hence, the aim of this chapter is to showcase the potential of earables as a personal-identification device. First and foremost, we sought to assess whether our in-ear microphone-based approach is capable of recognizing the user in such a way that if others (i.e. *impostors*) were using the device, it would be able to spot it. We thus consider both **(i)** *replace attacks* (a different user mistakenly tries to use the earables) and **(ii)** *mimic attacks* (a malicious attacker deliberately tries to use the earbuds by actually impersonating the user, i.e., simulating the user's gait). As highlighted by our evaluation, different users can be distinguished very clearly by our system, thus making mimic attacks even more unfeasible. To further clarify that, for instance, let us assume there is a very well-trained impostor, who can accurately mimic the gait of the legitimate user, generating the very same vibration patterns whenever their feet hit the ground. However, unfortunately for the impostor, even given all the most favorable conditions, when the vibrations propagate through their body and bones, all the way to the impostor's ear canal, they will inevitably be different from those belonging to the legitimate user. This is because, as discussed in Section 2.2.1, the human body and skeleton act as a natural modulator. Hence, in the remainder of the paper we will focus on showing the performance of our system against the more common replace attacks.

5.4 Implementation

This section provides the design details our prototype and describes the data collection procedures we followed.

5.4.1 EarGate prototype

To have full control on the data (sampling rate as well as unlimited access), we decided to leverage the same prototype we had built for OESense (Section 4.4.1). During the whole process, we were driven by the requirements we stated in section 5.2.2.

5.4.2 Data collection

After having obtained clearance for carrying out the studies from the Ethics Board of our institution, we recruited 31 subjects. Out of the 31 participants, 15 of them were females and the remainder 16 were males. The mean age of the participants was 26.6±5.8. All the participants received a voucher in exchange of their time. After having taken COVID-19 related precautions (i.e., face masks, hands sanitizing, etc.), we admitted the participants, one at a time, into the room of the experiment (a 12 $\times$ 6 square meters room). We instructed the participants to walk in circles following the perimeter walls of the room. The room was quiet, with a noise level of approximately $30dB$. We asked the volunteers to keep their normal pace and walking style. As a gait cycle is composed of two consecutive steps, the subjects always started with their left foot and finished with the right foot when walking. We considered different ground materials, as tiles and carpet, and multiple conditions, with the participants walking barefoot, with slippers, with sneakers, and while speaking. Factoring in all these variables allows us to further assess the robustness, and generalizability, of EarGate. For each of these conditions, the participants walked continuously for 1.5 minutes (a session). During every walk, each subject counted the number of steps they took. This served as ground truth. Walking at normal speed, all the volunteers made $156 - -176$ steps per 1.5 minutes-long session. Eventually, each participant performed 8 (2 ground material $\times$ (3 footwear + 1 speaking)) different walking sessions, accounting for a grand total of 52046 steps (i.e., 26023 gait cycles).

5.5 Performance Evaluation

In this section, we present the data collection procedure, the training methodology, as well as different variables we consider while assessing the performance of our system.

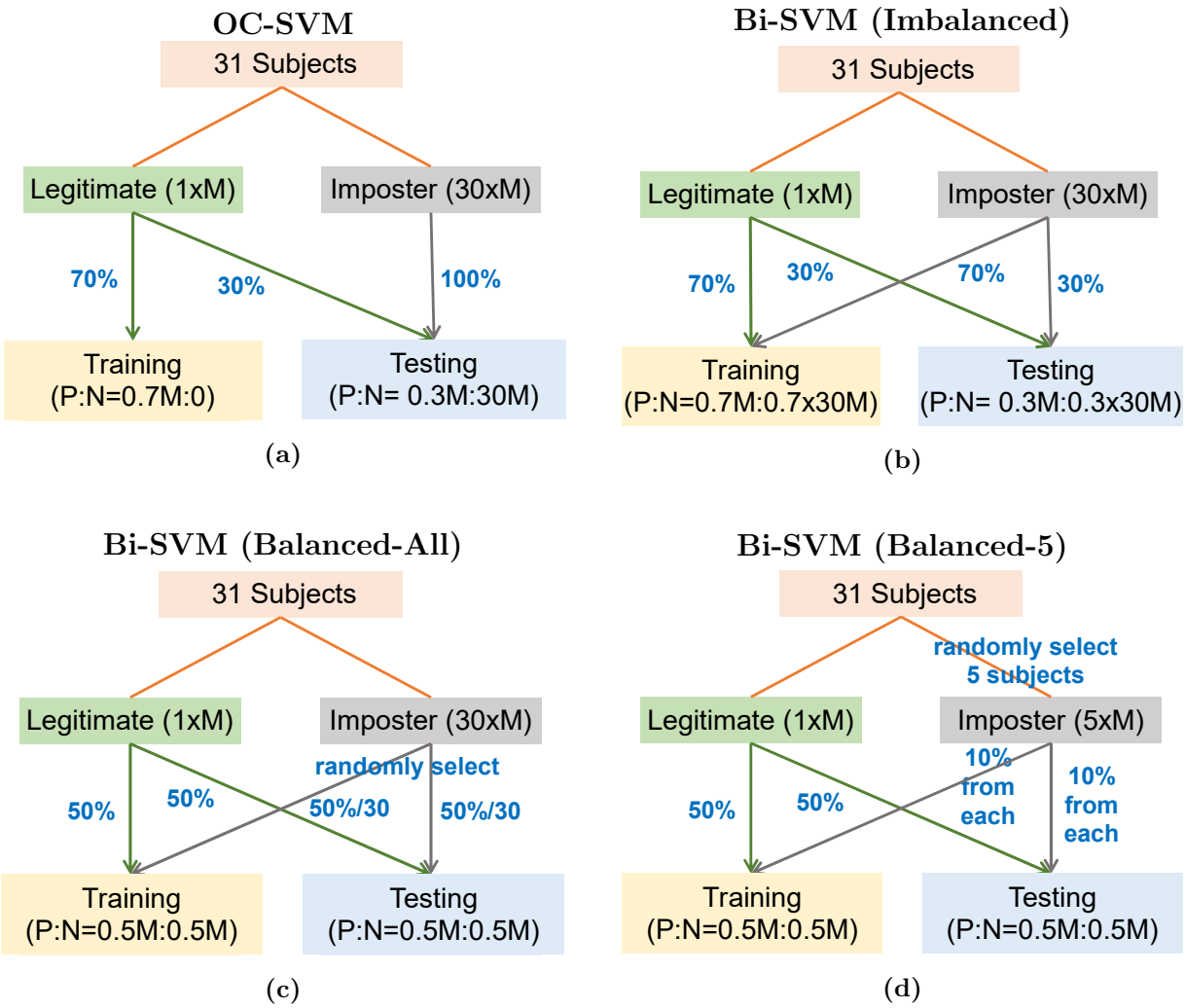
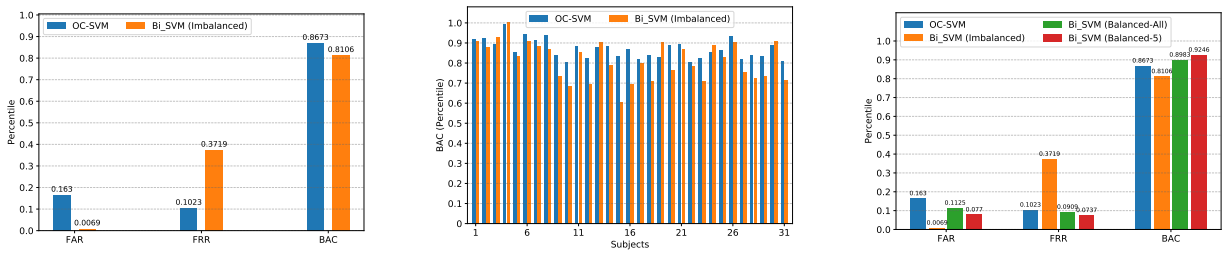


Fig. 5.5.1: Training and testing data splitting scheme for 5.5.1a one-class SVM, 5.5.1b imbalanced binary SVM, 5.5.1c balanced binary SVM with all subjects' data, and 5.5.1d balanced SVM with part of subjects' data. $P : N$ represents the ratio between positive (legitimate) and negative (impostor) gait.



(a) (b) (c)

Fig. 5.5.2: 5.5.2a Overall performance of EarGate (averaged across 31 subjects), 5.5.2b BAC of each individual using OC-SVM and Bi-SVM (Imbalanced), 5.5.2c Comparison of the four proposed training-testing protocol, which also reflects the impact of data imbalance.

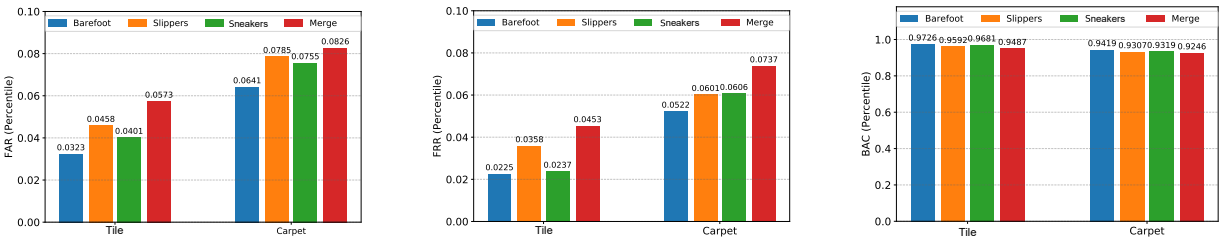
5.5.1 Metrics

To assess the goodness of EarGate, we rely on three metrics commonly used to evaluate identification tasks:

- **False Acceptance Rate** (FAR): describes the system's likelihood of successfully identifying a non-legitimate-user;
- **False Rejecting Rate** (FRR): also known by the name of False Negative Rate (FNR), it indicates the identification system's likelihood of rejecting the legitimate-user;
- **Balanced Accuracy** (BAC): given we also assess the performance of our system in the case of unbalanced training and testing sets, we consider BAC to gauge the real accuracy of our system. Specifically, BAC is defined as $\frac{TPR+TNR}{2}$, where TPR and TNR are True Positive and True Negative Rate, respectively. The True Positive Rate of an identification system is its goodness in recognizing legitimate users, whilst True Negative Rate indicates the value of the system in protecting the user from attackers.

5.5.2 Training-testing protocols

To evaluate our system, we propose four different training-testing protocols, as shown in Figure 5.5.1. The number of gait cycles collected from each of the 31 subjects is denoted as M. The first two schemes (5.5.1a and 5.5.1b) involve different amounts of positive and negative data during training and testing. Such data imbalance might affect the performance. Thus, we further propose two protocols (5.5.1c and 5.5.1d) to ensure balanced positive and negative samples during training and testing. All the four schemes consider



(a) **(b)** **(c)**

Fig. 5.5.3: Performance comparison 5.5.3a FAR, 5.5.3b FRR, and 5.5.3c BAC, of different footwear and ground material.

adversarial attacks during testing. Notably, informed by the physiological explanation of gait (Section 2.2.4), we did not review mimic attacks: by being so strongly correlated with the muscular-skeletal structure of the individual, gait is extremely hard to mimic – if not impossible. Moreover, the modulation of the in-ear (bone-conducted) audio by the skeleton of the subject constitutes an additional barrier against mimic attacks.

- **(a)** One-Class SVM (**OC-SVM**): one subject is iteratively selected as the legitimate user, while the rest of the users are regarded as impostors. Training dataset only consists of 70% ($0.7 \times M$) data from the legitimate user, and the testing dataset is composed of 30% ($0.3 \times M$) legitimate user data and all impostor data ($30 \times M$). The scheme is reported in Figure 5.5.1a.

- **(b)** Imbalanced Binary SVM (**Bi-SVM (Imbalanced)**): one subject is iteratively selected as the legitimate user, while the remainder users are regarded as impostors. Training dataset consists of 70% legitimate user's data ($0.7 \times M$) and 70% impostors' data ($30 \times 0.7 \times M$), and the testing dataset is composed of 30% ($0.3 \times M$) legitimate user's data and 30% ($30 \times 0.3 \times M$) impostors' data. The scheme is reported in Figure 5.5.1b.

- **(c)** Balanced Binary SVM with all subjects' data (**Bi-SVM (Balanced-All)**): one subject is iteratively selected as the legitimate user, while the reminder are regarded as impostors. Training dataset consists of 50% ($0.5 \times M$) of the legitimate user's data and the same number ($0.5 \times M$) of gait cycles that are randomly selected from the 30 impostors. The testing dataset is composed of the rest 50% ($0.5 \times M$) of the legitimate user's data and another $0.5 \times M$ randomly selected impostor data. The scheme is reported in Figure 5.5.1c.

- **(d)** Balanced Binary SVM with part of subjects' data (**Bi-SVM (Balanced-5)**):

one subject is iteratively selected as the legitimate user and five (5/30) subjects are randomly selected as impostors. Training dataset consists of 50% ($0.5 \times M$) legitimate user's data and 10% from each of the five impostors ($5 \times 0.1 \times M$). The testing dataset is composed of the rest 50% ($0.5 \times M$) of the legitimate user's data and another 10% ($5 \times 0.1 \times M$) data from each of the impostors. The scheme is reported in Figure 5.5.1d.

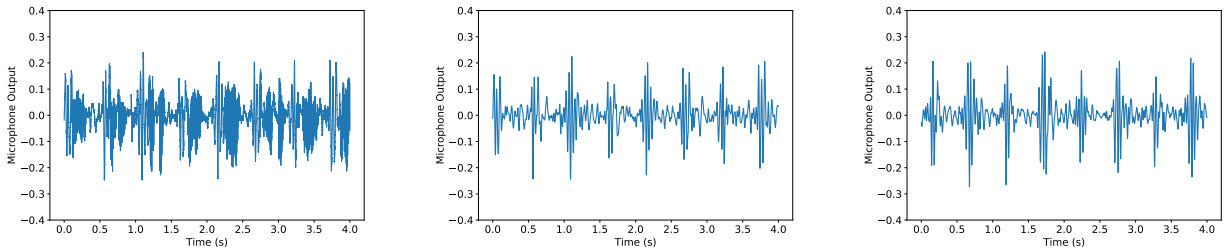
In the remainder of this section, we present our experimental results under various conditions.

5.5.3 Overall performance

We first evaluate the overall performance of EarGate by combining all the collected data together, i.e., different ground materials and footwear. Figure 5.5.2a reports the results obtained with the first two training protocols (OC-SVM and Bi-SVM (Imbalanced)), which are averaged over the 31 subjects. We can observe that with both methods, EarGate can achieve over 80% balanced accuracy (BAC). Notably, both the FAR and the FRR are relatively high. This might be due to the imbalanced dataset and the variability across the different waking conditions. We will discuss these more in depth in the following subsections. Figure 5.5.2b plots the BAC of each individual with the two protocols. We can see that although the performance varies among subjects, most of the subjects achieve over 80% BAC. In addition, we found that the best training-testing protocol is subject-dependent as some users obtain higher BAC with OC-SVM, while others achieve better performance with Bi-SVM (Imbalanced). Thus, it is necessary to optimize the training protocol for each individual.

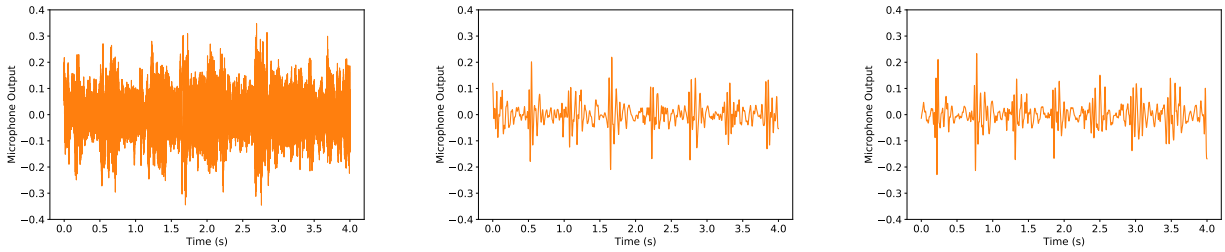
5.5.4 Impact of data imbalance

Intuitively, Bi-SVM is expected to perform better than OC-SVM. The reason is that OC-SVM is similar to a clustering problem and the model has to learn the correlation within the data without information on the outliers. While for Bi-SVM, the inclusion of benchmark (impostor) data provides additional information about the negative samples, so that the model can learn a more accurate representation of the legitimate user. However, as observed in Figure 5.5.2a, Bi-SVM performs worse. Specifically, the FRR is dramatically higher than the FAR. Such phenomenon is also observed in [161] and we suspect this originates from the imbalanced positive and negative samples during training and testing. Hence,



(a) (b) (c)

Fig. 5.5.4: Impact of user speaking. We take one subject walking on tiles as an example, 5.5.4a original signal with speaking, 5.5.4b low-pass filtered signal with speaking, 5.5.4c filtered signal without speaking.



(a) (b) (c)

Fig. 5.5.5: Impact of music playback. We take one subject walking on tiles as an example, 5.5.5a original signal with music playing, 5.5.5b low-pass filtered signal with music playing, 5.5.5c filtered signal without music.

we decided to re-run all the experiments using the proposed two training schemes with balanced samples (Figure 5.5.1c and Figure 5.5.1d).

As shown in Figure 5.5.2c, the large difference between the FAR and the FRR disappears as soon as the positive and negative classes are balanced. As expected, Bi-SVM (Balanced-All) and Bi-SVM (Balanced-5) yield better performance (lower FAR/FRR and higher BAC), with the BAC significantly enhanced, from 81% to 92.5%. Notably, Bi-SVM (Balanced-5) performs even better if the impostor data is also balanced (10% from each impostor). Therefore, for the reminder of the evaluation, we will focus our analysis around the results we obtained with Bi-SVM (Balanced-5).

5.5.5 Impact of walking conditions

Different walking conditions, such as different footwear or ground materials, might affect the gait identification performance of EarGate. For instance, compared to walking on tiles,

walking on a carpet will result in longer stance phase and weaker vibrations when hitting the ground. These conditions inevitably introduce variations from the subject's standard gait and, therefore, make the identification task harder. In the next sections, we investigate the impact of footwear and ground material.

Footwear

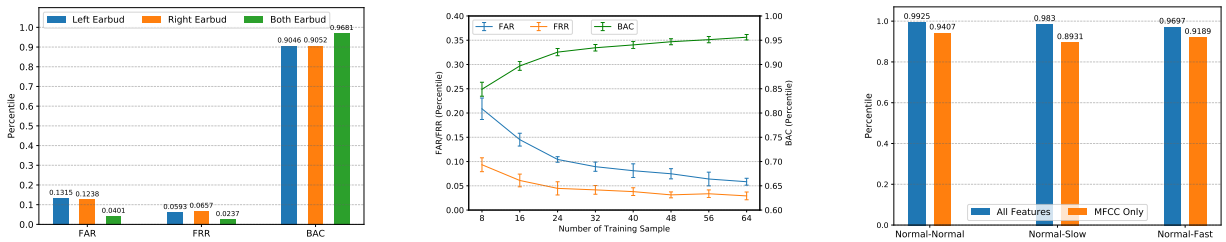
Figure 5.5.3 shows the FAR, FRR, and BAC with data collected when the subjects were barefoot, wearing slippers, wearing sneakers, as well as the combination of all these data, respectively. The results indicate that EarGate works well regardless of the footwear, with a BAC higher than 94%. Particularly, walking barefoot yields the best identification performance, followed by wearing sneakers. Interestingly, wearing slippers is the most challenging case, as slippers introduce more variations during walking. This observation is applicable to both the datasets collected on tiles as well as on carpet. In addition, when evaluating the data collected in a single session, the BAC increased significantly, jumping from 92.46% (Figure 5.5.2c) to 97.26%.

Ground material

We now explore the impact of ground material on the identification performance. As reported in Figure 5.5.3, EarGate achieves good performance with users walking both on tiles and carpet, with a BAC higher than 93%. In particular, for the same footwear (e.g., sneakers), tiles always yield better performance (3% improvement on BAC) compared to carpet. This might be because soft carpet dampens part of the vibrations generated when hitting the ground, and, therefore, the signal-to-noise ratio (SNR) is lower.

5.5.6 Impact of human speech

Human speech might also have an impact on the proposed identification system. On one hand, the voice produced when people speak will be captured by the in-ear microphone, thereby polluting the recorded acoustic gait data. On the other hand, while speaking, the movement of mouth and jaw modifies the structure of human body. Consequently, the generated vibrations will experience a slightly different propagation path, resulting in different modulations of the gait signal. Thus, we asked the subjects to speak while walking on tiles with sneakers (the most common case for daily walking) and we assessed whether the performance of EarGate could be impacted.



(a) **(b)** **(c)**

Fig. 5.5.6: Impact of 5.5.6a fusing data from both the earbuds, 5.5.6b different training size, and 5.5.6c different pace and features, on FAR, FRR, and BAC.

Figure 5.5.4 plots the raw microphone data from the left earbud. It is clear that the gait signal is polluted by high-frequency human speech. Fortunately, the actual gait signal we are interested in can be distinguished in the frequency domain as **(1)** the generated vibrations are in low frequencies and **(2)** the occlusion effect mainly emphasizes the low-frequency components of the bone-conducted sound. Figure 5.5.4b illustrates the low-pass filtered version of Figure 5.5.4a. In addition, we plot the filtered signal of the same participant walking without speaking in Figure 5.5.4c. Visually, the two filtered signals look quite similar and the impact of human speech has been completely removed. Using the Bi-SVM (Balanced-5) protocol, we run the experiments on the collected dataset with users speaking during walking. The results are satisfactory for both tiles (FAR=7.73%, FRR=4.37%, BAC=93.95%) and carpet (FAR=10.73%, FRR=6.98%, BAC=91.15%), denoting how human speech has little impact on EarGate.

5.5.7 Impact of music playback

The general purpose of earbuds is leisure/entertainment, particularly music playback. Due to the vicinity (less than 1 centimeter) of the speaker and in-ear microphone, we investigated whether music playback induces interference in the gait identification system. To explore that, we asked one subject to walk when the earbud was playing music at an appropriate volume. Figure 5.5.5a plots the original signal from the left earbud, where the target gait signal is overwhelmed by the music. Similarly to human speech, these audible sounds are in higher frequency bands and EarGate captures gait signals which are low frequencies ($< 50Hz$). Thus, after applying a low-pass filter, music, as with speech, can easily be eliminated, and the gait signal of interest can clearly be observed in Figure 5.5.5b. As we have done previously, we plot a trace (low-pass filtered) of the same subject walking without playing the music in Figure 5.5.5c, which shows high similarity with signals

Table 5.5.1: Identification performance improvements achieved with transfer learning.

Tasks	Type of Features	SVM			XGBoost		
		FAR	**FRR**	**BAC**	**FAR**	**FRR**	**BAC**
Brick Barefeet	**VGGish embeddings**	0,730	0,231	0,520	0,093	0,092	**0,907**
	Handcrafted Features	0,081	0,023	**0,948**	0,077	0,070	0,927
	VGGish embeddings + hancrafted features	0,069	0,017	**0,957**	0,065	0,067	0,934
Brick Shoe	**VGGish embeddings**	0,645	0,333	0,511	0,133	0,088	**0,889**
	Handcrafted Features	0,048	0,022	**0,965**	0,060	0,059	0,940
	VGGish embeddings + hancrafted features	0,032	0,022	**0,973**	0,073	0,055	0,936
Brick Slipper	**VGGish embeddings**	0,640	0,316	0,522	0,052	0,113	**0,917**
	Handcrafted Features	0,085	0,026	**0,945**	0,071	0,064	0,932
	VGGish embeddings + hancrafted features	0,056	0,016	**0,964**	0,048	0,073	0,940
Carpet Barefeet	**VGGish embeddings**	0,716	0,202	0,541	0,132	0,147	**0,861**
	Handcrafted Features	0,096	0,034	**0,935**	0,089	0,101	0,905
	VGGish embeddings + hancrafted features	0,052	0,036	**0,956**	0,085	0,097	0,909
Carpet Shoe	**VGGish embeddings**	0,589	0,378	0,517	0,160	0,126	**0,857**
	Handcrafted Features	0,116	0,051	**0,917**	0,104	0,085	0,906
	VGGish embeddings + hancrafted features	0,056	0,048	**0,948**	0,093	0,086	0,910
Carpet Slipper	**VGGish embeddings**	0,548	0,404	0,524	0,129	0,141	**0,865**
	Handcrafted Features	0,069	0,043	**0,944**	0,115	0,107	0,889
	VGGish embeddings + hancrafted features	0,050	0,038	**0,956**	0,130	0,104	0,883

in Figure 5.5.5b.

To further demonstrate how EarGate is robust against different types of music and phone calls, we conducted a similar analysis to that done for OESense in Section 4.6.5. Ultimately, our analysis shows EarGate is compatible with general purpose earbuds, without introducing any mutual interference. Moreover, given many existing earbuds already features in-ear microphones for active noise cancellation (ANC), EarGate can be operated concurrently. Specifically, the data from in-ear microphones can be delivered to two independent pipelines: noise cancellation algorithm to actively generate the anti-noise and user identification framework to recognize the user.

5.5.8 Sensor multiplexing

As we have widely discussed in this dissertation, unlike other wearable devices, earbuds possess the unique advantage of being able to sense with both the left and right earbud simultaneously, providing a sensor multiplexing gain. Taking the data collected when subjects walked on tiles with sneakers as an example, we study the performance of using solely the left or right earbud, as well as the achievable gain with both earbuds. As presented in Figure 5.5.6a, when a single earbud is used, the left and right one achieves quite similar performance on the three metrics, suggesting both of them are effective in detecting users' gait. When combining the features extracted from both earbuds, the

BAC increases from 90.5% to 96.8%, denoting a great multiplexing gain when using both earbuds.

5.5.9 Impact of training size

To gauge the overhead of EarGate during enrollment phase, we study the impact of training data size on the identification performance as the minimal number of gait cycles required for training is the actual burden on users. Specifically, we select 80 gait cycles for each subject from the tiles-sneakers dataset and run the experiment with Bi-SVM (Balanced-5) protocol. 20% (i.e., 16) gait cycles are fixed as the testing dataset to ensure a fair comparison. Then, we continuously increase the number of training samples from 10% (i.e., 8) to 80% (i.e., 64). We run the experiment iteratively selecting one subject as the legitimate user. Figure 5.5.6b shows the averaged performance over the 31 subjects. As expected, the performance improves with the increase of training samples. The BAC reaches 93.5% when 32 gait cycles are used for training. Based on the walking speed of participants, the corresponding overhead for data collection is around $30s$ continuous walk, which would be acceptable for most people.

In addition, EarGate can adapt its identification model by re-training it with new gait samples. For example, if a new gait cycle is recognized as positive/negative with high confidence, it can be added to the training set for model re-training. Thus, the performance of EarGate is expected to improve continuously after it has been put into practical use.

5.5.10 Transfer learning

We sought to further improve the performance of EarGate by looking at more complex, deep learning based approaches. Although the performance of traditional deep learning techniques are hindered by the modest size of our dataset (the number of available data points is not sufficient for the model to properly learn the weights[1]), transfer learning may come handy [162]. In doing so, we leveraged VGGish, an audio-specific pre-trained model released by Google [163]. VGGish is a convolutional neural network (CNN) that is trained using a large-scale YouTube audio dataset. Concretely, we use VGGish to automatically extract audio features (*embeddings*) from the raw in-ear audio data. At this point, we combined the handcrafted features with the embeddings extracted by VGGish and trained two different classifiers: SVM (as in our previously reported results) and XGBoost, an

[1]Considering VGGish as an example, the number of trainable parameters is $4,499,712$.

advanced decision-tree based machine learning algorithm [164]. We report our findings in Table 5.5.1. As we can appreciate from Table 5.5.1, XGBoost consistently outperforms SVM when solely using the VGGish embeddings, without any handcrafted feature. Conversely, by feeding handcrafted features to SVM we achieve better results (both in terms of BAC as well as, in most cases, in terms of FAR and FRR). Notably, by combining the embeddings extracted with transfer learning and the manually engineered features, we can further boost the performance of EarGate, increasing the BAC and lowering both FAR and FRR. This is due to the model learning from both the carefully selected features as well as from the more abstract representation of the data generated by VGGish. Interestingly, while in terms of accuracy the benefit of transfer learning is clear, it is also important to bear in mind the accuracy versus power consumption trade-off. The performance achievable with only the handcrafted features are indeed only marginally lower than those achieved by combining the transfer learning embeddings and the manually crafted features, in face of an inevitably higher power consumption (caused by the execution of the VGGish model). However, if a larger dataset is available, the performances associated with transfer learning could further improve.

5.5.11 Contribution of specific features

Our classifier has been trained on a variety of features, including many related to the frequency spectrum. For an evaluation of our scale, it might be that the walking cadence can be distinct for all the participants and, therefore, there is the concrete risk that the model learns to strongly weight the frequency features closely associated with people's cadence. Ultimately, this would be problematic because walking cadence will not be distinct on the broader population level and, furthermore, people move at different cadences (e.g., when walking alongside someone).

To confirm that our classifier is not simply a cadence classifier, we asked one subject to walk at three *speeds*: slow ($1.59 step/s$), normal ($1.97 step/s$), fast ($2.13 step/s$). After training the model with normal-speed walking data, we tested it on all the three instances. Figure 5.5.6c reports the BAC for normal, slow, and fast pace respectively: 99.25%, 98.30%, and 96.97% BAC. These results clearly show that the model is not biasing towards cadence as the identifier, and is instead learning something more fundamental to the user's movement.

Given this, we also consider the value of the different features. After training our classifier

Table 5.5.2: Power consumption and latency measurement of EarGate.

Scheme	Operation	Power (mW)	Latency (ms)	Energy (mJ)
On-device identification	MicRecd	120	1000	
	LowPassFilt	635	1.83	168.59 (All)
	FeatExtr (All/MFCC)	655/651	71.98/23.62	136.82 (MFCC)
	Inference	644	0.44	
Raw Data Offloading	MicRecd	120	1000	
	TX [OS+Air] (WiFi)	334	9.49+12.8	123.17 (WiFi)
	TX [OS+Air] (BT)	478	148.41+128	190.94 (BT)
Feature Offloading	MicRecd	120	1000	
	LowPassFilt	635	1.83	168.91 (WiFi, All)
	FeatExtr (All/MFCC)	655/651	71.98/23.62	172.27 (BT, All)
	TX [OS+Air] (WiFi)(All/MFCC)	332	1.81+0.59/0.39+0.26	136.67 (WiFi, MFCC)
	TX [OS+Air] (BT)(All/MFCC)	457	8.66+5.94/5.74+2.56	139.26 (BT, MFCC)

individually on each of the features listed in Section 5.3.3, we plot in Figure 5.5.6c the BAC achieved training our system with only Mel-Frequency Cepstral Coefficients (MFCC), as well as with all the features. Notably, among all the features taken individually, MFCC achieves the best results. Interestingly, when analyzing the computation time of different features, we find that most of the feature extraction time (89%) is actually consumed on a feature called '*tonnetz*' (with 6 values). We then repeat the experiment after removing this feature and the identification accuracy almost remains the same. Therefore, it is possible to use a smaller set of features and reduce the overall end-to-end latency of the system.

5.6 System considerations

To gauge the system-level performance of EarGate, we conducted a power-consumption and latency investigation using the same prototype described in Section 5.4. We assume the model is pre-trained and only consider the real-time identification overhead. Given EarGate can either run the identification framework on-device, or offload it, we consider three different schemes which would impact differently power consumption and latency:

1. **On-device identification**: microphone data recording (MicRecd), as well as all the gait identification procedures, including low-pass filtering (LowPassFilt), feature extraction (FeatExtr), and inference, are performed on-device.
2. **Offloaded identification (raw data offloading)**: the earable records microphone data and directly transmit the raw data via WiFi or Bluetooth (BT), without any

processing. The size of raw data is $16KB$ [2]. The identification process would be performed on the cloud or companion smartphones and possibly the result will be communicated back to the device.

3. **Offloaded identification (features offloading)**: both low pass filtering and feature extraction are carried out on the earable, right after recording the microphone data. Only the extracted features (either all features or only MFCC features) are transmitted via WiFi/BT and the result might be communicated back to the device. The size of features is $0.748KB$ for all features and $0.16KB$ for MFCC features [3].

Our aim here is to show the flexibility of our system to work, irrespective of the network architecture preferred or available.

For the offloading cases, we consider two types of radio frequency (RF) communications: Bluetooth (BT) and WiFi, as they are (or might soon be) the most commonly available radio chips in earables. For WiFi, we consider a typical uplink throughput of $10Mbps$ (similar to that of a domestic network) to compute the transmission latency over the air. Regarding Bluetooth, the version supported by the Raspberry Pi we use is BT 4.1. Compared to more recent BT standards (e.g., Bluetooth 5, available in Apple AirPods Pro), BT 4.1 offers less throughput ($1Mbps$ instead of $2Mbps$). As a consequence of that, the over-the-air transmission time reported is longer than what it would be if a more advanced version of BT were adopted. We measure the power consumption with a USB power meter. Latency measurements are obtained timing the execution of the software handling the operation of interest. The results are averaged over multiple measurements and presented in Table 5.5.2. The baseline power consumption of our Raspberry Pi (idle) is around $2,325mW$ and the values reported in the table are additional power consumption. The energy column computes the total energy required to perform the operations for one gait cycle.

On-device identification performs the whole pipeline including microphone recording (MicRecd), filtering (LowPassFilt), feature extraction (FeatExtr), and inference on the device. The power column indicates that numerical computations (LowPassFilt, FeatExtr, and Inference) are intensive and more power-hungry than microphone recording. The latency column shows that most of the processing time is spent on feature extraction. Regard-

[2] With a sampling rate of $4000Hz$ and gait cycle length of 1 second, the data from two microphones is $2 \times 4000 \times 2B = 16KB$.

[3] All features includes 187 features from each microphone data and the size is $2 \times 187 \times 2B = 0.748\ KB$, MFCC features include 40 features and the size is $2 \times 40 \times 2B = 0.16\ KB$.

less of the time for data acquisition, the overall identification latency is within $100ms$. Concretely, the energy required for a one-time on-device identification is $168.58mJ$ (using all the features) and $136.82mJ$ (using MFCC features only, the most effective features as described in Section 5.5.11).

When offloading the raw data to the cloud, only MicRecd is performed on-device. Here, we consider the latency as the sum of TX OS (the time that Pi requires to write the data in the buffer of the chosen interface, either Bluetooth or WiFi) and TX Air (the time it takes for the data to propagate over-the-air). We can observe that the latency is largely dependent on the throughput of the network. The energy required for WiFi offloading is $123.17mJ$, whilst for BT is $190.94mJ$. Conversely, when offloading the features, also LowPassFilt and FeatExtr are done on-device. Here, following Section 5.5.11, we compare the energy efficiency of sending all the features or just the MFCC features. Given there are 187 features in total and only 40 are MFCC features, the transmission latency for MFCC features only is shorter, overall below $100ms$.

In summary, we can conclude that (1) both on-device identification and features offloading guarantee a time delay (after data acquisition) of less than 100 ms, showing EarGate can work in real-time scenarios both on-device and while offloading features; (2) with respect to the energy consumption, all the schemes consume less than $200mJ$ for one-time identification; (3) the latency and energy consumption for raw data offloading are strictly dependent on the quality of the communication link. Thus, offloading raw data would be the best option when the network is stable. Notably, in this section we focus on the SVM-based pipeline, only considering the case where handcrafted features are used to train the model. We do this for several reasons. First, at the moment, deploying a complex model like VGGish on resource-constrained devices (like earbuds) is an extremely challenging task and there are no publicly available tools to supporting it. Second, given the performance improvements of the transfer learning approach over the handcrafted features one are marginal (Section 5.5.10). Third, running VGGish to extract the embeddings would entail far more operations than simply leveraging a limited set of handcrafted features to train SVM and, therefore, it is fair to assume that power and latency figures would be considerably higher than those for the traditional machine learning pipeline.

5.7 Discussion

In this section, we talk through the limitations of EarGate, the possible improvements, and the future directions we plan on pursuing. Despite the promising results, we are aware of some of the limitations of our approach. First and foremost, in order for our system to work, initial user data (in the enrollment phase) are required. Concretely, this means there has to be an enrollment phase during which the system acquires data about the legitimate user. While we acknowledge it would be ideal if such a phase did not have to take place, most of the well-established biometric identification schemes, like facial recognition and fingerprint, do require some initial user data. Besides, in our evaluation (Figure 5.5.6b) we show we only need a limited amount of user data to start offering acceptable performance.

Second, whilst Figure 5.5.3 shows that the impact of different footwear is marginal, when the model is trained on them; to maintain high performance, the model should be partially re-trained to add new pairs of shoes to the legitimate user data. This could be done by the user walking and manually committing the new gait cycles to re-train the model; or it could happen in background if the user changes shoes while wearing their earables (provided the user has been successfully identified beforehand). We believe the latter is a reasonable assumption, especially for hearing aid users as such devices are continuously worn throughout the day. Hopefully, this will soon be the case for leisure earables, too, which, once the advancements in materials will guarantee better comfort, could also be worn for a longer amount of time. Further, although gait may slightly change over the years, continuously adapting the model (like we do for different shoes) could relieve this issue, too. Alternatively, the impact of different footwear could be obviated by mean of the combination of a general model and a personalized model. For instance, an auto-encoder [165] could be trained on all the users' data to obtain a general model.

Lastly, one other concern could be related to the obstruction of the ear canal orifice, and the consequent impact on audible sounds (which will result muffled due to the presence of an obstructing body). We are aware this could be a potential safety issue, therefore, similarly to the AirPods Pro *Transparency Mode*, it is possible to do the same with EarGate. By playing back the audio recorded by the external mic, the use will be able to hear as if there were no earbuds obstructing their ear canal. Notably, this does not affect our system as we can easily filter it out (leveraging the difference in frequency), like we did for music. In addition, as we discussSection 4.6.6), we believe the power consumption and latency can be further reduced when specialized audio chips (e.g., Apple H1 Chip) are used and more

advanced engineering work (e.g., dedicated PCB design) is implemented.

5.8 Conclusion

In this chapter, we presented EarGate, an earable identification system based on user gait. Exploiting the occlusion effect, EarGate enables detection of human acoustic gait from an in-ear facing microphone. Experimenting with 31 subjects, we demonstrated that EarGate achieves robust and acceptable identification performance (up to 97.26% BAC, with low FAR and FRR of 3.23% and 2.25% respectively) under various practical conditions. Moreover, EarGate will not affect the general functionality of earbuds and is robust to high-frequency noises like music playback and human speech. We envision that EarGate can be an effective and robust way of identifying (standalone or companion with smartphone) users using earables. Particularly, its unobtrusiveness makes it an appealing replacement for FaceID for users wearing face masks during the COVID-19 pandemic.

Having shown the power of in-ear microphones for motion sensing and identification, the following chapter explores a new dimension of personal-scale sensing, investigating the potential of in-ear photoplethysmography for earable-based cardiac vital sign monitoring. To this end, Chapter 6 researches how to enable in-ear photoplethysmography (PPG), looking for the best location around the ear where to sense PPG, as well as the extent at which it in-ear PPG is robust against motion artifacts.

'Darkness took me. And I strayed out of thought and time.
Stars wheeled overhead, and every day was as long as the life age of the earth.
But it was not the end. I felt light in me again.'

Gandalf

Chapter 6

In-ear Photoplethysmography Sensing

6.1 Introduction

After exploring the new sensing avenues opened by in-ear microphones in Chapter 4 and Chapter 5, we sought to investigate the capabilities of earables for personal-scale health and wellbeing applications. To this end, in this chapter, we report an exploration of in-ear photoplethysmography (PPG) towards aiding the design of a sensory earable for vital signs monitoring. Tracking changes in vital signs, such as cardiovascular functions, heart rate, oxygen saturation, and blood pressure, through photoplethysmography (PPG) is common across wearables like smartwatches [166]. As we discussed in Section 2.3.2, photoplethysmography, better know by its acronyn as PPG, is an optical technique used to infer blood volumetric changes in the peripheral circulation. PPG is indeed a remarkable signal, which not only carries a wealth of clinical information (such as heart rate, heart rate variability, blood oxygen saturation, respiration rate, blood pressure, and artery characteristics [167–169]), but can also be used for non-medical applications such as authentication [170] and drowsiness detection [171]. Our decision to study in-ear PPG draws upon a set of technical, biological and usability-oriented observations.

Technically, PPG is relatively straightforward to implement and mechanically integrate into an earable, requiring only LEDs and photodiodes. Next, unlike more complex modalities like electrooculography (EOG) or electroencephalogram (EEG), PPG's output signal is

easy to interpret. Finally, as we mentioned above, PPG signals can derive various vital signs, including heart rate, heart rate variability, blood oxygen saturation, and respiration rate, to name a few. Biologically, several blood vessels surround the human ear, and some directly connect to main arteries (i.e., the carotid artery). This unique property is critical and very beneficial for the goodness of an optical-based PPG sensor that measures blood volume changes. In addition, the head is generally less susceptible to motion artifacts due to the musculoskeletal system's natural vibration damping. Usability-wise, modern earables are lightweight, ergonomically comfortable, and non-invasive. These properties are imperative to ensure continuous and longitudinal usage of earables, and PPG, due to its mechanical simplicity, is the most appropriate sensor for seamless integration to existing forms.

The accuracy of PPG measurements has been a subject of extensive research in recent studies [172], as PPG is a common feature in various wrist-worn consumer-grade wearables. In medical settings, prior research studied PPG on the tip of fingers, forehead, or ear lobes. These efforts mainly quantified PPG signals' inaccuracies stemmed from diverse skin types and motion artifacts [173]. Key insights of these studies suggest PPG signal is relatively accurate across skin tones, while its accuracy drops up to 30% in the presence of motion artifacts. There have been several attempts to study in-ear PPG; however, these explorations looked primarily at single sensor placement, e.g., from the ear canal [16] and from behind the ear [18]. Moreover, these mostly explored PPG sensing in or around the ear focusing on specific applications. To date, no studies have systematically validated in-ear PPG under various motion conditions across a range of different positions. To this end, we begin our investigation of in-ear PPG sensing by providing evidence towards answering *1) where is the optimal placement of an in-ear PPG sensor? and 2) what is the impact of motion artifacts at different placements?*

To address our first question, we revisit established literature on ear anatomy to identify three plausible locations in and around the ear as sensor placement alternatives, namely, behind the auricle or pinna (referred to as behind-the-ear, BTE); the concha (in-the-ear, ITE); and the first part of the ear canal (in-the-canal, ITC). We then collect PPG signals across these locations from a first cohort of 12 individuals to generate four vital signs - heart rate and heart rate variability, blood oxygen saturation, and respiration rate. We quantify in-ear PPG's quality in inferring these vital signs by comparing the extracted signals with ground truth collected with medical-grade devices. Our analysis shows that ITC represents the best location with the least error variation. We then sought to answer

our second initial question by analyzing the goodness of the in-ear PPG signal under the presence of motion artifacts (speaking, walking, and running). Our results suggest that ITC presents the smallest errors and the least inter-subject variability across the motions under examination. However, it is nonetheless heavily impacted by motion artifacts (errors up to 29.84%, 24.09%, 3.28%, and 30.80% respectively for HR, HRV, SpO_2, and RR) due to the signal crossover as reported in past studies, albeit for different body placements.

Ultimately, our findings (published in [4]) confirm that PPG sensing is particularly challenging in presence of either ambient light or motion. While the former can be mitigated by ambient light rejection modules (often already implemented in-hardware), to date, there still is no unanimously agreed technique to mitigate the latter without a considerable loss of information. Previous work considered only motion artifacts (MA) deriving from all-body movements, like walking or running [17, 172, 174, 175]. However, the head and face consist of an intricate mesh of muscles and blood vessels which contract and relax with each of their movements. This has the potential to induce unwanted noise and motion artifacts in the PPG signals recorded from the ear. The interaction between these motions and the signals recorded from in-ear PPG sensors is completely unexplored. Hence, after assessing the optimal location for in-ear PPG sensing, this chapter aims at *providing the research community with a novel, multi-modal, dataset, which, for the first time, will allow to study the impact of body and head/face movements on both the morphology of the PPG wave captured at the ear, as well as on the vital signs estimation.* To accurately collect in-ear PPG data, coupled with a 6 degrees-of-freedom (DoF) motion signature, we leveraged the knowledge acquired through our preliminary investigation to prototype and build a flexible research platform for in-the-ear data collection. The platform is centered around a novel ear tip design which includes a 3-channel PPG (green, red, infrared) and a 6-axis (accelerometer, gyroscope) motion sensor (IMU) co-located on the same ear-tip. This allows the simultaneous collection of spatially distant (i.e., one tip in the left and one in the right ear) PPG data at multiple wavelengths and the corresponding motion signature, for a total of 18 data streams. Inspired by the Facial Action Coding Systems (FACS) [103], we then consider a set of potential sources of motion artifact (MA) caused by natural facial and head movements. Specifically, we gather data of head movements (nodding, shaking, tilting), eyes movements (vertical eyes movements, horizontal eyes movements, brow raiser, brow lowerer, right eye wink, left eye wink), and mouth movements (lip puller, chin raiser, mouth stretch, speaking, chewing). Given their relevance to earables and daily life, we also collect motion and PPG data under activities, of different intensity, which entail the

movement of the entire body (walking and running). Together with in-ear PPG and IMU data we collect several vital signs including, heart rate, heart rate variability, breathing rate and raw ECG, from a medical-grade chest device.

With ∼17 hours of data from a second cohort of 30 participants of mixed gender and ethnicity (mean age: 28.9 years, standard deviation: 6.11 years), our dataset empowers the research community to analyze the morphological characteristics of in-ear PPG signals with respect to motion, device positioning (left ear, right ear), as well as a set of configuration parameters and their corresponding data quality/power consumption trade-off. We envision such dataset could open the door to innovative filtering techniques to mitigate, and eventually eliminate, the impact of MA on in-ear PPG.

We run a set of preliminary analysis on the data, considering both handcrafted features, as well as a DNN (Deep Neural Network) approach. Ultimately, we observe statistically significant morphological differences in the PPG signal across different types of motions when compared to a situation where there is no motion. We also discuss a 3-classes classification task and show how full-body motions and head/face motions can be discriminated from a still baseline (and among themselves). These results represents the first step towards the detection of corrupted PPG segments and show the importance of studying how head/face movements impact PPG signals in the ear.

To the best of our knowledge this is the first in-ear PPG dataset that covers a wide range of full-body and head/facial motion artifacts. Being able to study the signal quality and motion artifacts under such circumstances will serve as reference for future research in the field, acting as a stepping stone to fully enable PPG-equipped earables.

In summary, the contributions of this chapter include:

- An investigation of the best location around the ear to sense PPG.
- A novel ear tip design with co-located multi-channel PPG and motion sensors for in-the-canal data collection –informed by our findings.
- A first of its kind multi-modal PPG dataset under a wide range of motion artifacts induced from full-body movements and, more importantly, head/facial movements. For the first time, the PPG (and IMU) signal is collected simultaneously from spatially distant locations (left and right ear).
- A preliminary analysis of how full-body and head/facial movements impact in-ear

PPG.

6.2 Ear anatomy and sensor placement

As we have discussed in Section 2.3.1, the ear is the human sensorium dedicated to hearing. It is composed of three regions: outer ear, middle ear (i.e., where sound waves are modulated), and inner ear (i.e., where the modulated sound waves are transmitted to the brain). In this dissertation, we focus our attention on the outer ear, easily accessible for the placement of ear-worn devices.

To identify plausible placements for a PPG sensor in the outer ear, two aspects shape our design considerations: quality of signal and device comfort. From a signal quality perspective, a PPG sensor detects changes in blood volume flow, and as such, it is essential to place it in an area well supplied by blood vessels. The main components of the outer ear are the **auricle** (or pinna), the visible part of the ear; the **concha**, the depression in the auricle leading to the ear canal orifice; and the **ear canal** itself (Figure 6.2.1). All these three areas are supplied with blood by branches of the external carotid artery, a major artery of the head and neck. Several capillaries called **perforating branches** (Figure 6.2.1) supply blood to the auricle (on both sides, behind, as well as, in the proximity of the ear canal orifice) and the concha. Additionally, behind the auricle flows the **superior auricular artery**. Another major blood vessel in the very proximity of the ear, particularly close to the ear canal, is the **superficial temporal artery** [176]. The high concentration of blood vessels in the outer ear naturally poses multiple alternatives for a PPG sensor placement.

To minimize our search space, we turn our attention to ergonomic comfort. Ideally, the earable should enable continuous monitoring in natural situations while ensuring ease of use and comfort for prolonged use. Borrowing notions from the rich literature on earables, including hearing aid ergonomics [177], we identify three plausible alternatives: 1) behind the auricle, 2) in the concha, or 3) inside the ear canal.

Hence, we annotate Figure 6.2.1 highlighting in green the three vantage points defined as: *behind-the-ear* (BTE) leveraging the superior auricular artery; *in-the-ear* (ITE) for the perforating branches and *in-the-canal* (ITC) for the superficial temporal artery.

We exclude other areas of the ear such as the lobe or behind it (i.e., the skin of the neck), even if supplied by the same main artery, because placing an earable on these areas is unnatural and potentially uncomfortable, especially for prolonged use.

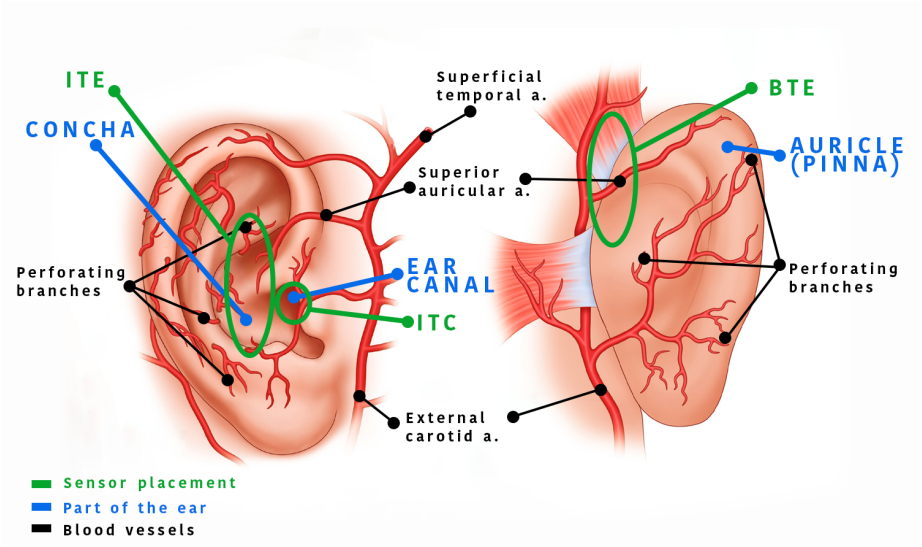


Fig. 6.2.1: Anatomy of the human ear with annotated the chosen sensor placements.

6.3 PPG and vital signs extraction

Photoplethysmogram (PPG) is an optical measurement technique consisting of an LED and a photodetector used to derive a heart rate signal by detecting blood volume changes in the region under examination (Section 2.3.2). There are two types of PPG sensors: one measuring light transmission through the tissues (used on the extremities as fingertips and ear lobes) and the other measuring the light reflected by the tissues (as in smartwatches and wrist-worn fitness trackers). For the scope of this dissertation, we focus on light reflectance PPG sensors given their mechanical simplicity in being integrated in an earable form, as they do not require the placement of electronic components on two opposite sides of the skin.

Concretely, light reflectance PPG sensors measure how light intensity varies whenever there is a blood volume change. Bones, muscles, and tissues absorb light at a constant rate. Therefore, arterial blood volume variations directly map into a decrease in the light intensity (i.e., the voltage of the signal) measured at the photodetector [78]. For a more in-depth explanation of how photoplethysmography works, refer to Section 2.3.2.

6.3.1 Vital signs estimation from PPG

PPG signals are commonly used in medical and free-living settings to extract vital signs, i.e., reliable bio-markers to indicate health outcomes (Section 2.3.3). The most common

types of bio-markers extracted from PPG can be grouped as those related to **cardiovascular activity** (e.g., heart rate, heart rate variability, blood pressure, hyper and hypovolemia), **respiration** (heart rate signals are modulated by respiratory activity [178]), and **blood oxygen saturation** (SpO_2). These bio-markers constitute an enormous wealth of information, allowing medical practitioners, for instance, to infer the fitness of a patient, the presence (or absence) of respiration as well as cardiovascular conditions, and their mental health (stress, cognitive load, and mental fatigue) [179]. In this chapter, we study three vital signs and the respective processing pipelines we have developed to extract them from raw PPG signals collected from different locations in, and, around the ear.

Heart rate (HR) and heart rate variability (HRV)

HR measures the heart contractions that push blood through the arteries. On the other hand, HRV is the variation in the time interval between consecutive heartbeats (contractions). In medical settings, HR is usually measured either with a finger pulse oximeter or with ECG and ranges between 60 and 100 bpm in healthy subjects. Contrary, HRV is extracted from the R-peaks detected in an ECG signal and is often used when referring to a person's fitness and mental health. High inter-beat interval (ibi) denotes high fitness (700-1000 ms), while lower HRV often indicates a higher stress level.

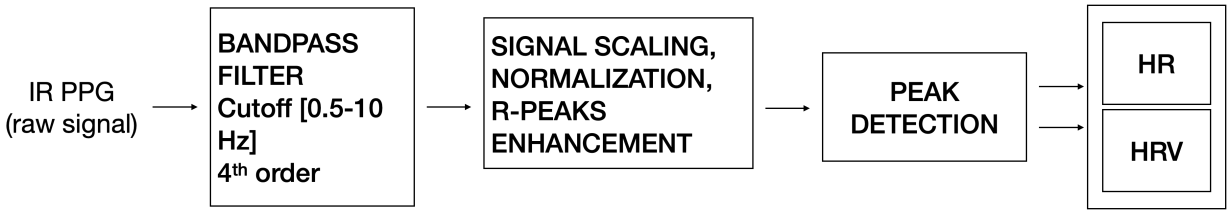
Extraction pipeline: As depicted in Figure 6.3.1a, we start by processing both the raw infrared PPG and the raw ECG (our ground truth signal) with a band-pass and a notch filter, respectively. This step preserves the information content at typical heart rate frequencies and discards higher frequencies. The filtered signals are then scaled and normalized, and the R-peaks amplitude is enhanced. From the resulting signals we detect all the peaks corresponding to heart beats and compute the mean HR and HRV.

Reporting metrics: We use *mean bpm* (beats per minute) for HR, as the average HR during a session and *mean inter-beat-interval* (IBI) for HRV, as the average time between successive heartbeats (taken across a session).

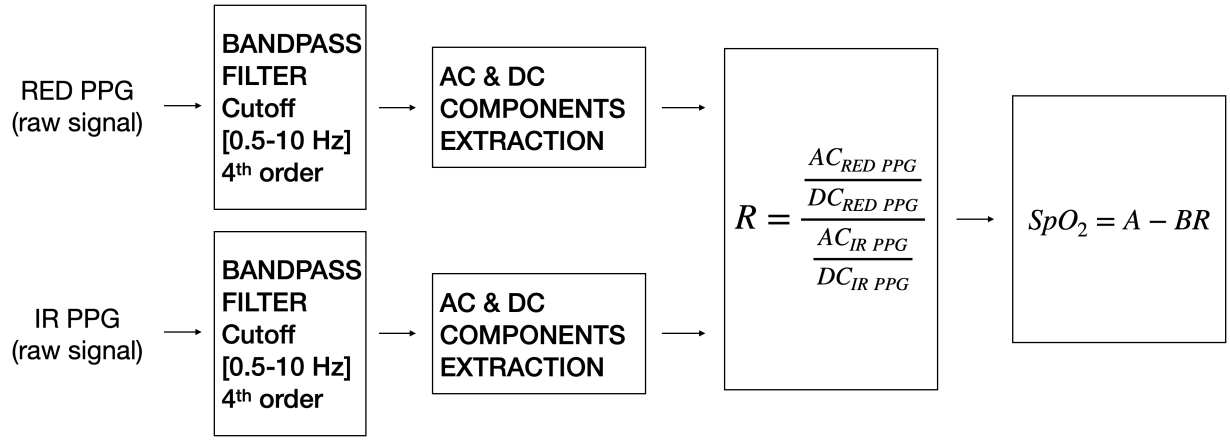
SpO_2

Blood Oxygen Saturation indicates the percentage of oxygen-saturated hemoglobin in the blood. In healthy subjects, it usually is between 95% and 100%. SpO_2 is commonly measured with finger pulse oximeters.

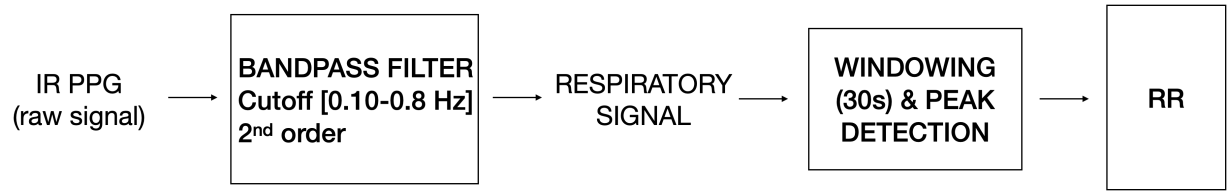
Extraction pipeline: To extract SpO_2 from PPG, we first filter the PPG readings with



(a) Pipeline for resting heart rate and heart rate variability.



(b) Pipeline for blood oxygen saturation.



(c) Pipeline for respiration rate.

Fig. 6.3.1: Processing pipelines used to extract vital signs.

the same band-pass filter used to pre-process the data before extracting HR and HRV. We then sought to isolate the AC (pulsatile) and DC (non-pulsatile) components [180] (for a more in-depth explanation of the AC and DC component of PPG, please refer to Section 2.3.2). We adopt the window minimum approach [78] and compute R, the ratio of the ratios of the AC and DC components of the red and infrared PPG signals (Figure 6.3.1b):

$$R = \frac{\frac{AC_{RED\ PPG}}{DC_{RED\ PPG}}}{\frac{AC_{IR\ PPG}}{DC_{IR\ PPG}}}. \tag{6.1}$$

We then calculate the SpO_2 by applying the equation:

$$SpO_2 = A - BR, \tag{6.2}$$

where A and B are calibration coefficients [181] we empirically set as $A = 120$ and $B = 24$.

Reporting metrics: We use *mean spo2* for SpO_2, as the average percentage of oxygen saturation during a session.

Respiration Rate (RR)

Respiration rate is the number of breaths a person takes per minute. In adult subjects, it usually varies between 10 and 20 breaths per minute.

Extraction pipeline: We extract a respiration rate (RR) following a frequency-based approach [178] (Figure 6.3.1c). First, we extract a respiratory signal from the raw infrared PPG signal. Considering a normal breathing range is $10 - -20$ breaths per minute [182], the non-respiratory frequencies are removed with a band-pass filter. Once left with the respiratory signal, we segment it in 30 seconds long windows, find the peaks and compute the breath per minute value for every window.

Reporting metrics: We use *mean rr* for a respiration rate, as the average number of breaths per minute within a session.

6.4 Experimental setup

To identify the optimal positioning for an in-earable PPG sensor, we run a first data collection campaign featuring twelve subjects.

6.4.1 Study population

Twelve individuals (2 females, 24-40 years of age, mean = 30.4) were recruited to participate in the study. None of the participants had any heart or respiratory condition and were in good health. All participants were briefed about the study and voluntarily consented to take part in it (no compensation was offered). The study received IRB approval before its beginning.

6.4.2 Devices

To collect PPG signals from the ear, we use a Cosinuss Two[1] which features an infra-red and a red PPG sensor on the same ear-tip. The raw data for both wavelengths is provided by the device with a sampling rate of $100Hz$. We modify the Cosinuss' ear-tip to provide better adhesion with the skin for the BTE and ITE conditions and use surgical tape for attachment. For the ITC condition, we mount the ear-tip on an existing earable (i.e., eSense [2]) to mimic an off-the-shelf earbud.

For ground truth purposes, we use a portable ECG (heart's electrical activity) chest band (Zephyr Bioharness 3.0[2]) and a medical grade pulse oximeter to be worn on the finger (Masimo Health MightySat-Rx[3]). We use the ECG signal to derive the ground truth for HR and HRV while RR and SpO_2 are provided directly by the Zephyr and the pulse oximeter, respectively.

The participants wore the portable ECG band and the pulse oximeter on their chest and index finger for the whole experiment. Synchronized data from our PPG sensor and the ground truth devices has been collected for the entire duration of the study. The devices used during the data collection are depicted in Figure 6.4.1.

6.4.3 Data collection protocol

Given our first objective is to assess the best positioning for a PPG sensor around the ear, and the impact of common motions, we collect PPG data from the three locations identified earlier (ITC, ITE and BTE), first in a resting condition (without any movement) and then under three motion conditions of increasing intensity: speaking, walking and running. Walking and running represent typical full-body sources of motion-induced noise for optical-based heart rate monitors, as shown by previous research [173]. Contrary, the speaking condition is instead unique to earables. The complex muscle movements while talking are likely to cause significant deformations of the tissues around the ear and in the ear canal, with potential negative effect on the recorded PPG signal.

For the resting and speaking conditions we followed the wearable device validation guideline stipulated by the Consumer Technology Association [183] and measured PPG while seated in the upright position. Notably, during the resting condition, the participants were asked

[1] https://www.cosinuss.com/
[2] https://www.zephyranywhere.com/
[3] https://www.masimo.com/

to breath normally without moving, while during the speaking condition they were asked to read aloud an article provided by the investigators. For the motion conditions the participants were asked to walk and run at a comfortable pace around the room ensuring the intensity was higher for running compared to walking. We measured an average acceleration of $0.17g$ ($\sigma = 0.05g$) and $1.04g$ ($\sigma = 0.11g$) while walking and running, respectively.

For each motion condition we recorded 5 minutes of data at the three sensor placements we identified. The length of the sessions was chosen so that they were long enough to observe changes in the characteristics of good-quality vital signs, without being too tedious for the data collection volunteers.

6.4.4 Data analysis

The PPG data is processed according to the pipelines described in Section 6.3.1 in order to extract the vitals we are considering. The PPG-derived bio-markers are then compared to the values obtained from the ground truth devices. We compute the relative error as the difference between the ground truth value and the PPG-derived value for each bio-marker. We chose to look at the relative error to be able to observe whether the data present trends (e.g., PPG consistently over/under-estimating the HR).

6.5 Benchmarks

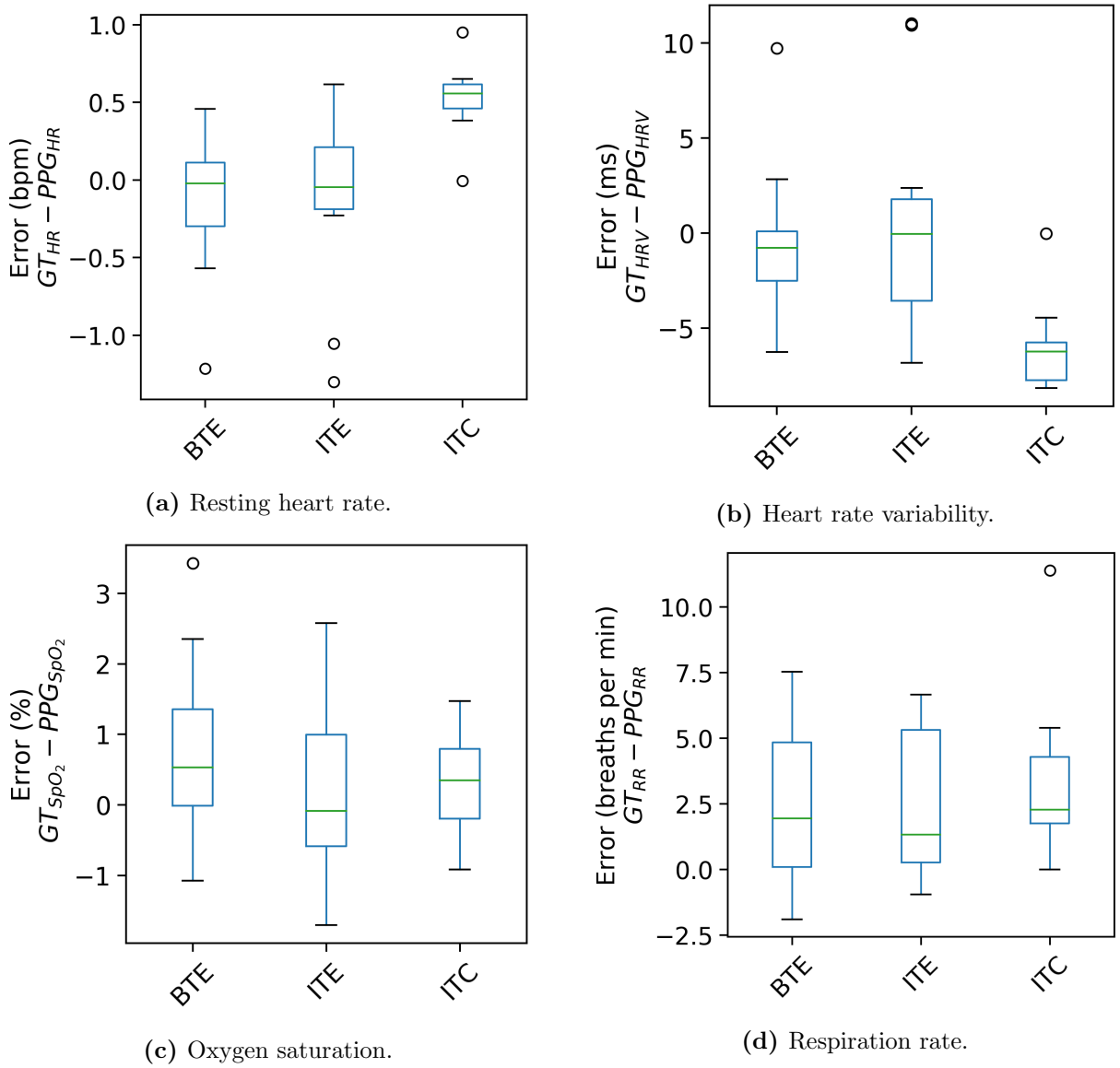
We begin by reporting our observations concerning placement variabilities and then reflect on the impact of motion artifacts on the accuracy of PPG-derived vital signs.

6.5.1 PPG placement

Figure 6.4.2 illustrates the inaccuracies we observed in the PPG-derived vital signs in comparison with ground truth. We notice that all three positions show comparable performance for different vital signs concerning accuracy. However, one-way repeated measures ANOVA tests performed on the three locations show significant differences on the mean error of each vital sign with test statistic F ranging from 11.0 for RR to 141.2 for HR with p-value < 0.05.

Looking closely at Figure 6.4.2, the ITC position shows the least variability (smaller interquartile range). Overall, we can observe that in-ear PPG shows acceptable performance

Fig. 6.4.1: On the left a participant wearing the ITC PPG; and on the right the devices used in the data collection (clockwise: Zephyr Bioharness 3.0, Masimo Health MightySat-Rx, Cosinuss two tip on eSense for ITC, modified Cosinuss two tip used for ITE and BTE).



(a) Resting heart rate.

(b) Heart rate variability.

(c) Oxygen saturation.

(d) Respiration rate.

Fig. 6.4.2: Impact of different ear-placements on the goodness of PPG-extracted bio-markers in a resting condition.

to extract reliable HR, HRV, and SpO_2 with errors ranging across ± 0.5 bpm, $\pm 5ms$, and $\pm 2\%$, respectively. For instance, errors of a few ms when estimating IBI for HRV are acceptable given that they account for about 1% of the actual IBI of healthy subjects at rest — usually being between 700 to $1000ms$, depending on age and medical conditions. Similarly, for SpO_2 data, medical-grade pulse-oximeters usually have a 2% accuracy, which is in line with what we observe for the ITC location.

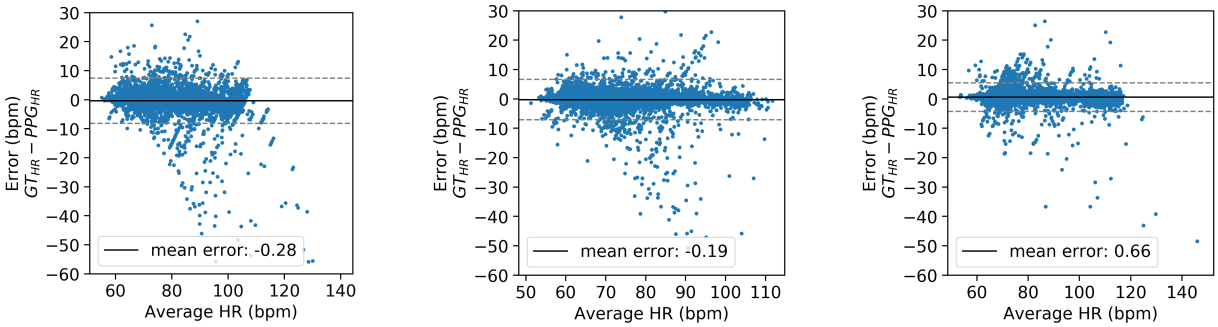
However, for the respiration rate, we notice a relatively larger error across the three locations, with a mean error around 3 breaths per minute and large variation. This represents a 30% to 15% error on average if we consider typical breathing rates of 10 to 20 breaths per minute at rest [182].

To further analyze the data we employ Bland-Altman (BA) plots, which depict the difference between ground truth bio-marker and PPG-extracted bio-marker on the y-axis ($ECG_{HR} - PPG_{HR}$) and their average on the x-axis ($(ECG_{HR} + PPG_{HR})/2$). Instead of using the mean value of each bio-marker during the entire session, for this analysis, we use the bio-marker values computed with a 1 second granularity. BA plots are typically employed to study the agreement between two measurement methods and allow us to investigate how the two measurements of a bio-marker differ (i.e., y-axis) at different magnitudes of the bio-marker (i.e., x-axis).

Figure 6.5.1 depicts the BA plots for the HR in three locations. While in the interest of space we only report the the BA plots for HR, we observed similar patterns for the other bio-markers. The plot shows how the mean error of HR is consistently small across all three locations. Diversely from ITE and BTE, where the data-points are sparse and the confidence intervals large (± 1.96 of the standard deviation of the error), for the ITC location the majority of data-points fall within the confidence intervals around the mean error. The BTE and ITE locations show larger errors as the HR increases (i.e., in the range $70 - -100 bpm$), with several data points showing negative errors. This implies that HR estimated from PPG tends to overestimate the actual value in this range. This result suggests an overall agreement between the PPG and the ground truth readings at rest, confirming our previous observation of less error variability for the ITC location.

Key takeaways: Our analysis suggests that, in a resting condition, while there is marginal difference in the median error across the three locations, the in-the-canal position represents a good placement option for in-ear PPG, as it shows consistently the least variation. We argue that this is enabled by the in-the-canal location and the form factor of the corresponding device, i.e., an ear-tip. We could observe that, even when participants wear the ear-tip naturally without special care from the investigators, it automatically ensures good sensor-to-skin contact and improved shielding from ambient light. Devices built for the other locations instead currently provide less tight contact to the skin, thereby resulting in higher variation across wearing events.

Our results also highlight that the PPG-extracted respiration rate has a large error margin.



(a) Resting heart rate BTE. **(b)** Resting heart rate ITE. **(c)** Resting heart rate ITC.

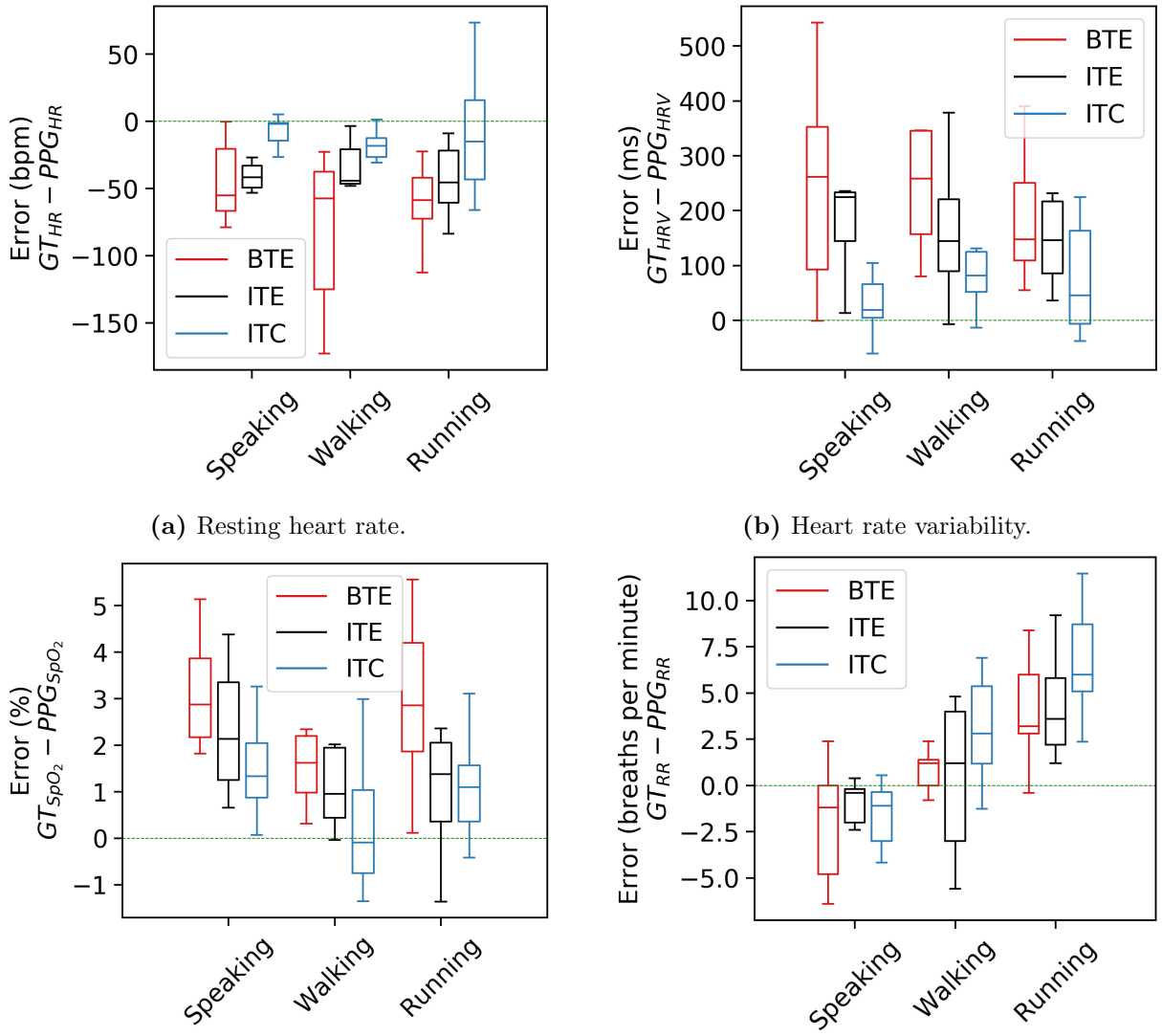
Fig. 6.5.1: Difference between ground truth and PPG heart rate plotted against the mean of the two measurements (Bland-Altman plot). In the interest of space, only data for HR is shown. The solid black line represents the mean error, and the dashed gray lines the 1.96 SD boundaries (95% limits of agreement).

We consider this error could be mitigated with more sophisticated processing techniques or by leveraging other sensing modalities (e.g. IMU-based breathing detection [184]).

6.5.2 Impact of motion artifacts

The form, functions, and applications of an earable naturally imply suspected measurement errors of in-ear PPG, which are typically caused by motion artifacts, i.e., sensor displacement due to head and body movement. Consequently, we sought to understand how the PPG sensor suffers from these artifacts. Specifically, we investigate the impact of three motion activities that are commonly performed while wearing earables - speaking (micro motion), walking (mild motion) and running (intense motion).

Figure 6.5.2 illustrates the box plots of the errors between ground truth and PPG-extracted vital signs for different artifacts, across the three locations. At a glance we notice that the median error for the ITC location tends to be lower than the other locations for the three motion conditions. This is further confirmed when comparing the mean absolute errors of the various bio-markers. For instance, HR extracted from the ITC location shows a 12.52%, 27.14%, and 29.84% error respectively for speaking, walking, and running. These figures are substantially lower than those of the other two locations: 62.60%, 105.98%, 56.62% and 58.29%, 58.66%, 44.00% respectively for BTE and ITE (similar differences can be observed for HRV and SpO_2). An exception to this is the respiration rate for which ITC shows similar or slightly higher median errors (with BTE showing a lower mean absolute error). For the respiration rate we also notice a clear trend where the error increases with



(a) Resting heart rate.

(b) Heart rate variability.

(c) Blood oxygen saturation.

(d) Respiration rate.

Fig. 6.5.2: Impact of different external artifacts (i.e. speaking, light motion, and intense motion) on the goodness of BTE, ITE, and ITC PPG-extracted bio-markers.

the increase of motion intensity.

ITE and BTE instead, show high median errors and large variations across all three motion conditions, suggesting that they are less suitable locations for in-ear vital signs measurement, even in presence of moderate motion (e.g., speaking and walking). This could be a consequence of the difficulty in attaching the sensor firmly in these locations and the more pronounced deformation of the skin and tissues during motion activities which results in corrupted PPG signals.

Notably, the estimation of SpO_2 shows similar median error across the three motions with relatively limited variability for all three locations. We hypothesize that this relative consistency in SpO_2 measurements is due to the fact that participants' oxygen saturation did not change significantly during the data collection period, hence only a limited range of values could be recorded (e.g., between 96% and 100%). We leave the investigation of other SpO_2 ranges through the use of specialized equipment to future work.

Key takeaways: Our results suggest that in-ear PPG, despite being located on a more stable part of the body (i.e., the head compared to the wrist), is still significantly affected by motion, with absolute errors for ITC up to 29.84%, 24.09%, 3.28%, and 30.80%, respectively for HR, HRV, SpO_2, and RR. Based on our population, ITC seems to provide the lowest errors across the motion conditions, with ITE and BTE showing large errors even with moderate motion. This is probably due to a poorer adhesion with the skin for the ITE and BTE locations compared to ITC which benefits from the enclosed space of the ear canal. These observations call for attention to sophisticated signal processing pipelines mitigating signal cross-over issues, ideally guided by additional sensing modalities (e.g., co-located IMU) and advanced industrial design to ensure stability and good skin contact, especially for the BTE and ITE placements.

6.6 Outlook: in-the-canal PPG

Our results distilled that the in-the-canal placement is a plausible design choice striking the best trade-offs. However, we need to cater for a careful mechanical design for robust data acquisition and signal processing to mitigate motion artifacts. We reflect on these two aspects in this concluding section.

6.6.1 Form factor and ear-tip design

From a mechanical construction perspective, the in-the-canal placement of a PPG sensor uncovers interesting design challenges. On the one hand, it liberates us from a specific form design, as ear-tip or earmold is a default feature of any earable, be it a lifestyle earbud or a medical-grade hearing aid. Besides, in-the-canal placement ensures good skin contact and natural shielding from ambient light –both imperative to achieve a robust sensing performance.

However, on the other hand, designing an ear-tip with an integrated PPG sensor is exceptionally challenging. Firstly, the ear-tip has to guarantee good sealing for robust data acquisition. Secondly, the ear-tip must have a stable and steady fit across users to ensure minimal sensor displacement, minimizing motion artifacts. Finally, the ear-tip must be comfortable for prolonged use. However, this well sealed, stable, and comfortable fit must not come at the expense of additional occlusion effects, beyond what is already mitigated today with state-of-the-art audio engineering, e.g., pass-through audio. Besides, given every single human ear is different in size and shape, it is incredibly hard to design a cost-effective and universal ear-tip with integrated PPG using existing materials, e.g., rubber, silicone, or foam. Consequently, we call attention to material engineers and industrial designers to carefully consider the facets mentioned above in developing next-generation sensory ear-tips.

6.6.2 Signal processing pipeline

The accuracy of wrist-worn PPG under full-body motion artifacts has been extensively studied in recent literature, with results reporting up to 30% of error rate in extracting vital signs [173]. Even though the head is the stationary part of the human body, we still noticed a similar error margin with in-ear PPG. This signal cross-over effect is caused by the fact that a PPG sensor mistakenly detects the cardiovascular cycle in the presence of a periodic signal caused by repetitive motion, for instance, walking or running. Contemporary mitigation strategies consider wavelet transform, independent component analysis (ICA), template matching, and adaptive filtering techniques. While these techniques are certainly useful, we advocate for adaptive and context-aware signal processing strategies. In particular, we suggest using motion sensing to remove motion artifacts. Given that the PPG signal is correlated with motion intensity and that an inertial measurement unit (IMU) is a common feature in modern earable, we can leverage motion-awareness to trigger

PPG operations, thus avoiding corrupted PPG segments while saving energy. However, as of today, there is very limited availability of datasets which spatially couple PPG and IMU.

We acknowledge that the relatively small size of the study population and the preliminary investigation of SpO_2 ranges can be regarded as limitations of this preliminary study. Nonetheless, having characterized in-ear PPG, under various full-body motion conditions across a range of different positions and gave us the intuitions we needed to proceed with the development of a novel sensory earable platform. Additionally, our investigation led us to the realization that the literature completely lacks an analysis of how more subtle motion artifacts impact PPG. Notably, this might be a game-stopper for in-ear PPG sensing. Hence, leveraging our custom-made earable (featuring in-the-canal PPG), we collect a one-of-its-kind dataset to study how head movements and facial expressions affect the quality of in-ear PPG. In the remainder of this chapter we first describe the panorama of available in-ear PPG datasets, we then present our PPG-equipped sensory earable and the rigorous methodology we followed to collect EarSet: a multi-modal dataset for studying the impact of head and facial movements on in-ear PPG signals.

6.7 Existing PPG motion artifacts datasets

In the first half of this chapter, we have seen the importance of studying the effect of motion artifacts on the PPG signal.

To this end, researchers often rely on data collected under controlled conditions which induce motion artifacts on the PPG signals. Table 6.7.1 summarizes few of the well-known and publicly available datasets for PPG sensor data collected at different locations.

Wrist PPG motion artifacts: Wrist is the most popular location for studying PPG signals owing to the popularity of wrist-worn fitness trackers among the consumers. IEEE PPG dataset [188] was one of the first publicly available PPG datasets that studies how motion artifacts induce noise in the resulting wrist PPG signals and how different signal processing techniques can help in removing motion artifacts from full-body motions. PPGDaLiA [186], [185] and [187] are datasets that aim to improve activity monitoring by identifying daily activities like walking, running, sitting, cycling etc. from wrist PPG signals. [189] released a PPG dataset consisting of wrist PPG signals collected during walking and running to help promote motion artifact removal techniques for PPG signals.

Table 6.7.1: PPG datasets publicly available for motion artifact studies

Datasets	PPG sensor location	Motion being studied	Additional sensor data	Number of participants
Activity monitoring [185]	Wrist	Squat exercises, stepper exercises, and resting	3-channel PPG 3-axis accelerometer	7
PPG DaLiA [186]	Wrist	Daily life activities like sitting, walking, cycling, driving, working etc.	3-channel PPG Electrocardiogram (ECG) Electrodermal activity (EDA) 3-axis accelerometer Respiration rate Body temperature	15
Effect of exercises on PPG signals [187]	Wrist	Walking, running and biking	3-channel PPG Chest ECG 3-axis accelerometer 3-axis low noise accelerometer 3-axis gyroscope	23
Motion artifact removal in PPG signals (IEEE signal processing cup) [188]	Wrist	Random physical exercises without labels	3-channel PPG Chest ECG 3-axis accelerometer	12 (training dataset) 10 (test dataset)
Motion artifact cancellation [189]	Wrist	Walking and running	3-channel PPG Chest ECG 3-axis accelerometer 3-axis gyroscope	24
WESAD (Stress detection) [190]	Wrist and Chest	Intense physical activity and mental exercises to induce stress	Wrist PPG Wrist accelerometer Wrist electrodermal activity (EDA) Body temperarture Chest ECG Chest accelerometer Chest EMG Chest Respiration	17
BIDMC dataset [191]	Finger	No exercise involved	Finger PPG Pneumography (Respiration)	53
FatigueSet [192]	In-Ear	Running on a treadmill to induce physical fatigue	In-Ear PPG In-Ear IMU sensor Chest ECG Chest Respiration sensor Wrist PPG Wrist EDA Wrist IMU Body temperature sensor	12
Motion tolerant heart rate and Blood pressure monitoring [18]	Outside ear	Exercising on a bike	Ear PPG Ear ECG Ambulatory blood pressure monitor	14
EarSet	**Stereo In-ear**	**16 different facial and head motions**	**In-Ear PPG (Both Left and right)** **In-Ear IMU (Both Left and right)** **Chest band ECG** **Chest band Respiration sensor**	**30**

In addition, WESAD [190] and BIDMC [191] datasets consists of PPG signals collected for the purpose of stress detection and respiration rate estimation respectively. It is evident that the there are sufficient data-sets publicly available to study the effect of motion artifacts on wrist PPG signals. However, the same cannot be applicable for PPG signals collected from the ear since earables have been explored only during recent years unlike wrist-worn fitness trackers.

EarSet: Very few openly available datasets include PPG data from the ear [18,192]. However, there are no publicly available datasets that explore the effect of facial expressions and head movements on earables. Recently, [18] proposed a solution for how motion artifacts can be removed for accurate heart rate and blood pressure estimation with PPG sensors placed on the ear lobes. However, they only study the effect of body motion artifacts on the acquired PPG signals. Hence, there is a strong need for an open-source dataset studying the effect of facial motions on in-ear PPG signals.

To this aim, we propose *EarSet*, a unique earable dataset that provides in-ear PPG signals as well as in-ear IMU signals acquired from both the left and right ear during 16 different facial and head movements (plus a baseline session). We firmly believe this will help to study and gain a thorough understanding of how artifacts arising from the face, mouth, eye, and head motions affect in-ear PPG signals. Further, this could in turn facilitate the development of novel filtering techniques ensuring a clean in-ear PPG signal for accurate and real-time health sensing. EarSet is the first dataset to provide in-ear PPG signals for both left and right ear in presence of face/head motions as well as full-body motions (i.e., walking and running). We also provide ground-truth ECG signals and other vital signs collected using a chest band ECG device which could be used to support the development of vitals estimation from in-ear PPG signals.

6.8 EarSet: methodology

In this section we detail the methodology and the devices followed to collect EarSet.

6.8.1 Study population

Thirty individuals (18 males, 12 females, $20-49$ years of age, mean age: 28.9 years, standard deviation: 6.11 years) were recruited and voluntarily took part in the study. Within the population of our study, none of the participants had any underlying heart or

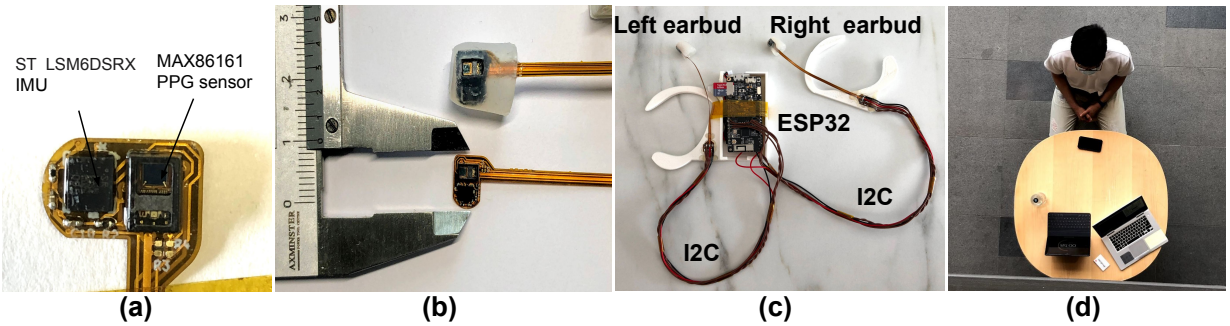


Fig. 6.8.1: (a) Flexible PCB implementation of our earbud with a MAX86161 PPG sensor and a co-located ST LSM6DSRX IMU (b) An in-ear soft earbud was realized by embedding the in-ear flexible PCB board into a transparent silicone mould. (c) Head-worn data acquisition device consisting of an ESP32 micro-controller collecting data from in-ear PPG and IMU sensors in the left and right ear. (d) A participant wearing our earbud based prototype and taking part in the data collection protocol.

respiratory condition, and were in good health at the time of the study. Before taking part to the study, we briefed all the participants who then gave their written consent by completing a consent-form. Every participant received a gift card as a compensation upon completion of the study. The study received IRB approval prior commencing.

6.8.2 Data collection devices

Given the lack of existing open-source in-ear PPG platforms, and inspired by the findings we reported earlier in this chapter, we designed a custom head-worn prototype (Figure 6.8.1(c)) to collect in-ear PPG signals with well-known and affordable hardware components. The prototype consists of an ESP32 micro-controller collecting sensor data from both the left and right ears. To facilitate PPG signal acquisition from inside the ear, we fabricated a flexible PCB board consisting of a MAXM86161 [4] PPG sensor and ST-LSM6DSRX [5] IMU as shown in Figure 6.8.1(a). The flexible PCB board is interfaced via the I2C protocol to the ESP32 micro-controller for data acquisition. The MAXM86161 is a well-known 3-channel PPG sensor (green – $520-550$nm, red – 660nm, infrared – 880nm) catered for in-ear sensing applications. The IMU continuously records 3-axis accelerometer and 3-axis gyroscope data to provide motion signals for in-ear motions occurring while making a facial expression or head movements. Both sensors are sampled at a 100Hz frequency. As shown in Figure 6.8.1(b), the flexible PCB containing the PPG sensor and

[4] https://www.maximintegrated.com/en/products/sensors/MAXM86161.html
[5] https://www.st.com/en/mems-and-sensors/lsm6dsrx.html

Table 6.8.1: PPG parameters and relative sensor current draw.

Conf.	LED Current (mA)	Pulse Width (us)	Integration Time (us)	Current Draw (mA)
1	16	21.3	14.8	1.62
2	32	21.3	14.8	1.81
3	16	123.8	117.3	2.66
4	32	123.8	117.3	3.78

the IMU was coated with soft silicone to resemble a typical ear tip to provide comfort while wearing the device, as well as remain firm within the ear during various face/head motions. We used a transparent soft silicone gel to prevent any distortions in the acquired PPG signals.

PPG signal quality is not only affected by sensor motion but also by the sensor configuration. Typically, sensors allow to change several parameters which affect the acquired signal and consequently the power consumed by the sensor. Given this trade-off, often, optimal parameters for signal quality are not the most efficient in terms of power consumption. To explore this aspect of PPG sensing, we configured our device to change the sensor parameters every 30 seconds. This way, by collecting data for 2 minutes for each motion session we could cycle through 4 different set of configurations (Table 6.8.1). In particular, the MAXM86161 allows to change three parameters: LED current which determines the brightness of the three LEDs, pulse width which is the time each LED is kept on during a measurement and the integration time which is the period during which the photodiode is active and sampling the reflected light[6]. As shown in Table 6.8.1, we have chosen 4 configurations that offer distinct power consumption profiles and should result in diverse SNR characteristics. To best of our knowledge this is the first PPG dataset that offers data collected with different sensor configurations.

On the other hand, to collect vital signs from a reliable source which is not affected by motion artifacts, as we did to determine the optimal positioning for the in-ear PPG sensor, we rely on a Zephyr Bioharness 3.0[7]. The participants wore the portable ECG band on their chest for the whole experiment.

[6]Notice that pulse width and integration time cannot be controlled individually and only 4 combinations of the two parameters are available in the sensor.

[7]https://www.zephyranywhere.com/

Table 6.8.2: Sensor data collected from each wearable device.

Sensor	**Units/Range**	**Sampling Rate**
Earable prototype (one per ear)		
Accelerometer	g {-2:+2}	100 Hz
Gyroscope	°/s {-500:+500}	100 Hz
PPG - green, infrared, and red channels	-	100 Hz
Zephyr BioHarness 3.0 chest band		
Accelerometer	bits {0-4094}	100 Hz
Breathing sensor raw output	bits {1:16777215}	25 Hz
Breathing rate	breaths per minute {4:70}	1 Hz
Breath-to-breath interval	ms	-
ECG raw waveform	bits {0:4095}	250 Hz
Heart rate	beats per minute {25:240}	1 Hz
Heart rate variability	ms {0:65534}	1 Hz
RR interval	ms {0:32767}	-
Posture	degrees from vertical {-180:180}	1 Hz

6.8.3 Data collection protocol

Since the aim of this chapter is to provide the research community with a dataset to better understand the impact of motion artifacts on in-ear PPG, we sampled PPG data under a number of different conditions.

After being briefed about the study, the participants wore our in-ear data collection device on the head placing the ear tips in the left and right ear canal (Figure 6.8.1) and the ECG device on their chest. Starting form a resting pose (participants sitting still without any motion), we progressively asked the participants to repetitively carry out individual movements. We consider two main classes of motions: head/face movements and full-body movements. By looking at the inherent nature of the motions, head/face movements can be further categorized into *one-shot* and continuous movements. The selection process for the one-shot motion artifacts was informed by both anatomy principles [103] and previous work [104, 105]. In building our dataset, we look at AUs that entail the movement of the head, the eyes (and the adjacent muscles), and the mouth. Specifically, we selected: (1) nod; (2) shake; and (3) tilt as head movements. The eyes movements we considered were respectively: (4) vertical eyes movements; (5) horizontal eyes movements; (6) brow raiser; (7) brow lowerer; (8) right eye wink; and (9) left eye wink. Finally, we investigated: (10) lip puller; (11) chin raiser; and (12) mouth stretch as mouth movements. These are all motions that are not normally performed continuously, and that are often performed in normal social interactions as well as in the form of psychosomatic tics. Notably, before performing each and every motion, the investigator demoed the gesture to the participants. Besides, we also accounted for head/face continuous movements caused by common activities such as (13) chewing; and (14) speaking. Notably, together with the one-shot movements, these are sources of noise which are unique to ear-worn devices. In fact, when performing these, the complex mesh of facial muscles moves substantially and, therefore, these activities are likely to cause significant deformations of the tissues in and around the ear. On the other hand, we also considered full-body activities such as (15) walking and (16) running, which give rise to well known sources of noise [173] in the PPG signal. The list of all the considered motion artifacts is reported in Table 6.8.3.

Ultimately, for all the conditions but the full-body movements (walking and running), we followed the wearable device validation guideline stipulated by the Consumer Technology Association [183] and measured PPG while seated in the upright position. During the resting condition, we instructed the participants to breath normally without moving. The

Table 6.8.3: List of the considered motion artifacts.

Class	Muscle Group	One-Shot	Artifact Name
Still	n/a	n/a	Still
Head/Face	Head	✓	Nod
		✓	Shake
		✓	Tilt
	Eyes	✓	Vertical Eyes Movements
		✓	Horizontal Eyes Movements
		✓	Brow Raiser
		✓	Brow Lowerer
		✓	Right Eye Wink
		✓	Left Eye Wink
	Mouth	✓	Lip Puller
		✓	Chin Raiser
		✓	Mouth Stretch
		×	Chewing
		×	Speaking
Full-Body	n/a	×	Walking
		×	Running

speaking condition consisted in a conversation with the investigator, where the participant described a recent event to the investigator. The chewing condition was assessed recording PPG data while the participant was chewing gum. For the full-body motion conditions, the participants were asked to walk and run at a set pace on a treadmill. We set the speed of the treadmill at $5kph$ and $8kph$ while walking and running, respectively. For each motion condition we recorded 2 minutes of data, automatically changing the configuration of the PPG parameters every 30 seconds using the values described in Section **??**. The length of the sessions was carefully chosen to be long enough to yield good-quality vital signs and yet not too tedious/harmful to the participants. Finally, we instructed the participants to repeat the one-shot movements roughly every 5 seconds.

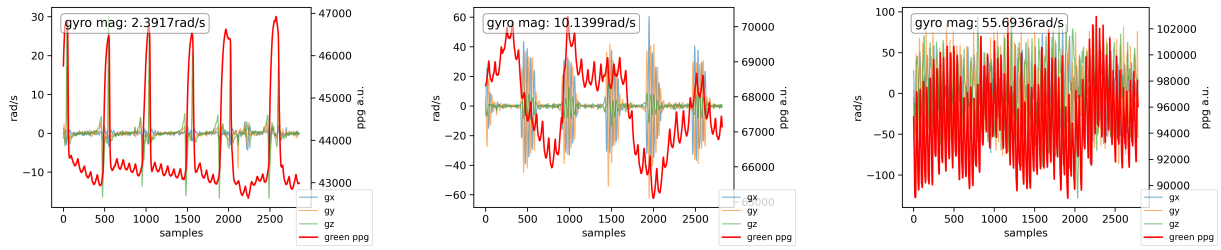
6.9 Results

In this section we present the analysis we performed to explore our dataset. Specifically, the PPG data (3 channels: green, red, infrared) recorded from the left and the right ears are processed independently. The data are aligned and stored in Pandas Data Frames. Each Data Frame is then re-sampled at 100Hz to make sure the sampling rate is consistent. The start and the end of each Data Frame is trimmed to ensure they have the same length. Notice that our exploration focuses on the 4th set of configuration parameters (LED current $32mA$; pulse width $123.8\mu s$; integration time $117.3\mu s$) as decribed in Section 6.8.2.

6.9.1 Dataset outlook and template matching

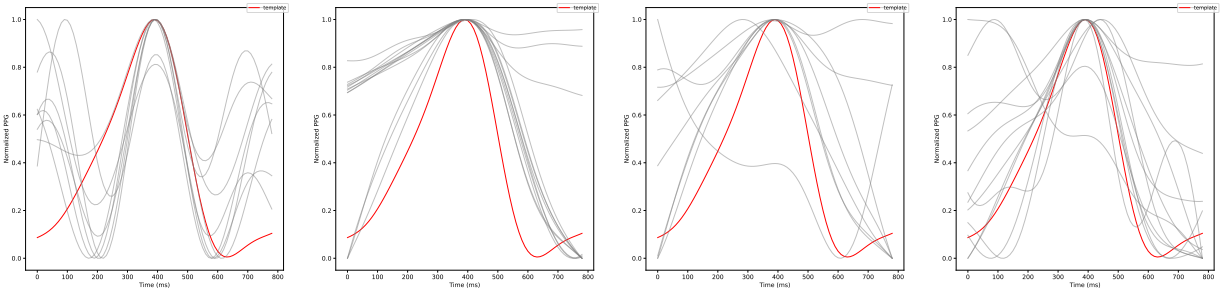
We begin our analysis by looking qualitatively at how the different motion artifacts impact the PPG signal. In Figure 6.9.1 we can appreciate at a glance how two diverse facial movements, such as lip puller (Figure 6.9.1a) and nod (Figure 6.9.1b), have a very different impact on the PPG trace when compared to a full-body movement like running (Figure 6.9.1c) – in which the signal is dominated by the running cadence rather than by the cardiac signal. Notably, we can observe substantial differences even among the two facial movements: while the impact of the lip puller appears very localized and aligned with the motion (as we can see from the variations along the gyroscope axes), the nod seems to have a more prolonged impact on the DC component of PPG trace. Manually inspecting the data we also noticed that for some combinations [participant, motion] the PPG was not affected by artifacts. In particular, the vertical and horizontal movement of the eyes did not cause any artifact on the PPG signals. This is due to the limited involvement of the facial muscles, and especially of those near the ears, during eye movements. Similarly, for the left and right eye wink motions, some participants could not perform the motion with both eyes or not at all. In other cases, the wink was subtle and hence did not result in any artifact in the corresponding PPG signal. For the rest of the analysis we filtered out these [participant, motion] combinations for which the PPG was not affected by motion.

To deepen our investigation, and gain a better visual understanding of how the various motion artifacts affect the morphology of the PPG pulses, we relied on a template matching analysis [80]. In doing so, we build a template pulse by taking the average of all the pulses of each user when still. Then plot the template pulse in red and use it as a reference against all the PPG pulses present in each motion session (plotted in gray). Figure 6.9.2 depicts the template matching analysis for shake (Figure 6.9.2a), brow raiser (Figure 6.9.2b), lip



(a) Lip Puller. **(b)** Nod. **(c)** Running.

Fig. 6.9.1: Samples of green PPG and IMU (gyroscope) data under different motion artifacts.



(a) Shake. **(b)** Brow Raiser. **(c)** Lip Puller. **(d)** Mouth Stretch.

Fig. 6.9.2: Template matching of PPG pulses from user 12 for four different motions.

puller (Figure 6.9.2c), and mouth stretch (Figure 6.9.2d). The plots show how each of the considered movement affects the morphology of the PPG pulse differently, resulting in subtle, yet notable artifacts. As we have seen in Section 2.3.3, many applications rely on morphological features computed on the PPG signals. Hence, such artifacts in the morphology of each pulse could lead to erroneous vitals estimation. We believe that our dataset represents a good resource for a more in depth study and characterization of this issue for an emerging class of devices – earables equipped with heath-related sensors.

6.9.2 Handcrafted metrics extraction and statistical analysis

We sought to proceed our exploration of the dataset with the extraction of the set of handcrafted features, as described in Section 2.3.3. Notably, for all of the metrics, but Perfusion Index, we apply a 4th order Butterworth band-pass filter (low-cut = 0.4Hz, high-cut = 4Hz) to smooth the PPG signal. To fairly compare the metrics values of different artifacts, we normalized their values (min-max normalization). We chose to normalize the metrics values of all the artifacts of each user. Specifically, normalizing every user independently allows us to retain the subject-dependent characteristics of the motion and

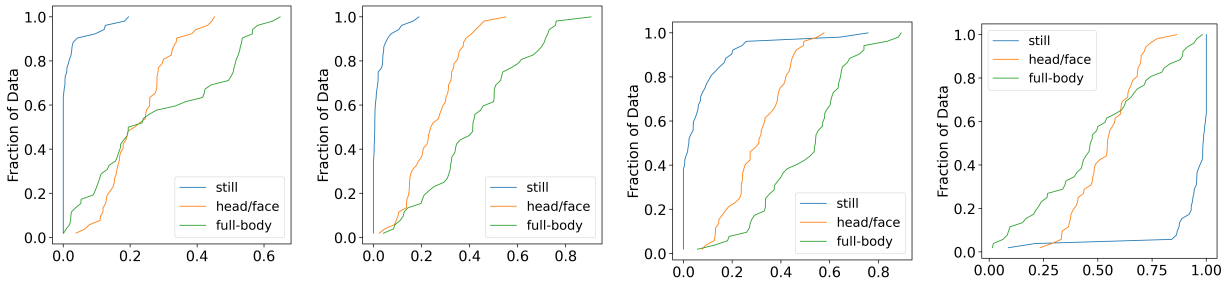
of the blood vessels morphology of the user.

Figure 6.9.3 reports the empirical cumulative distribution function (ECDF) of how head/face and full-body movements impact the Peak-to-Peak Magnitude Variance (Figure 6.9.3a), Peak-Time Interval Variance (Figure 6.9.3b), Perfusion Index (Figure 6.9.3c), and the Spectral Kurtosis (Figure 6.9.3d) of the in-ear PPG signal. Similar patterns can be observed for other metrics. For this analysis, we consider the (normalized) metrics computed for both the left and the right PPG for all the users. We can observe how the metrics values for still are mostly consistent across the entire population. On the other hand, although they show a different behavior, both head/face and full-body movements appear to have more widespread distributions. This is especially true for full-body movements. Notably, the findings of the spectral kurtosis analysis (Figure 6.9.3d) are also aligned to the literature [80], showing higher values for clean PPG signal. This can be explained by the presence of sharper peaks in the Fourier spectrum of clean (still in our case) PPG. These results suggest that different motion categories (i.e., head/face and full-body) create diverse artifacts in the PPG signal, and therefore it might be necessary to adopt dedicated approaches when applying signal filtering techniques. Our dataset represents a good source of data to start exploring this avenue.

Finally, to further discern whether there exist differences between the individual motions, we looked at the Mean Absolute Error (MAE) between all the metrics extracted form the PPG data under the various motion artifact and the still baseline. As we can see from Figure 6.9.4, for the majority of the PPG metrics, there are statistically significant differences between the still baseline and most of the artifacts. As expected, more intense head/face movements, like tilt and mouth stretch, present greater differences in the metrics computed from the still baseline. This is even clearer when looking at full-body movements. Besides, a comparison of data from the left (Figure 6.9.4b) and right (Figure 6.9.4a) ear hints to differences between the PPG collected from these two locations. Multi-site PPG signals from the ears have been largely understudied so far. We believe our dataset is the perfect starting point to further explore this area.

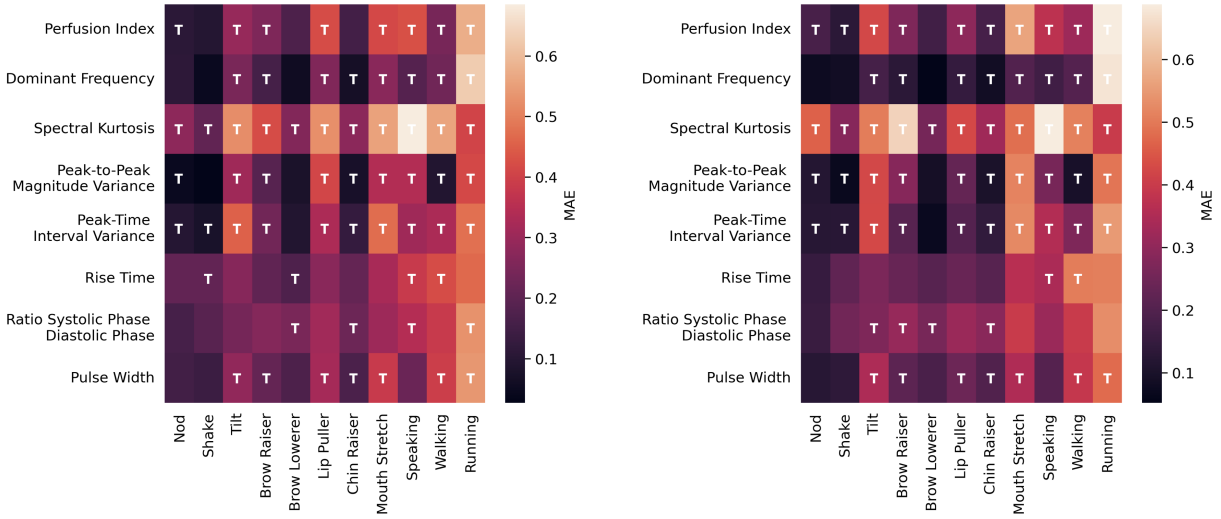
6.9.3 DNN-based motion artifacts classification

An important task when working with physiological sensor data involves detecting segments that are corrupted by noise with the intent of trying to clean them with filtering techniques or discard them completely to avoid errors during the bio-markers estimation. In this



(a) Peak-to-Peak Mag. Var. **(b)** Peak-Time Interval Var. **(c)** Perfusion Index. **(d)** Spectral Kurtosis.

Fig. 6.9.3: Empirical Cumulative Distribution Function (ECDF) of how the various classes of motion artifacts impact some of the handcrafted metrics extracted from PPG.



(a) Left ear. **(b)** Right ear.

Fig. 6.9.4: Heatmaps of how the various motion artifacts impact the handcrafted metrics extracted from the green PPG signal (Figure 6.9.4a left ear; Figure 6.9.4b right ear). The values reported in the heatmaps are the Mean Absolute Error (MAE) with respect to the still baseline. The heatmaps' cells are annotated with a T whenever there is statistically significant difference between the still baseline signal and the MA-corrupted one ($p < 0.05$).

Table 6.9.1: PPG classification model architecture.

Type / Stride / Padding	# Filter	Filter Size
Conv1D / S1 / Same	16	10
Conv1D / S2 / Valid	16	10
Conv1D / S1 / Same	32	10
Conv1D / S2 / Valid	32	10
Conv1D / S1 / Same	64	10
Conv1D / S2 / Valid	64	10
Global Avg. Pool	n/a	n/a
FC / S1	128	n/a
Softmax / S1	3	n/a

section we explore how our dataset can support the development of an automatic pipeline to detect PPG segments affected by motion artifacts and classify the nature of the artifact. For this purpose we rely on a Deep Neural Network (DNN) approach since it has been often used for this task in recent work [193, 194]. As preliminary evaluation we aim to distinguish clean PPG segments (i.e., when the participant is not moving) from segments affected by head/face or full-body artifacts. We have seen in Section 6.9.2 that these three categories (i.e., still, head/face and full-body) present distinct characteristics across various PPG metrics, hence we want to explore if these can be learned automatically from the PPG data. The classification of these three macro categories represents the first step towards a more complex pipeline used to detect more motion artifacts types, potentially using a hierarchical detection pipeline where the classification is refined at each step. We believe our dataset can support this task and we leave it as future work.

Model: Inspired from recent research on PPG classification [193, 194], we designed a model with 6 1-D convolutional layers with Rectified Linear Unit (ReLU) activation for feature extraction, and a single fully connected layer as final classifier. At alternating layers we use a stride of 2 to reduce the dimension of intermediate feature maps without using pooling layers. Table 6.9.1 reports the details of the model. The total number of trainable parameters is $88,451$ making it a small model that could potentially be deployed on wearable devices running on batteries.

Data pre-processing: For this evaluation we use only the data from the green wavelength and, similarly to our previous analysis, we focus on the 4th set of configuration parameters and filter the signal with a 4th order Butterworth band-pass filter with cutoff frequencies of [0.4, 4]Hz. We then split the data with 7 seconds windows with 70% overlap (i.e., 5

seconds), this balances the need for having sufficient data in each window to correctly model the artifact with the granularity of each classification. Each window is then associated with the corresponding class from the three we aim to classify, i.e., *still*, *head/face* and *full-body*. Similarly with our previous analysis (Section 6.9.2) we aggregate all the motions related to the face and head into a single class, while walking and running are considered as a separate class (*full-body*).

Evaluation strategy: To train and evaluate the model we adopted a 5-fold cross-validation approach. We created 5 folds such data for each fold the training and testing data would be taken from different participants, hence creating independent training and testing sets. However, the resulting train set is skewed towards more samples for the head/face class given that more artifacts are aggregated into a single class. For this reason we first down-sampled the majority class and then up-sampled the minority classes with the SMOTE techniques [195], which is a popular approach for data augmentation, in order to balance the training set. The testing set remained imbalanced, instead.

The model has been trained and tested on each fold using the categorical cross entropy loss function and RMSprop optimizer. Early stopping has been put in place to avoid overfitting of the model.

Results: The model performs reasonably well with an average accuracy across the 5 folds of 87% ($\sigma = 3\%$) and average F1 score of 79% ($\sigma = 15\%$), demonstrating that our dataset can well support automatic recognition tasks on PPG data.

Figure 6.9.5 shows the cumulative confusion matrix of the test sets across the 5 folds. We notice that the performance of the model is slightly lower for the still class. This could be due to the fact that the still class is always the minority class and, even despite data augmentation to re-balance the training set, there might be not sufficient data for an accurate modeling of that class. This is our preliminary effort in this direction to assess the quality of the dataset in supporting such tasks. We believe that a more sophisticated data augmentation approach could improve the classification accuracy of the still class and the overall accuracy of the model. We leave this exploration to future work.

6.10 Conclusion

In this chapter, we presented EarSet, a multi-modal (PPG, IMU, ECG) dataset for exploring and understanding the impact of head/facial and full-body movements on in-ear

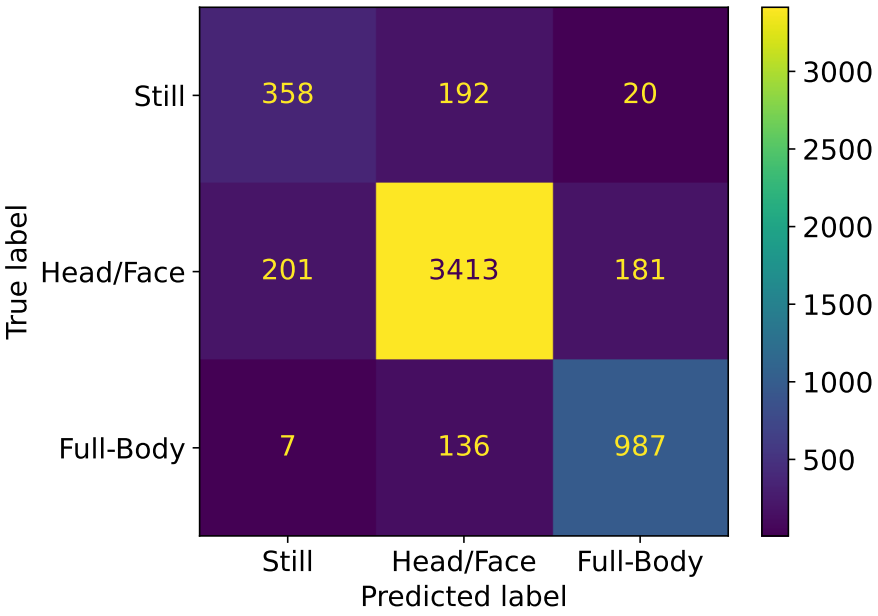


Fig. 6.9.5: Cumulative confusion matrix of the test sets across the 5 folds normalized by the total number of instances in each class.

PPG. We run an analysis based on the data we collected from 30 participants, across 17 2-minutes-long sessions. Our results suggest different head/facial movements cause diverse, often hard-to-predict, artifacts. We investigate eight handcrafted PPG features to study the differences between the still baseline and the various artifacts. Our exploration confirms our qualitative observations. Finally, we research whether our dataset could support the development of automatic pipelines to detect motion artifacts in PPG traces. We show it is possible to distinguish reasonably well still, head/face, and full-body movements, demonstrating that our dataset can enable researchers to investigate automatic recognition tasks. Ultimately, we believe there is room to develop multi-stage pipelines to detect and classify motion types, perhaps leveraging a combination of the in-ear PPG as well as the IMU data. Given the availability of several bio-markers from a reliable chest band, our dataset also enables the investigation of vitals estimation from the ears, both in static and in noisy conditions. This is a great opportunity given the rapidly growing interest in health sensing from the ears and head.

'Even the smallest person can change the course of the future.'

Galadriel

Chapter 7

Final Remarks

In this dissertation, we presented six original pieces of work investigating foundational challenges in mobile personal-scale sensing: understanding which sensing modalities can be integrated in earables without altering their functioning, whilst providing high-quality data to facilitate personal-scale applications; developing dedicated pipelines that can guarantee low latency (often quasi-real-time) in processing the collected data while, more importantly, ensuring minimal energy waste. Our premise was that emerging forms of wearables – such as earables – had the potential to collect high quality data from the human head, thus providing a different vantage point to perform several tasks. Ultimately, this dissertation demonstrated how earables can be used for personal-scale sensing: showing their great potential in monitoring fitness and vital signs (through photoplethysmography), recognizing human activities, providing context-aware information, and augmenting the user experience and awareness in general (e.g., by mean of inertial measurement units and in-ear microphones). In this concluding chapter, we briefly summarize our key contributions, providing an answer to the research questions we posed at the beginning of this manuscript (Chapter 1). Finally, we reflect on the shortcomings of the works presented in this dissertation, and suggest potential future research directions.

7.1 Summary of contributions

This dissertation aimed to answer three fundamental research questions in mobile personal-scale sensing.

7.1.1 Head motion tracking and addition of a calibrated magnetometer

In Chapter 3, we investigated earables for inertial sensing, showing how earables can be used for accurate head motion tracking. Our inertial head motion tracking system combines multiple streams of data (i.e., linear acceleration and angular velocity) coming from both earbuds (left and right), achieving results precise up to a few degrees, also under realistic situations (e.g., with the subjects speaking or chewing).

However, without continuously re-calibrating the accelerometer and gyroscope with the aid of a magnetometer, it is extremely hard to have absolute orientations and accurate tracking of sustained motions over time. Because of that, we sought to understand the factors that hinder the presence of a magnetometer in commercially available earables. We found that despite RF communications (i.e., Bluetooth) yielding unpredictable, variable interference on the magnetometer readings, having an appropriate calibration can be leveraged to remove the induced interference. A comprehensive evaluation of our proposed calibration routine shows how our user-transparent approach performs well under different conditions, achieving convincing results with errors below 3° for the majority of the experiments. By doing so, we proved how earables can be successfully augmented with a magnetometer, thus unlocking a number of personal-scale sensing applications, e.g., navigation and augmenting the user experience of their surroundings.

7.1.2 In-ear microphone-based general motion sensing and user identification

In Chapter 4 and Chapter 5, we explored the potential of in-ear facing microphones for personal-scale human motion sensing. We showed how in-ear microphones can be used instead of inertial sensors in sensing light motions when the latter fail. Leveraging the occlusion effect, a natural phenomenon that boosts the low-frequency components of bone-conducted sounds, we proposed OEsense, an acoustic-based in-ear system for general human motion sensing, and EarGate, an acoustic-gait-based user identification system.

OESense achieves promising results across three different personal-scale applications: step counting, activity recognition, and hand-to-face gesture interaction. Specifically, thanks to the natural boost operated by the occlusion effect, OESense reaches 99.3% step counting recall, 98.3% recognition recall for 5 activities, and 97.0% recall for five tapping gestures

on human face, respectively. The addition of such functionalities is compatible with the earbuds' fundamental functionalities (e.g., music playback and phone calls). More importantly, in terms of energy, OESense has a minimal overhead which is more than acceptable denoting encouraging potential to be integrated into future earbuds.

On the other hand, with EarGate, we leveraged the good performance achieved by OESense at step counting to build a system for gait-based identification. In particular, EarGate aims at extracting acoustic gait from the sounds induced by walking and propagated through the musculoskeletal system in the body. We showed how EarGate can achieve up to 97.26% Balanced Accuracy (BAC) with very low False Acceptance Rate (FAR) and False Rejection Rate (FRR) of 3.23% and 2.25%, respectively. As OESense, also EarGate yield a minimal overhead, both in terms of power consumption as well as latency. Ultimately, with OESense and EarGate, we demonstrated how earables equipped with in-ear microphones can be leveraged to unlock a number of personal-scale applications ranging from activity recognition to user identification and authentication.

7.1.3 In-ear photoplethysmography: best location for in-ear PPG sensing and robustness to motion artifacts

To answer the third of our research questions, in Chapter 6, we studied in-ear photoplethysmography (PPG) aiming to uncover a new domain of personal-scale sensing opportunities to track cardiac healthcare and fitness with earables. We reported an in-depth characterization of the best placement to enable accurate vital signs measurements from in-ear PPG, specifically looking at heart rate (HR), heart rate variability (HRV), blood oxygen saturation (SpO_2), and respiration rate (RR). We found that embedding a PPG sensor in the ear-tip of an earable yields the lowest inter-subject variability. Unfortunately, our results also suggests the absolute error increases substantially up to 29.84%, 24.09%, 3.28%, and 30.80% respectively for HR, HRV, SpO_2, and RR, during motion activities.

Having identified the in-the-canal (ITC) placement as the best location for in-ear PPG sensing, we designed a novel ear tip featuring a 3-channel PPG coupled with a 6-axis motion sensor (IMU) co-located on the same tip. This enabled us to study the motion signatures of a number of different motion artifacts, ranging from both full body motions to more subtle facial expressions. In particular, understanding the latter – which are unique to earables – is critical to the success of earables as the next wearable platform for accurate and reliable cardiovascular health monitoring. To this end, we collected a

unique dataset with 30 people and 18 different motion artifacts which will greatly support research efforts towards better characterizing in-ear vital signs sensing, making it more accurate and robust. By doing so, we paved the way to unlocking the full potential of the next-generation PPG-equipped earables for personal-scale sensing.

7.2 Limitations and future directions

The work we presented has several potential implications impacting various communities and stakeholders. For instance, researchers could use the methods, ideas, and datasets collected as a starting point for their own studies, perhaps applying them to new forms of wearable devices. At the same time, engineers could embed the algorithmic pipelines and the models presented in this dissertation into the new generation of products, targeting smart earbuds devices which understand the context at the user-scale, predicting mental health issues, fitness, and metabolic health. Finally, medical practitioners could benefit from the outputs of these new devices to gain a continuous snapshot of their patients' health away from hospital settings. In the reminder of this section, we will detail the major shortcomings of the works presented in this dissertation, discussing the possible room for improvements and the new future directions this research could foster.

7.2.1 Dataset size

One of the biggest limitations of the works described in this dissertation is the modest size of the datasets we collected. Although the results have been statistically validated, small datasets may yield less general conclusions. Further, a small population study inevitably affects the depth of the analysis that can be carried out on the dataset itself. However, collecting big datasets is a remarkable challenge stemming from multiple factors. First and foremost, the lack of commercially available earable research platforms, together with the lack of application programming interfaces (APIs) to access raw data, forced us to prototype our own custom devices most of the time. However, the intrinsic fragility of early stage research-level prototypes introduces additional inevitable hurdles, further complicating the task. As a matter of fact, research prototypes often break, yielding poor quality data, and often forcing the participants to redo part of the data collection. Ultimately, this also affects the researchers, who have to look for hacks out of their comfort zone to fix their prototypes. Secondly, because of their very time consuming nature, it is very hard to recruit participants to take part in data collection exercises. In addition to that, most

of the research presented in this dissertation has been done during the COVID-19 pandemic adding even more complexity with recruiting participants and performing intricate data collection sessions while maintaining COVID-19 protocols. While dataset like EarSet (Chapter 6) opens up novel opportunities in the mobile sensing and earable computing space, when collecting data there are often confounding factors that are either hard to take into account, or that are left behind for a number of reasons (including time constraints). For example, skin tone is an additional factor that could affect the data quality of PPG [173]. EarSet offers diversity in skin tones, however without uniform distribution among the six categories of pigmentation [196].

Future work should consider expanding the datasets presented in this dissertation in order to include additional participants to uniformly cover different ethnic groups, genders, and ages. Moreover, despite it being an even greater challenge in itself, it would be valuable to collect data from people suffering from heterogeneous underlying health conditions, to increase the diversity of the data. In fact, all the data collected for this dissertation belong to participants who were healthy at the time of the data collection and had no known underlying conditions.

7.2.2 Machine learning algorithms

In this dissertation we have often relied on machine learning (ML) models to fulfill complex tasks. Examples of these are Logistic Regression (LR), Support Vector Machine (SVM), K Nearest Neighbours (KNN), Decision Tree (DT), and Random Forest (RF) in OESense (Chapter 4), Support Vector Machine in EarGate (Chapter 5). The choice to leverage basic ML models like the ones listed above was dictated mostly by the limited processing capabilities available on the prototype we built. At the same time, another key reason we opted for traditional machine approaches over more advanced Deep Learning (DL) techniques lies in the limited size of the datasets. However, relying on basic ML models such those mentioned above, comes with a number of shortcomings. For example, traditional ML approaches usually yield lower accuracy (and have worse generalization abilities). Additionally, unlike deep learning, traditional ML models, like SVM, do require manual feature extraction and engineering. Ultimately, as we have seen for both OESense and EarGate, in Chapter 4 and Chapter 5 respectively, the feature extraction phase is indeed the most power hungry and time consuming step of the pipeline. In addition to the basic models we discussed above, in this dissertation we also make use of transfer learning [162, 197] to

boost the performance of EarGate Chapter 5 and of a shallow deep neural network to run preliminary analysis on EarSet Chapter 6. However, once again, the limited amount of data available for training hindered the performances of these approaches.

Possible future research should aim to explore data augmentation and more advanced, low-power deep learning techniques to try further boosting the performances whilst, for example, waiving the time spent and the power used in extracting features. Specifically to EarGate, for instance, an auto-encoder [165] could be trained on all the users' data to obtain a more general model. By working in an unsupervised manner, the auto-encoder could learn representations of the data without the need for labels. The following stage would consist of automatically extracting a pool of features from the auto-encoder. Finally, such features could be used to train personalized models, that would then learn even better representations of each user. This could be done, for example, by running a k-neighbor search on the auto-encoder space, making sure user A is the closest neighbor of themselves, e.g., regardless the footwear. Alternatively, another interesting research direction would be exploring teacher-student architectures [198,199] to distill complex models into constrained devices as earables.

7.2.3 New modalities and forms

Ultimately, in this dissertation we have discussed in-depth three different modalities: inertial sensors and magnetometers; microphones; photoplethysmography sensors. While these unleash a wealth of fascinating opportunities, opening the door to numerous personal-scale applications, there are many more modalities whose in-ear potential is well worth investigating more. Examples include, but are not restricted to, in-ear EEG and EOG electrodes, temperature sensors, as well as infrared sensors (e.g., for posture monitoring).

Besides exploring new modalities, the success of earables as sensing and computing devices suggest that new forms of wearable devices should also be taken into account. To this end, future research could transfer the methods described in, and the lessons learned through, this dissertation on to different forms of wearables, such as smart-rings, smart patches/plasters, 3D-printed smart-tattoos, etc.. Finally, this dissertation did not research in-depth multi-devices collaboration paradigms, e.g., to maximize battery life and accuracy. Future research efforts should consider multi-modal multi-device collaboration as a potential research avenue to boost the capabilities of earables as a personal-scale sensing platform.

List of Figures

2.2.1 Anatomy of the human ear and explanation of the occlusion effect. When the orifice of the ear canal is occluded, sounds are trapped inside the ear canal, resulting in the amplification of their low frequency components. . . 17

2.3.1 Blood vessels around the outer ear (Figure 2.3.1a). Pulsatile (AC) and non-pulsatile (DC) components of a typical PPG signal (infrared wavelength in this example). The DC component indicates the light absorption from the tissues, the bones, and the static blood in veins and capillaries; the AC component reflects the pulsations of the arterial blood caused by the cardiac activity (Figure 2.3.1b). 22

2.3.2 Typical time domain signal features extracted from a PPG signal. 25

3.3.1 Example of the Gimbal Lock: two out of three degrees of freedom collide on the same plane (graphic credits: MathsPoetry, CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0, via Wikimedia Commons). 36

3.4.1 3.4.1a Mean error and standard deviation of the head movements estimation of 10 silent volunteers. 3.4.1b Impact of chewing activity on the mean error and standard deviation of the head movements estimation of 10 volunteers. 3.4.1c Impact of speech on the mean error and standard deviation of the head movements estimation of 10 volunteers. 38

3.5.1 Magnetometer readings from a STEVAL-STLCS01V1 device at different distances from one eSense earbud. Notice how the scale of the plots changes dramatically from top to bottom. 41

3.5.2 We run two experiments where a volunteer was asked to shake his head while having the a STEVAL-STLCS01V1 close to the ear. (a) We first collected data without magnetic interference. (b) We then asked our volunteer to repeated the same movements with the STEVAL-STLCS01V1 placed inside the eSense case. (c) This way we could observe how correcting the offset introduced by the magnets, the magnetometer placed in the earbud is still able to record motion data. 42
3.6.1 Basic system setup (Figure 3.6.1a) and setup used to isolate the interference generated by BT streaming and speaker (Figure 3.6.1b). 45
3.7.1 Impact of audio streaming and consequent music playback on the magnetometers readings. 47
3.7.2 Impact of voice calls and music playback on the magnetometer readings. . 47
3.7.3 48
3.7.4 48
3.7.5 3.7.3 FFT of the magnetometer traces while playing (on high volume) pure tones at $20Hz$. 3.7.4 Impact of the interference generated by BT streaming and by the speaker on the heading estimated by the raw magnetometer traces. Although the magnitude of the interference on calibrated traces is smaller, for clarity in this figure we use raw readings. 48
3.8.1 Intuition behind the proposed calibration technique. 50
3.8.2 Auto-Calibration procedure: the calibration watcher checks if the phone is picked up (i.e., during an unlock event). If such event occurs, and the phone is unlocked in a position suitable for calibration, then the data is collected and stored. Once we have enough data to calibrate (i.e., at least two), then we perform the calibration procedure as described in Equation (3.7). If successful, before applying the updated calibration values we perform a sanity check to eliminate any potential interference or bad measurements; which, if found present, the newly calibrated parameters are discarded. . . 54
3.9.1 Setup used to benchmark the proposed calibration technique (3.9.1a) and volunteer wearing the Arduino as if they were earbuds (3.9.1b). This is the setup used for our in-the-wild use test. 55
3.9.2 Errors aggregated per testing angle when changing: 3.9.2a the spacing between two reference headings; 3.9.2b increasing the number of references; 3.9.2c the number of data-points fed to our algorithm. 57

3.9.3 Heading estimation (3.9.3a) and mean errors (3.9.3b) for small angles. . . 59
3.9.4 Heading estimation (3.9.4a) and mean errors (3.9.4b) for large angles. . . 59
3.9.5 In-the-wild heading estimation (3.9.5a) and mean errors (3.9.5b). 60

4.2.1 Comparison of signals from the accelerometer without (upper row) and with head movements (lower row). 67
4.2.2 Comparison of signals from the external facing microphone without (upper row) and with background noise (lower row). 68
4.3.1 Impact of occlusion effect on the frequency response. 69
4.3.2 Comparison of signals from the inward-facing microphone without (upper row) and with head movements (lower row). 69
4.3.3 Comparison of signals from the inward-facing microphone without (upper row) and with background noise (lower row). 70
4.3.4 A walking segment showing the performance of proposed step counting algorithm. 73
4.4.1 The developed data recording prototype (4.4.1a), and a participant wearing the device (4.4.1b). 74
4.5.1 Illustration of the designed tapping gestures (4.5.1a), gesture waveform from both earbuds of Subject 1 (4.5.1b). 77
4.6.1 Step counting performance for walking on brick (4.6.1a) and on carpet (4.6.1b). 80
4.6.2 Overall activity recognition performance (4.6.2a) and confusion matrix (4.6.2b). 81
4.6.3 Activity recognition performance for leave-one-out test (4.6.3a), individual performance with model personalization (4.6.3b), and average recognition performance with different amounts of personal data (4.6.3c). 81
4.6.4 Overall gesture recognition performance on 12 gestures (4.6.4a), individual gesture recognition performance with different gestures (4.6.4b), and impact of training data size averaged over 29 subjects (4.6.4c). 83
4.6.5 The original signal (4.6.5a), the low-pass filtered signal for participant walking during music playback (4.6.5b), the low-pass filtered signal for the same participant walking without music playback (4.6.5c). 84
4.6.6 Spectrum energy analysis of 100 songs. 85

5.2.1 Walking signals collected with a traditional outward facing microphone and an in-ear microphone under different conditions. Notably, thanks to the occlusion effect, the in-ear microphone data shows higher gain at low frequencies ($< 50Hz$) and more resilience towards environmental noise. . . 93
5.2.2 5.2.2a Extracted feature vectors for four different gait cycles from the same subject, and 5.2.2b gait cycles from four different subjects. Figure 5.2.2a clearly shows how the gait cycles captured by EarGate are consistent for each individual (i.e. high intra-class similarity). Figure 5.2.2b, on the other hand, shows how the gait cycles are distinguishable among subjects (i.e. high inter-class difference). 93
5.3.1 EarGate functioning. 95
5.3.2 (a) The low-pass filtered signal collected when one participant is walking, (b) a segment showing the performance of proposed gait segmentation algorithm. 97
5.5.1 Training and testing data splitting scheme for 5.5.1a one-class SVM, 5.5.1b imbalanced binary SVM, 5.5.1c balanced binary SVM with all subjects' data, and 5.5.1d balanced SVM with part of subjects' data. $P : N$ represents the ratio between positive (legitimate) and negative (impostor) gait. 101
5.5.2 5.5.2a Overall performance of EarGate (averaged across 31 subjects), 5.5.2b BAC of each individual using OC-SVM and Bi-SVM (Imbalanced), 5.5.2c Comparison of the four proposed training-testing protocol, which also reflects the impact of data imbalance. 102
5.5.3 Performance comparison 5.5.3a FAR, 5.5.3b FRR, and 5.5.3c BAC, of different footwear and ground material. 103
5.5.4 Impact of user speaking. We take one subject walking on tiles as an example, 5.5.4a original signal with speaking, 5.5.4b low-pass filtered signal with speaking, 5.5.4c filtered signal without speaking. 105
5.5.5 Impact of music playback. We take one subject walking on tiles as an example, 5.5.5a original signal with music playing, 5.5.5b low-pass filtered signal with music playing, 5.5.5c filtered signal without music. 105
5.5.6 Impact of 5.5.6a fusing data from both the earbuds, 5.5.6b different training size, and 5.5.6c different pace and features, on FAR, FRR, and BAC. . . . 107

6.2.1 Anatomy of the human ear with annotated the chosen sensor placements. 121
6.3.1 Processing pipelines used to extract vital signs. 123

6.4.1 On the left a participant wearing the ITC PPG; and on the right the devices used in the data collection (clockwise: Zephyr Bioharness 3.0, Masimo Health MightySat-Rx, Cosinuss two tip on eSense for ITC, modified Cosinuss two tip used for ITE and BTE). 127
6.4.2 Impact of different ear-placements on the goodness of PPG-extracted biomarkers in a resting condition. 128
6.5.1 Difference between ground truth and PPG heart rate plotted against the mean of the two measurements (Bland-Altman plot). In the interest of space, only data for HR is shown. The solid black line represents the mean error, and the dashed gray lines the 1.96 SD boundaries (95% limits of agreement). 130
6.5.2 Impact of different external artifacts (i.e. speaking, light motion, and intense motion) on the goodness of BTE, ITE, and ITC PPG-extracted biomarkers. 131
6.8.1 (a) Flexible PCB implementation of our earbud with a MAX86161 PPG sensor and a co-located ST LSM6DSRX IMU (b) An in-ear soft earbud was realized by embedding the in-ear flexible PCB board into a transparent silicone mould. (c) Head-worn data acquisition device consisting of an ESP32 micro-controller collecting data from in-ear PPG and IMU sensors in the left and right ear. (d) A participant wearing our earbud based prototype and taking part in the data collection protocol. 137
6.9.1 Samples of green PPG and IMU (gyroscope) data under different motion artifacts. 143
6.9.2 Template matching of PPG pulses from user 12 for four different motions. 143
6.9.3 Empirical Cumulative Distribution Function (ECDF) of how the various classes of motion artifacts impact some of the handcrafted metrics extracted from PPG. 145
6.9.4 Heatmaps of how the various motion artifacts impact the handcrafted metrics extracted from the green PPG signal (Figure 6.9.4a left ear; Figure 6.9.4b right ear). The values reported in the heatmaps are the Mean Absolute Error (MAE) with respect to the still baseline. The heatmaps' cells are annotated with a T whenever there is statistically significant difference between the still baseline signal and the MA-corrupted one ($p < 0.05$). 145

6.9.5 Cumulative confusion matrix of the test sets across the 5 folds normalized by the total number of instances in each class. 148

Bibliography

[1] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9):140–150, 2010.

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

[3] Takashi Nakamura, Valentin Goverdovsky, and Danilo P Mandic. In-ear eeg biometrics for feasible and readily collectable real-world person authentication. *IEEE Transactions on Information Forensics and Security*, 13(3):648–661, 2017.

[4] Andrea Ferlini, Alessandro Montanari, Chulhong Min, Hongwei Li, Ugo Sassi, and Fahim Kawsar. In-ear ppg for vital signs. *IEEE Pervasive Computing*, pages 1–10, 2021.

[5] Andrea Ferlini, Alessandro Montanari, Andreas Grammenos, Robert Harle, and Cecilia Mascolo. Enabling in-ear magnetic sensing: Automatic and user transparent magnetometer calibration. *The 19th International Conference on Pervasive Computing and Communications (PerCom 2021)*, 2021.

[6] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. Earbit: using wearable sensors to detect eating episodes in unconstrained environments. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 1(3):1–20, 2017.

[7] Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, et al. Auracle:

Detecting eating episodes with an ear-mounted sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–27, 2018.

[8] Chulhong Min, Akhil Mathur, and Fahim Kawsar. Exploring audio and kinetic sensing on earable devices. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, pages 5–10, 2018.

[9] Fahim Kawsar, Chulhong Min, Akhil Mathur, Alessandro Montanari, Utku Günay Acer, and Marc Van den Broeck. eSense: Open Earable Platform for Human Sensing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, pages 371–372, 2018.

[10] Jay Prakash, Zhijian Yang, Yu-Lin Wei, and Romit Roy Choudhury. STEAR: Robust Step Counting from Earables. In *Proceedings of the 1st International Workshop on Earable Computing*, pages 36–41, 2019.

[11] Siddharth Rupavatharam and Marco Gruteser. Towards in-ear inertial jaw clenching detection. In *Proceedings of the 1st International Workshop on Earable Computing*, pages 54–55, 2019.

[12] Erika Bondareva, Elín Rós Hauksdóttir, and Cecilia Mascolo. Earables for detection of bruxism: a feasibility study. 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*, pages 146–151, 2021.

[13] Andrea Ferlini, Alessandro Montanari, Cecilia Mascolo, and Robert Harle. Head motion tracking through in-ear wearables. In *Proceedings of the 1st International Workshop on Earable Computing*, EarComp'19, pages 8–13, New York, NY, USA, 2019. Association for Computing Machinery.

[14] Sheng Shen, Mahanth Gowda, and Romit Roy Choudhury. Closing the gaps in inertial motion tracking. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pages 429–444, 2018.

[15] Pengfei Zhou, Mo Li, and Guobin Shen. Use it free: Instantly knowing your phone attitude. In *Proceedings of the 20th annual international conference on Mobile computing and networking*, pages 605–616, 2014.

[16] Markus Lueken, Xiaowei Feng, Boudewijn Venema, Berno JE Misgeld, and Steffen Leonhardt. Photoplethysmography-based in-ear sensor system for identification of

increased stress arousal in everyday life. In *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 83–86. IEEE, 2017.

[17] Boudewijn Venema, Johannes Schiefer, Vladimir Blazek, Nikolai Blanik, and Steffen Leonhardt. Evaluating innovative in-ear pulse oximetry for unobtrusive cardiovascular and pulmonary monitoring during sleep. *IEEE journal of translational engineering in health and medicine*, 1:2700208–2700208, 2013.

[18] Qingxue Zhang, Xuan Zeng, Wenchuang Hu, and Dian Zhou. A machine learning-empowered system for long-term motion-tolerant wearable monitoring of blood pressure and heart rate with ear-ecg/ppg. *IEEE Access*, 5:10547–10561, 2017.

[19] Dong Ma, Andrea Ferlini, and Cecilia Mascolo. Oesense: Employing occlusion effect for in-ear human sensing. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '21, page 175–187, New York, NY, USA, 2021. Association for Computing Machinery.

[20] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. Eargate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 337–349, 2021.

[21] Furong Yang, Andrea Ferlini, Davide Aguiari, Davide Pesavento, Rita Tse, Suman Banerjee, Gaogang Xie, and Giovanni Pau. Revisiting wifi offloading in the wild for v2i applications. *Computer Networks*, 202:108634, 2022.

[22] Ashwin Ahuja, Andrea Ferlini, and Cecilia Mascolo. *PilotEar: Enabling In-Ear Inertial Navigation*, page 139–145. Association for Computing Machinery, New York, NY, USA, 2021.

[23] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo. Motion-resilient heart rate monitoring with in-ear microphones. *arXiv preprint arXiv:2108.09393*, 2021.

[24] Andrea Ferlini, Wei Wang, and Giovanni Pau. Corner-3d: A rf simulator for uav mobility in smart cities. In *Proceedings of the ACM SIGCOMM 2019 Workshop on Mobile AirGround Edge Computing, Systems, Networks, and Applications*, pages 22–28, 2019.

[25] Alexandru Solot and Andrea Ferlini. Leader-follower formations on real terrestrial robots. In *Proceedings of the ACM SIGCOMM 2019 Workshop on Mobile AirGround Edge Computing, Systems, Networks, and Applications*, pages 15–21, 2019.

[26] Davide Aguiari, Andrea Ferlini, Jiannong Cao, Song Guo, and Giovanni Pau. C-continuum: Edge-to-cloud computing for distributed ai. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1053–1054. IEEE, 2019.

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

[28] Oliver Amft, Mathias Stäger, Paul Lukowicz, and Gerhard Tröster. Analysis of chewing sounds for dietary monitoring. In *International Conference on Ubiquitous Computing*, pages 56–72. Springer, 2005.

[29] Oliver J Woodman. An introduction to inertial navigation. Technical report, University of Cambridge, Computer Laboratory, 2007.

[30] Steven M LaValle, Anna Yershova, Max Katsev, and Michael Antonov. Head tracking for the oculus rift. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 187–194. IEEE, 2014.

[31] J Metge, R Mégret, A Giremus, Y Berthoumieu, and T Décamps. Calibration of an inertial-magnetic measurement unit without external equipment, in the presence of dynamic magnetic disturbances. *Measurement Science and Technology*, 25(12):125106, 2014.

[32] Roelof Versteeg, Mark McKay, Matt Anderson, Ross Johnson, Bob Selfridge, and Jay Bennett. Feasibility study for an autonomous uav-magnetometer system. Technical report, IDAHO NATIONAL LAB IDAHO FALLS, 2007.

[33] Douglas Samuel Jones. *The theory of electromagnetism*. Elsevier, 2013.

[34] Ke Han, He Han, Zhifeng Wang, and Feng Xu. Extended kalman filter-based gyroscope-aided magnetometer calibration for consumer electronic devices. *IEEE Sensors Journal*, 17(1):63–71, 2016.

[35] Muhammad Tahir, Abdullah Moazzam, and Khurram Ali. A stochastic optimization approach to magnetometer calibration with gradient estimates using simultaneous perturbations. *IEEE Transactions on Instrumentation and Measurement*, 68(10):4152–4161, 2018.

[36] Andreas Grammenos, Cecilia Mascolo, and Jon Crowcroft. You are sensing, but are you biased? a user unaided sensor calibration approach for mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–26, 2018.

[37] Raj Sodhi, Jay Prunty, George Hsu, and Becky Oh. Automatic calibration of a three-axis magnetic compass, November 18 2008. US Patent 7,451,549.

[38] Hongfeng Pang, Shitu Luo, Mengchun Pan, Qi Zhang, and Ruifang Xie. Calibration of three-axis magnetometer diversionary error based on equipment and lms adaptive algorithm. In *Sixth International Symposium on Precision Engineering Measurements and Instrumentation*, volume 7544, page 75445P. International Society for Optics and Photonics, 2010.

[39] Erin L Renk, M Rizzo, W Collins, Fuju Lee, and Dennis S Bernstein. Calibrating a triaxial accelerometer-magnetometer-using robotic actuation for sensor reorientation during data collection. *IEEE Control Systems Magazine*, 25(6):86–95, 2005.

[40] Frédéric Camps, Sébastien Harasse, and André Monin. Numerical calibration for 3-axis accelerometers and magnetometers. In *2009 IEEE International Conference on Electro/Information Technology*, pages 217–221. IEEE, 2009.

[41] Nathaniel Bowditch. *The American practical navigator: an epitome of navigation/vol. 1.* 2018.

[42] Ahmed Wahdan, Jacques Georgy, Walid F Abdelfatah, and Aboelmagd Noureldin. Magnetometer calibration for portable navigation devices in vehicles using a fast and autonomous technique. *IEEE Transactions on Intelligent Transportation Systems*, 15(5):2347–2352, 2014.

[43] Timo Pylvänäinen. Automatic and adaptive calibration of 3d field sensors. *Applied Mathematical Modelling*, 32(4):575–587, 2008.

[44] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 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*, pages 679–689, 2017.

[45] Denys JC Matthies, Bernhard A Strecker, and Bodo Urban. EarFieldSensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1911–1922, 2017.

[46] Fahim Kawsar, Romit Roy Choudhury, and Ganesh Ananthanarayanan. Pervasive video and audio. *IEEE Pervasive Computing*, 20(2):7–8, 2021.

[47] Zhijian Yang, Yu-Lin Wei, Sheng Shen, and Romit Roy Choudhury. Ear-ar: indoor acoustic augmented reality on earphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–14, 2020.

[48] Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Brown, Jagmohan Chauhan, Ting Dang, Andreas Grammenos, Apinan Hasthanasombat, Andres Floto, et al. Sounds of covid-19: exploring realistic performance of audio-based digital testing. *NPJ digital medicine*, 5(1):1–9, 2022.

[49] Ting Dang, Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Andres Floto, et al. Covid-19 disease progression prediction via audio signals: A longitudinal study. *arXiv preprint arXiv:2201.01232*, 2022.

[50] Junjue Wang, Kaichen Zhao, Xinyu Zhang, and Chunyi Peng. Ubiquitous keyboard for small mobile devices: harnessing multipath fading for fine-grained keystroke localization. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, pages 14–27, 2014.

[51] Yanzhi Ren, Chen Wang, Jie Yang, and Yingying Chen. Fine-grained sleep monitoring: Hearing your breathing with smartphones. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 1194–1202. IEEE, 2015.

[52] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. reathPrint: Breathing acoustics-based user authentication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 278–291, 2017.

[53] Jian Liu, Yan Wang, Gorkem Kar, Yingying Chen, Jie Yang, and Marco Gruteser. Snooping keystrokes with mm-level audio ranging on a single phone. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 142–154, 2015.

[54] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. SoundWave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1911–1914, 2012.

[55] Wei Wang, Alex X Liu, and Ke Sun. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pages 82–94, 2016.

[56] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th annual international conference on mobile systems, applications, and services*, pages 15–28, 2017.

[57] Kévin Carillo, Olivier Doutres, and Franck Sgard. Theoretical investigation of the low frequency fundamental mechanism of the objective occlusion effect induced by bone-conducted stimulation. *The Journal of the Acoustical Society of America*, 147(5):3476–3489, 2020.

[58] Michael A Stone, Anna M Paul, Patrick Axon, and Brian CJ Moore. A technique for estimating the occlusion effect for frequencies below 125 hz. *Ear and hearing*, 35(1):49, 2014.

[59] Stefan Stenfelt. Acoustic and physiologic aspects of bone conduction hearing. In *Implantable bone conduction hearing aids*, volume 71, pages 10–21. Karger Publishers, 2011.

[60] Roman Schlieper, Song Li, Stephan Preihs, and Jürgen Peissig. The relationship between the acoustic impedance of headphones and the occlusion effect. In *Audio Engineering Society Conference: 2019 AES INTERNATIONAL CONFERENCE ON HEADPHONE TECHNOLOGY*. Audio Engineering Society, 2019.

[61] Stefan Stenfelt and Sabine Reinfeldt. A model of the occlusion effect with bone-conducted stimulation. *International journal of audiology*, 46(10):595–608, 2007.

[62] Marcos Serrano, Barrett M Ens, and Pourang P Irani. Exploring the use of hand-to-face input for interacting with head-worn displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3181–3190, 2014.

[63] Takashi Kikuchi, Yuta Sugiura, Katsutoshi Masai, Maki Sugimoto, and Bruce H Thomas. EarTouch: turning the ear into an input surface. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–6, 2017.

[64] Koki Yamashita, Takashi Kikuchi, Katsutoshi Masai, Maki Sugimoto, Bruce H Thomas, and Yuta Sugiura. CheekInput: turning your cheek into an input surface by embedded optical sensors on a head-mounted display. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, pages 1–8, 2017.

[65] Juyoung Lee, Hui-Shyong Yeo, Murtaza Dhuliawala, Jedidiah Akano, Junichi Shimizu, Thad Starner, Aaron Quigley, Woontack Woo, and Kai Kunze. Itchy Nose: discreet gesture interaction using EOG sensors in smart eyewear. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 94–97, 2017.

[66] Xuhai Xu, Haitian Shi, Xin Yi, Wenjia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K Dey. EarBuddy: Enabling On-Face Interaction via Wireless Earbuds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[67] Changsheng Wan, Li Wang, and Vir V Phoha. A survey on gait recognition. *ACM Computing Surveys (CSUR)*, 51(5):1–35, 2018.

[68] Maria De Marsico and Alessio Mecca. A survey on gait recognition via wearable sensors. *ACM Computing Surveys (CSUR)*, 52(4):1–39, 2019.

[69] Guoying Zhao, Guoyi Liu, Hua Li, and Matti Pietikainen. 3d gait recognition using multiple cameras. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 529–534. IEEE, 2006.

[70] Lee Middleton, Alex A Buss, Alex Bazin, and Mark S Nixon. A floor sensor system for gait recognition. In *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, pages 171–176. IEEE, 2005.

[71] Wei Wang, Alex X Liu, and Muhammad Shahzad. Gait recognition using wifi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 363–373, 2016.

[72] Davrondzhon Gafurov, Kirsi Helkala, and Torkjel Søndrol. Biometric gait authentication using accelerometer sensor. *JCP*, 1(7):51–59, 2006.

[73] Dong Ma, Guohao Lan, Weitao Xu, Mahbub Hassan, and Wen Hu. Simultaneous energy harvesting and gait recognition using piezoelectric energy harvester. *IEEE Transactions on Mobile Computing*, 2020.

[74] Jürgen T Geiger, Maximilian Kneißl, Björn W Schuller, and Gerhard Rigoll. Acoustic gait-based person identification using hidden markov models. In *Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge and Workshop*, pages 25–30, 2014.

[75] Yingxue Wang, Yanan Chen, Md Zakirul Alam Bhuiyan, Yu Han, Shenghui Zhao, and Jianxin Li. Gait-based human identification using acoustic sensor and deep neural network. *Future Generation Computer Systems*, 86:1228–1237, 2018.

[76] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. EarEcho: Using Ear Canal Echo for Wearable Authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–24, 2019.

[77] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. Facial expression recognition using ear canal transfer function. In *Proceedings of the 23rd International Symposium on Wearable Computers*, pages 1–9, 2019.

[78] William S Johnston et al. *Development of a signal processing library for extraction of SpO2, HR, HRV, and RR from photoplethysmographic waveforms*. PhD thesis, Worcester Polytechnic Institute., 2006.

[79] Michael T Petterson, Valerie L Begnoche, and John M Graybeal. The effect of motion on pulse oximetry and its clinical significance. *Anesthesia & Analgesia*, 105(6):S78–S84, 2007.

[80] Christina Orphanidou. Signal quality assessment in physiological monitoring: state of the art and practical considerations. 2017.

[81] Leandro Giacomini Rocha, Muqing Liu, Dwaipayan Biswas, Bram-Ernst Verhoef, Sergio Bampi, Chris H Kim, Chris Van Hoof, Mario Konijnenburg, Marian Verhelst, and Nick Van Helleputte. Real-time hr estimation from wrist ppg using binary lstms. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4. IEEE, 2019.

[82] Delaram Jarchi, Dario Salvi, Carmelo Velardo, Adam Mahdi, Lionel Tarassenko, and David A Clifton. Estimation of hrv and spo2 from wrist-worn commercial sensors for clinical settings. In *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 144–147. IEEE, 2018.

[83] Yetong Cao, Huijie Chen, Fan Li, and Yu Wang. Crisp-bp: continuous wrist ppg-based blood pressure measurement. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 378–391, 2021.

[84] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 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*, pages 1–17, 2019.

[85] Sheng Lu, He Zhao, Kihwan Ju, Kunson Shin, Myoungho Lee, Kirk Shelley, and Ki H Chon. Can photoplethysmography variability serve as an alternative approach to obtain heart rate variability information? *Journal of clinical monitoring and computing*, 22(1):23–29, 2008.

[86] Christoph Hoog Antink, Yen Mai, Mikko Peltokangas, Steffen Leonhardt, Niku Oksala, and Antti Vehkaoja. Accuracy of heart rate variability estimated with reflective wrist-ppg in elderly vascular patients. *Scientific reports*, 11(1):1–12, 2021.

[87] Joachim Behar, Aoife Roebuck, Mohammed Shahid, Jonathan Daly, Andre Hallack, Niclas Palmius, John Stradling, and Gari D Clifford. Sleepap: an automated obstructive sleep apnoea screening application for smartphones. *IEEE journal of biomedical and health informatics*, 19(1):325–331, 2014.

[88] Henri Korkalainen, Juhani Aakko, Brett Duce, Samu Kainulainen, Akseli Leino, Sami Nikkonen, Isaac O Afara, Sami Myllymaa, Juha Töyräs, and Timo Leppänen. Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea. *Sleep*, 43(11):zsaa098, 2020.

[89] Alberto G Bonomi, Fons Schipper, Linda M Eerikäinen, Jenny Margarito, Ronald M Aarts, Saeed Babaeizadeh, Helma M de Morree, and Lukas Dekker. Atrial fibrillation detection using photo-plethysmography and acceleration data at the wrist. In *2016 computing in cardiology conference (cinc)*, pages 277–280. IEEE, 2016.

[90] Shamim Nemati, Mohammad M Ghassemi, Vaidehi Ambai, Nino Isakadze, Oleksiy Levantsevych, Amit Shah, and Gari D Clifford. Monitoring and detecting atrial fibrillation using wearable technology. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3394–3397. IEEE, 2016.

[91] Hsiao-Huang Chang, Chuan-Chih Hsu, Chia-Yuen Chen, Wai-Keung Lee, Hao-Teng Hsu, Kuo-Kai Shyu, Jia-Rong Yeh, Pin-Jun Lin, and Po-Lei Lee. A method for respiration rate detection in wrist ppg signal using holo-hilbert spectrum. *IEEE Sensors Journal*, 18(18):7560–7569, 2018.

[92] Harishchandra Dubey, Nicholas Constant, and Kunal Mankodiya. Respire: A spectral kurtosis-based method to extract respiration rate from wearable ppg signals. In *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 84–89. IEEE, 2017.

[93] David Perpetuini, Antonio Maria Chiarelli, Lidia Maddiona, Sergio Rinella, Francesco Bianco, Valentina Bucciarelli, Sabina Gallina, Vincenzo Perciavalle, Vincenzo Vinciguerra, Arcangelo Merla, et al. Multi-site photoplethysmographic and electrocardiographic system for arterial stiffness and cardiovascular status assessment. *Sensors*, 19(24):5570, 2019.

[94] Fen Miao, Xurong Wang, Liyan Yin, and Ye Li. A wearable sensor for arterial stiffness monitoring based on machine learning algorithms. *IEEE Sensors Journal*, 19(4):1426–1434, 2018.

[95] Maarten Falter, Werner Budts, Kaatje Goetschalckx, Véronique Cornelissen, Roselien Buys, et al. Accuracy of apple watch measurements for heart rate and energy expenditure in patients with cardiovascular disease: Cross-sectional study. *JMIR mHealth and uHealth*, 7(3):e11889, 2019.

[96] Ananta Narayanan Balaji, Chen Yuan, Bo Wang, Li-Shiuan Peh, and Huilin Shao. ph watch-leveraging pulse oximeters in existing wearables for reusable, real-time mon-

itoring of ph in sweat. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, pages 262–274, 2019.

[97] Hugo F Posada-Quintero, Natasa Reljin, Aurelie Moutran, Dimitrios Georgopalis, Elaine Choung-Hee Lee, Gabrielle EW Giersch, Douglas J Casa, and Ki H Chon. Mild dehydration identification using machine learning to assess autonomic responses to cognitive stress. *Nutrients*, 12(1):42, 2020.

[98] Parastoo Dehkordi, Ainara Garde, Behnam Molavi, J. Mark Ansermino, and Guy A. Dumont. Extracting instantaneous respiratory rate from multiple photoplethysmogram respiratory-induced variations. *Frontiers in Physiology*, 9, 2018.

[99] Md Nazmul Islam Shuzan, Moajjem Hossain Chowdhury, Muhammad E. H. Chowdhury, M. Monir Uddin, Amith Khandakar, Zaid B. Mahbub, and Naveed Nawaz. A novel non-invasive estimation of respiration rate from photoplethysmograph signal using machine learning model, 2021.

[100] Ali Tazarv and Marco Levorato. A deep learning approach to predict blood pressure from ppg signals, 2021.

[101] Samuel Huthart, Mohamed Elgendi, Dingchang Zheng, Gerard Stansby, and John Allen. Advancing ppg signal quality and know-how through knowledge translation—from experts to student and researcher. *Frontiers in Digital Health*, page 49, 2020.

[102] Bhanupriya Mishra and Neelam Sobha Nirala. A survey on denoising techniques of ppg signal. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–8. IEEE, 2020.

[103] E Friesen and P Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 1978.

[104] Seungchul Lee, Chulhong Min, Alessandro Montanari, Akhil Mathur, Youngjae Chang, Junehwa Song, and Fahim Kawsar. Automatic smile and frown recognition with kinetic earables. In *Proceedings of the 10th Augmented Human International Conference 2019*, AH2019, New York, NY, USA, 2019. Association for Computing Machinery.

[105] Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. Expressear: Sensing fine-grained facial expressions with earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–28, 2021.

[106] Zhongxu Hu, Yiran Zhang, Yang Xing, Yifan Zhao, Dongpu Cao, and Chen Lv. Toward human-centered automated driving: a novel spatial-temporal vision transformer-enabled head tracker. 2022.

[107] György Wersényi. Effect of emulated head-tracking for reducing localization errors in virtual audio simulation. *IEEE transactions on audio, speech, and language processing*, 17(2):247–252, 2009.

[108] Giuseppe Angelo Zito, Dario Cazzoli, Loreen Scheffler, Michael Jäger, René Martin Müri, Urs Peter Mosimann, Thomas Nyffeler, Fred W Mast, and Tobias Nef. Street crossing behavior in younger and older pedestrians: an eye-and head-tracking study. *BMC geriatrics*, 15(1):1–10, 2015.

[109] Barbara G Shinn-Cunningham and Virginia Best. Selective attention in normal and impaired hearing. *Trends in amplification*, 12(4):283–299, 2008.

[110] Antoine Favre-Felix, Carina Graversen, Renskje K Hietkamp, Torsten Dau, and Thomas Lunner. Improving speech intelligibility by hearing aid eye-gaze steering: Conditions with head fixated in a multitalker environment. *Trends in hearing*, 22:2331216518814388, 2018.

[111] Janne Haverinen and Anssi Kemppainen. Global indoor self-localization based on the ambient magnetic field. *Robotics and Autonomous Systems*, 57(10):1028–1035, 2009.

[112] Hua Huang, Hongkai Chen, and Shan Lin. Magtrack: Enabling safe driving monitoring with wearable magnetics. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, pages 326–339, 2019.

[113] Ilknur Umay, Barış Fidan, and Billur Barshan. Localization and tracking of implantable biomedical sensors. *Sensors*, 17(3):583, 2017.

[114] Islam SM Khalil, Alaa Adel, Dalia Mahdy, Mina M Micheal, Mohanad Mansour, Nabila Hamdi, and Sarthak Misra. Magnetic localization and control of helical robots for clearing superficial blood clots. *APL bioengineering*, 3(2):026104, 2019.

[115] Chihwen Cheng, Xueliang Huo, and Maysam Ghovanloo. Towards a magnetic localization system for 3-d tracking of tongue movements in speech-language therapy. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 563–566. IEEE, 2009.

[116] Mark S George, Sarah H Lisanby, and Harold A Sackeim. Transcranial magnetic stimulation: applications in neuropsychiatry. *Archives of General Psychiatry*, 56(4):300–311, 1999.

[117] James Diebel. Representing attitude: Euler angles, unit quaternions, and rotation vectors. *Matrix*, 58(15-16):1–35, 2006.

[118] William Rowan Hamilton. *Lectures on Quaternions: Containing a Systematic Statement of a New Mathematical Method; of which the Principles Were Communicated in 1843 to the Royal Irish Academy; and which Has Since Formed the Subject of Successive Courses of Lectures, Delivered in 1848 and Sub Sequent Years, in the Halls of Trinity College, Dublin: With numerous Illustrative Diagrams, and with Some Geometrical and Physical Applications*. University Press by MH, 1853.

[119] Manon Kok, Jeroen D Hol, and Thomas B Schön. Using inertial sensors for position and orientation estimation. *arXiv preprint arXiv:1704.06053*, 2017.

[120] Shashi Poddar, Vipan Kumar, and Amod Kumar. A comprehensive overview of inertial sensor calibration techniques. *Journal of Dynamic Systems, Measurement, and Control*, 139(1), 2017.

[121] Flavio Ribeiro, Dinei Florencio, Philip A Chou, and Zhengyou Zhang. Auditory augmented reality: Object sonification for the visually impaired. In *2012 IEEE 14th international workshop on multimedia signal processing (MMSP)*, pages 319–324. IEEE, 2012.

[122] Aki Härmä, Julia Jakka, Miikka Tikander, Matti Karjalainen, Tapio Lokki, Jarmo Hiipakka, and Gaëtan Lorho. Augmented reality audio for mobile and wearable appliances. *Journal of the Audio Engineering Society*, 52(6):618–639, 2004.

[123] Colin N Hansen. *Understanding active noise cancellation*. CRC Press, 2002.

[124] Stefan Liebich, Jan-Gerrit Richter, Johannes Fabry, Christopher Durand, Janina Fels, and Peter Jax. Direction-of-arrival dependency of active noise cancellation headphones. In *ASME 2018 Noise Control and Acoustics Division Session presented*

at INTERNOISE 2018. American Society of Mechanical Engineers Digital Collection, 2018.

[125] Freescale Semiconductor. Xtrinsic mag3110 three-axis, digital magnetometer. *MAG3110 Datasheet*, 2013.

[126] Getting Heading and Course Information apple developer. https://developer.apple.com/documentation/corelocation/getting_heading_and_course_information. Accessed: 2020-05-08.

[127] Daniel Hintze, Philipp Hintze, Rainhard D Findling, and René Mayrhofer. A large-scale, long-term analysis of mobile device usage characteristics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–21, 2017.

[128] Magnetic declination. https://en.wikipedia.org/wiki/Magnetic_declination. Accessed: 2020-06-12.

[129] Talat Ozyagcilar. Calibrating an ecompass in the presence of hard and soft-iron interference. *Freescale Semiconductor Ltd*, pages 1–17, 2012.

[130] Arduino nano 33 ble imu. https://www.st.com/resource/en/datasheet/lsm9ds1.pdf. Accessed: 2020-05-07.

[131] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[132] Muhammad Farooq and Edward Sazonov. Accelerometer-based detection of food intake in free-living individuals. *IEEE sensors journal*, 18(9):3752–3758, 2018.

[133] Abhinav Parate, Meng-Chieh Chiu, Chaniel Chadowitz, Deepak Ganesan, and Evangelos Kalogerakis. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, pages 149–161, 2014.

[134] Mohammad Omar Derawi. Accelerometer-based gait analysis, a survey. *Nor Informasjonssikkerhetskonferanse NISK*, 1, 2010.

[135] Warren Gay. *Raspberry Pi hardware reference*. Apress, 2014.

[136] AirPods Pro. https://www.apple.com/uk/airpods-pro/, Online. (Accessed on January 10, 2021).

[137] TDK InvenSense and PACKAGING Cut Tape CT. Mpu-6050. *TDX Invensense*, 2020.

[138] Librosa. https://librosa.org/, Online. (Accessed on November 12, 2020).

[139] Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data. *arXiv preprint arXiv:2006.05919*, 2020.

[140] Scikit-learn. https://scikit-learn.org/stable/index.html, Online. (Accessed on November 12, 2020).

[141] Honor Magic Earbuds. https://www.hihonor.com/global/products/accessories/honor-magic-earbuds/, Online. (Accessed on November 12, 2020).

[142] MINISO Marvel Earphones. https://www.miniso-au.com/en-au/product/145169/marvel-earphones/, Online. (Accessed on November 12, 2020).

[143] Microphone SPU1410LR5H-QB. https://www.mouser.com/datasheet/2/218/SPU1410LR5H-QB-215269.pdf, Online. (Accessed on November 12, 2020).

[144] ReSpeaker Voice Accessory HAT. https://wiki.seeedstudio.com/ReSpeaker_4-Mic_Linear_Array_Kit_for_Raspberry_Pi/, Online. (Accessed on January 10, 2021).

[145] Marília Barandas, Duarte Folgado, Letícia Fernandes, Sara Santos, Mariana Abreu, Patrícia Bota, Hui Liu, Tanja Schultz, and Hugo Gamboa. Tsfel: Time series feature extraction library. *SoftwareX*, 11:100456, 2020.

[146] KL Yick, LT Tse, WT Lo, SP Ng, and J Yip. Effects of indoor slippers on plantar pressure and lower limb emg activity in older women. *Applied ergonomics*, 56:153–159, 2016.

[147] David R Bassett, Lindsay P Toth, Samuel R LaMunion, and Scott E Crouter. Step counting: a review of measurement considerations and health-related applications. *Sports Medicine*, 47(7):1303–1315, 2017.

[148] Wenqiang Chen, Maoning Guan, Yandao Huang, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. Vitype: A cost efficient on-body typing system through

vibration. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9. IEEE, 2018.

[149] Mehul P Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, 18(11):2385–2401, 2009.

[150] Billboard All-Time Top 100 Songs. https://www.billboard.com/articles/news/hot-100-turns-60/8468142/hot-100-all-time-biggest-hits-songs-list, Online. (Accessed on November 12, 2020).

[151] D Esteban, C Galand, Daniel Mauduit, and J Menez. 9.6/7.2 kbps voice excited predictive coder (vepc). In *ICASSP'78. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 307–311. IEEE, 1978.

[152] Frank Angione, Colin Novak, Chris Imeson, Ashley Lehman, Ben Merwin, Tom Pagliarella, Nikolina Samardzic, Peter D'Angela, and Helen Ule. Study of a low frequency emergency siren in comparison to traditional siren technology. In *Proceedings of Meetings on Acoustics 172ASA*, volume 29, page 030008. Acoustical Society of America, 2016.

[153] Matei-Sorin Axente, Ciprian Dobre, Radu-Ioan Ciobanu, and Raluca Purnichescu-Purtan. Gait recognition as an authentication method for mobile devices. *Sensors*, 20(15):4110, 2020.

[154] Andrew H Johnston and Gary M Weiss. Smartwatch-based biometric gait recognition. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE, 2015.

[155] Qi Lin, Weitao Xu, Guohao Lan, Yesheng Cui, Hong Jia, Wen Hu, Mahbub Hassan, and Aruna Seneviratne. Kehkey: Kinetic energy harvester-based authentication and key generation for body area network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–26, 2020.

[156] M Umair Bin Altaf, Taras Butko, and Biing-Hwang Fred Juang. Acoustic gaits: Gait analysis with footstep sounds. *IEEE Transactions on Biomedical Engineering*, 62(8):2001–2011, 2015.

[157] Yifan Zhang, Shuang Song, Rik Vullings, Dwaipayan Biswas, Neide Simões-Capela, Nick Van Helleputte, Chris Van Hoof, and Willemijn Groenendaal. Motion artifact re-

duction for wrist-worn photoplethysmograph sensors based on different wavelengths. *Sensors*, 19(3):673, 2019.

[158] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[159] Bendik B Mjaaland, Patrick Bours, and Danilo Gligoroski. Walk the walk: Attacking gait biometrics by imitation. In *International Conference on Information Security*, pages 361–380. Springer, 2010.

[160] Muhammad Muaaz and Rene Mayrhofer. Smartphone-based gait recognition: From authentication to imitation. *IEEE Transactions on Mobile Computing*, 16(11):3209–3221, 2017.

[161] Yongpan Zou, Meng Zhao, Zimu Zhou, Jiawei Lin, Mo Li, and Kaishun Wu. BiLock: User authentication via dental occlusion biometrics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–20, 2018.

[162] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[163] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[164] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[165] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1106–1115, 2015.

[166] Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their

potential future applications in health care. *International journal of biosensors & bioelectronics*, 4(4):195, 2018.

[167] Issei Imanaga, Hiroshi Hara, Samonn Koyanagi, and Kohtaro Tanaka. Correlation between wave components of the second derivative of plethysmogram and arterial distensibility. *Japanese heart journal*, 39(6):775–784, 1998.

[168] Lakshmanan Suganthi, M Manivannan, Brajesh Kumar Kunwar, George Joseph, and Debashish Danda. Morphological analysis of peripheral arterial signals in takayasu's arteritis. *Journal of clinical monitoring and computing*, 29(1):87–95, 2015.

[169] John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1, 2007.

[170] Tianming Zhao, Yan Wang, Jian Liu, Yingying Chen, Jerry Cheng, and Jiadi Yu. Trueheart: Continuous authentication on wrist-worn wearables using ppg-based biometrics. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 30–39. IEEE, 2020.

[171] B-G Lee, S-J Jung, and W-Y Chung. Real-time physiological and vision monitoring of vehicle driver for non-intrusive drowsiness detection. *IET communications*, 5(17):2461–2469, 2011.

[172] James AC Patterson, Douglas C McIlwraith, and Guang-Zhong Yang. A flexible, low noise reflective ppg sensor platform for ear-worn heart rate monitoring. In *2009 sixth international workshop on wearable and implantable body sensor networks*, pages 286–291. IEEE, 2009.

[173] Brinnae Bent, Benjamin A Goldstein, Warren A Kibbe, and Jessilyn P Dunn. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine*, 3(1):1–9, 2020.

[174] Ming-Zher Poh, Nicholas C Swenson, and Rosalind W Picard. Motion-tolerant magnetic earring sensor and wireless earpiece for wearable photoplethysmography. *IEEE Transactions on Information Technology in Biomedicine*, 14(3):786–794, 2010.

[175] Stefanie Passler, Niklas Müller, and Veit Senner. In-ear pulse rate measurement: a valid alternative to heart rate derived from electrocardiography? *Sensors*, 19(17):3641, 2019.

[176] Peter M Prendergast. Anatomy of the external ear. In *Advanced Cosmetic Otoplasty*, pages 15–21. Springer, 2013.

[177] S-K Stavrakos and Saeema Ahmed-Kristensen. Assessment of anthropometric methods in headset design. In *DS 70: Proceedings of DESIGN 2012, the 12th International Design Conference, Dubrovnik, Croatia*, 2012.

[178] Peter H Charlton, Timothy Bonnici, Lionel Tarassenko, David A Clifton, Richard Beale, and Peter J Watkinson. An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram. *Physiological measurement*, 37(4):610, 2016.

[179] Yekta Said Can, Niaz Chalabianloo, Deniz Ekiz, and Cem Ersoy. Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors*, 19(8):1849, 2019.

[180] Chungkeun Lee, Hang Sik Shin, and Myoungho Lee. Relations between ac-dc components and optical path length in photoplethysmography. *Journal of biomedical optics*, 16(7):077012, 2011.

[181] MPAK Shafique and PA Kyriacou. Photoplethysmographic signals and blood oxygen saturation values during artificial hypothermia in healthy volunteers. 33(12):2065, 2012.

[182] Marc A Russo, Danielle M Santarelli, and Dean O'Rourke. The physiological effects of slow breathing in the healthy human. *Breathe*, 13(4):298–309, 2017.

[183] Consumer Technology Association et al. Ansi/cta standard. physical activity monitoring for heart rate. Technical report, ANSI/CTA-2065, 2018.

[184] Tobias Röddiger, Daniel Wolffram, David Laubenstein, Matthias Budde, and Michael Beigl. Towards respiration rate monitoring using an in-ear headphone inertial measurement unit. In *Proceedings of the 1st International Workshop on Earable Computing*, pages 48–53, 2019.

[185] Giorgio Biagetti, Paolo Crippa, Laura Falaschetti, Leonardo Saraceni, Andrea Tiranti, and Claudio Turchetti. Dataset from ppg wireless sensor for activity monitoring. *Data in brief*, 29:105044, 2020.

[186] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep ppg: large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

[187] Delaram Jarchi and Alexander J Casson. Description of a database containing wrist ppg signals recorded during physical exercise with both accelerometer and gyroscope measures of motion. *Data*, 2(1):1, 2016.

[188] Chang Wei Tan, Christoph Bergmeir, Francois Petitjean, and Geoffrey I Webb. Ieeeppg dataset, June 2020.

[189] H Lee, H Chung, and J Lee. Motion artifact cancellation in wearable photoplethysmography using gyroscope. *IEEE Sensors Journal*, 19(3):1166–1175, 2018.

[190] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[191] Marco AF Pimentel, Alistair EW Johnson, Peter H Charlton, Drew Birrenkott, Peter J Watkinson, Lionel Tarassenko, and David A Clifton. Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Biomedical Engineering*, 64(8):1914–1923, 2016.

[192] Manasa Kalanadhabhatta, Chulhong Min, Alessandro Montanari, and Fahim Kawsar. Fatigueset: A multi-modal dataset for modeling mental fatigue and fatigability. In *International Conference on Pervasive Computing Technologies for Healthcare*, pages 204–217. Springer, 2022.

[193] Luke Everson, Dwaipayan Biswas, Madhuri Panwar, Dimitrios Rodopoulos, Amit Acharyya, Chris H Kim, Chris Van Hoof, Mario Konijnenburg, and Nick Van Helleputte. Biometricnet: Deep learning based biometric identification using wrist-worn ppg. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2018.

[194] Dwaipayan Biswas, Luke Everson, Muqing Liu, Madhuri Panwar, Bram-Ernst Verhoef, Shrishail Patki, Chris H Kim, Amit Acharyya, Chris Van Hoof, Mario Konijnenburg, et al. Cornet: Deep learning framework for ppg-based heart rate estimation

and biometric identification in ambulant environment. *IEEE transactions on biomedical circuits and systems*, 13(2):282–291, 2019.

[195] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[196] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

[197] Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J Wareham, and Cecilia Mascolo. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 69–78, 2021.

[198] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. Selfhar: Improving human activity recognition through self-training with unlabeled data. *arXiv preprint arXiv:2102.06073*, 2021.

[199] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.